Este documento forma parte de un conjunto de informes técnicos que explican la investigación llevada a cabo por los especialistas en estadística de Minitab para desarrollar los métodos y las verificaciones de datos que se utilizan en el Asistente de Minitab Statistical Software.

Pruebas de desviaciones estándar (dos o más muestras)

Revisión general

El Asistente de Minitab incluye dos análisis para comparar muestras independientes con el fin de determinar si la variabilidad defiere significativamente. La Prueba de desviaciones estándar de 2 muestras compara las desviaciones estándar de 2 muestras y la Prueba de desviaciones estándar compara las desviaciones estándar de más de 2 muestras. En este documento, nos referimos a los diseños de k muestras con k = 2 como diseños de 2 muestras y a los diseños de k muestras con k > 2 como diseños de múltiples muestras. Por lo general, estos dos tipos de diseño se estudian individualmente (véase el Apéndice A).

Debido a que la desviación estándar es la raíz cuadrada de la varianza, una prueba de hipótesis que compara desviaciones estándar es equivalente a una prueba de hipótesis que compara varianzas. Se han desarrollado numerosos métodos estadísticos para comparar varianzas de dos o más poblaciones. Entre estas pruebas, la prueba de Levene/Brown-Forsythe es la más robusta y la que más se utiliza. Sin embargo, el rendimiento de la prueba de Levene/Brown-Forsythe en lo que respeta a la potencia es menos satisfactorio que las propiedades del error Tipo I en los diseños de 2 muestras. Pan (1999) demuestra que para algunas poblaciones, incluida la población normal, la potencia de la prueba en los diseños de 2 muestras tiene un límite superior que pudiera estar muy por debajo de 1, independientemente de la magnitud de la diferencia entre las desviaciones estándar. En otras palabras, para estos tipos de datos es más probable que la prueba concluya que no existe diferencia entre las desviaciones estándar,

independientemente de qué tan grande sea la diferencia. Por estas razones, el Asistente utiliza una prueba nueva, la de Bonett, para la Prueba de desviaciones estándar de 2 muestras. Para la prueba de desviaciones estándar con diseños de múltiples muestras, el Asistente utiliza un procedimiento de comparación múltiple (MC).

La prueba de Bonett (2006), una versión modificada de la prueba de igualdad de dos varianzas de Layard (1978), mejora el rendimiento de la prueba con muestras pequeñas. Banga y Fox (2013A) obtienen los intervalos de confianza asociados con la prueba de Bonett y demuestran que son tan exactos como los intervalos de confianza asociados con la prueba de Levene/Brown-Forsythe y son más precisos para la mayoría de las distribuciones. Adicionalmente, Banga y Fox (2013A) determinaron que la prueba de Bonett es tan robusta como la de Levene/Brown-Forsythe y tiene más potencia para la mayoría de las distribuciones.

El procedimiento de comparación múltiple (MC) incluye una prueba general de la homogeneidad, o igualdad, de las desviaciones estándar (o varianzas) para múltiples muestras, que se basa en los intervalos de comparación para cada par de desviaciones estándar. Se obtienen los intervalos de comparación para que la prueba de comparación múltiple sea significativa si, y solo si, no se superpone por lo menos un par de los intervalos de comparación. Banga y Fox (2013B) demostraron que la prueba de comparación múltiple tiene propiedades de error Tipo I y Tipo II que se asemejan a la prueba de Levene/Brown-Forsythe para la mayoría de las distribuciones. Una importante ventaja de la prueba de comparación múltiple es la presentación gráfica de los intervalos de comparación, la cual proporciona una efectiva herramienta visual para identificar muestras con desviaciones estándar diferentes. Cuando solo existen dos muestras en el diseño, la prueba de comparación múltiple es equivalente a la prueba de Bonett.

En este documento, evaluamos la validez de la prueba de Bonett y de la prueba de comparación múltiple para diferentes distribuciones de datos y tamaños de muestra. Además, investigamos el análisis de potencia y tamaño de la muestra utilizado para la prueba de Bonett, que se basa en un método de aproximación basado en muestras grandes. Con base en estos factores, desarrollamos las siguientes verificaciones que el Asistente realiza automáticamente en sus datos y que muestra en la Tarjeta de informe:

- Datos poco comunes
- Normalidad
- Validez de la prueba
- Tamaño de la muestra (solo para la Prueba de desviaciones estándar de 2 muestras)

Métodos de las pruebas de desviaciones estándar

Validez de la prueba de Bonett y de la prueba de comparación múltiple

En su estudio comparativo de pruebas para determinar varianzas iguales, Conover, et al. (1981) hallaron que la prueba de Levene/Brown-Forsythe se encontraba entre las que tienen un mejor desempeño, con base en sus tasas de error Tipo I y Tipo II. Desde ese momento, se han propuesto otros métodos para probar la existencia de varianzas iguales en diseños de 2 y más muestras (Pan, 1999; Shoemaker, 2003; Bonett, 2006). Por ejemplo, Pan demuestra que pese a su robustez y simplicidad de interpretación, la prueba de Levene/Brown-Forsythe no tiene suficiente potencia para detectar diferencias importantes entre 2 desviaciones estándar cuando las muestras se originan de algunas poblaciones, incluida la población normal. Debido a esta importante limitación, el Asistente utiliza la prueba de Bonett para la Prueba de desviaciones estándar de 2 muestras (véase Apéndice A o Banga y Fox, 2013A). Para la prueba de desviaciones estándar con más de 2 muestras, el Asistente utiliza un procedimiento de comparación múltiple con intervalos de comparación que proporciona una presentación gráfica para identificar muestras con diferentes desviaciones estándar cuando la prueba de comparación múltiple es significativa (véase Apéndice A y Banga y Fox, 2013B).

Objetivo

En primer lugar, queríamos evaluar el rendimiento de la prueba de Bonett al comparar desviaciones estándar de dos poblaciones. En segundo lugar, queremos evaluar el rendimiento de la prueba de comparación múltiple al comparar las desviaciones estándar entre más de dos poblaciones. Específicamente, queríamos evaluar la validez de estas pruebas cuando se realizaran con muestras de diferentes tamaños provenientes de diferentes tipos de distribuciones.

Método

Los métodos estadísticos utilizados para la prueba de Bonett y la prueba de comparación múltiple se definen en el Apéndice A. Para evaluar la validez de las pruebas, necesitábamos examinar si las tasas de error Tipo I se aproximaban al nivel de significancia objetivo (alfa) bajo diferentes condiciones. Para ello, realizamos un conjunto de simulaciones para evaluar la validez de la prueba de Bonett al comparar las desviaciones estándar de 2 muestras independientes y otros conjuntos de simulaciones para evaluar la validez de la prueba de comparación múltiple al comparar desviaciones estándar de múltiples (k) muestras independientes, cuando $k > 2$.

Generamos 10,000 pares de múltiples (k) muestras aleatorias de varios tamaños de diferentes distribuciones, utilizando diseños tanto balanceados como no balanceados. A continuación, realizamos una prueba de Bonett bilateral para comparar las desviaciones estándar de las 2 muestras o realizamos una prueba de comparación múltiple para comparar las desviaciones estándar de las k muestras en cada experimento, utilizando un nivel de significancia de $\alpha = 0.05$. Contamos el número de veces sobre 10,000 réplicas que la prueba rechazó la hipótesis nula (cuando en realidad las deviaciones estándar verdaderas eran iguales) y comparamos esta proporción, que se conoce como el nivel de significancia simulado, con el nivel de significancia objetivo. Si la prueba funciona adecuadamente, el nivel de significancia simulado, que representa la verdadera tasa de error Tipo I, debería aproximarse considerablemente al nivel de significancia objetivo. Para obtener más detalles sobre los métodos específicos utilizados para las simulaciones de 2 muestras y de k muestras, véase el Apéndice B.

Resultados

Para las comparaciones de 2 muestras, las tasas de error Tipo I simuladas de la prueba de Bonett se aproximaron al nivel de significancia objetivo cuando las muestras tenían tamaños de moderado a grande, independientemente de la distribución y de si el diseño era balanceado o no balanceado. Sin embargo, cuando se extrajeron muestras pequeñas de poblaciones extremadamente asimétricas, la prueba de Bonett, por lo general, era conservadora y tenía tasas de error Tipo I que eran ligeramente inferiores al nivel de significancia objetivo; es decir, la tasa de error Tipo I objetivo.

Para las comparaciones de múltiples muestras, las tasas de error Tipo I de la prueba de comparación múltiple se aproximaron al nivel de significancia objetivo cuando las muestras tenían tamaños de moderado a grande, independientemente de la distribución y de si el diseño era balanceado o no balanceado. Para las muestras pequeñas y extremadamente asimétricas, sin embargo, la prueba era, por lo general, menos conservadora y tenía tasas de error Tipo I que eran mayores que el nivel de significancia objetivo cuando el número de muestras del diseño es muy elevado.

Los resultados de nuestros estudios coincidían con los de Banga y Fox (2013A) y (2013B). Concluimos que las pruebas de Bonett y comparación múltiple funcionan adecuadamente cuando el tamaño de la muestra más pequeña es por lo menos 20. Por lo tanto, utilizamos este requisito de tamaño de muestra mínimo en la verificación de validez de la prueba en la Tarjeta de informe del Asistente (véase la sección Verificación de datos).

Intervalos de comparación

Cuando una prueba utilizada para comparar una o más desviaciones estándar es estadísticamente significativa, lo cual indica que por lo menos una de las desviaciones estándar es diferente de las demás, el siguiente paso en el análisis es determinar cuáles muestras son estadísticamente diferentes. Una manera intuitiva de realizar esta comparación es graficar los intervalos de confianza asociados con cada muestra e identificar las muestras que no tengan intervalos que se superpongan. Sin embargo, las conclusiones obtenidas a partir de la gráfica

pudieran no coincidir con los resultados de las pruebas debido a que los intervalos de confianza individuales no están diseñados para las comparaciones.

Objetivo

Queríamos desarrollar un método para calcular intervalos de comparación individuales que se puedan utilizar como una prueba de la homogeneidad de las varianzas y como un método para identificar muestras con varianzas diferentes cuando la prueba general sea significativa. Un requisito crítico para el procedimiento de comparación múltiple es que la prueba general sea significativa si, y solo si, no se superpone por lo menos un par de los intervalos de comparación, lo cual indica que las desviaciones estándar de por lo menos dos muestras son diferentes.

Método

El procedimiento de comparación múltiple que utilizamos para comparar múltiples desviaciones estándar se obtiene de múltiples comparaciones en pareja. Cada par de muestras se compara utilizando la prueba de igualdad de desviaciones estándar de dos poblaciones de Bonett (2006). Las comparaciones en pareja utilizan una corrección de multiplicidad con base en una aproximación basada en muestras grandes que demuestra Nayakama (2009). Es preferible el uso de la aproximación basada en muestras grandes que de la comúnmente utilizada corrección de Bonferroni, debido a que el carácter conservador de esta corrección incrementa a medida que el número de muestras es mayor. Por último, los intervalos de comparación resultan de las comparaciones en pareja con base en el mejor procedimiento de aproximación de Hochberg et al. (1982). Para mayor información, véase el Apéndice A.

Resultados

El procedimiento de comparación múltiple satisface el requisito de que la prueba de igualdad de las desviaciones estándar es significativa si, y solo si, no se superpone por lo menos un par de los intervalos de comparación. Si la prueba general no es significativa, entonces se deben superponer todos los intervalos de comparación.

El Asistente muestra los intervalos de comparación en la Gráfica de comparación de desviaciones estándar en el Informe de resumen. Junto a esta gráfica, el Asistente muestra el valor p de la prueba de comparación múltiple, que es la prueba general para la homogeneidad de las desviaciones estándar. Cuando la prueba de desviaciones estándar es estadísticamente significativa, cualquier intervalo de comparación que no se superponga a por lo menos otro intervalo se marca en rojo. Si la prueba de desviaciones estándar no es estadísticamente significativa, entonces ninguno de los intervalos se marca en rojo.

Rendimiento de potencia teórica (diseños de 2 muestras solamente)

Las funciones de potencia teórica de las pruebas de Bonett y comparación múltiple son necesarias para planificar los tamaños de las muestras. Para los diseños de 2 muestras, se puede

derivar una función de potencia teórica aproximada de la prueba utilizando métodos teóricos basados en muestras grandes. Debido a que esta función resulta de métodos de aproximación basados en muestras grandes, necesitamos evaluar sus propiedades cuando la prueba se realiza utilizando muestras pequeñas provenientes de distribuciones normales y no normales. Cuando se comparan las desviaciones estándar de más de dos grupos, sin embargo, la función de potencia teórica de la prueba de comparación múltiple no se obtiene con facilidad.

Objetivo

Queríamos determinar si podíamos utilizar la función de potencia teórica partiendo de la aproximación basada en muestras grandes para evaluar los requisitos de potencia y tamaño de la muestra para la prueba de desviaciones estándar de 2 muestras en el Asistente. Para ello, necesitábamos evaluar si la función de potencia teórica aproximada refleja de manera correcta la potencia real de la prueba de Bonett cuando se realiza con datos provenientes de diversos tipos de distribuciones, incluidas las distribuciones normal y no normal.

Método

En el Apéndice C se deriva la función de potencia teórica aproximada de la prueba de Bonett para diseños de 2 muestras.

Realizamos simulaciones para estimar los niveles de potencia real (a los que nos referimos como niveles de potencia simulada) utilizando la prueba de Bonett. En primer lugar, generamos pares de múltiples muestras aleatorias de varios tamaños de diferentes distribuciones, incluidas las distribuciones normal y no normal. Para cada distribución, realizamos la prueba de Bonett con cada uno de los 10,000 pares de réplicas de las muestras. Para cada par de tamaños de muestra, calculamos la potencia simulada de la prueba para detectar una diferencia específica como la fracción de los 10,000 pares de muestras para los que la prueba es significativa. Para las comparaciones, también calculamos el nivel de potencia correspondiente utilizando la funcion de potencia teórica aproximada de la prueba. Si la aproximación funciona adecuadamente, los niveles de potencia teórica y simulada deberían ser similares. Para mayor información, véase el Apéndice D.

Resultados

Nuestras simulaciones demostraron que para la mayoría de las distribuciones, las funciones de potencia teórica y simulada de la prueba de Bonett son casi iguales para las muestras con tamaños pequeños y se asemejan cuando el tamaño mínimo de la prueba llega a 20. Para las distribuciones simétrica y casi simétrica con colas de ligeras a moderadas, los niveles de potencia teórica son ligeramente mayores que los niveles de potencia simulada (real). Sin embargo, para las distribuciones simétricas y las distribuciones con colas pesadas, éstos son menores que los niveles de potencia simulada (real). Para mayor información, véase el Apéndice D.

En general, los resultados demuestran que la función de potencia teórica proporciona una base adecuada para planificar tamaños de muestra.

Verificaciones de datos

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 y pueden afectar las probabilidades de hallar resultados estadísticamente significativos, especialmente cuando la muestra es pequeña. Los datos poco comunes pueden indicar problemas con la recolección de los datos o pudieran deberse a un comportamiento poco común del proceso que se está estudiando. Por lo tanto, estos puntos de datos con frecuencia merecen investigarse y se deberían corregir cuando sea posible. Los estudios de simulación demuestran que cuando los datos contienen datos atípicos, la prueba de Bonett y la de comparación múltiple son conservadoras (véase Apéndice B). Los niveles de significancia reales de las pruebas son notablemente mayores que el nivel objetivo, particularmente cuando se realiza el análisis con muestras pequeñas.

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 con base en el método descrito por Hoaglin, Iglewicz y Tukey (1986), que se utiliza 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 los dos cuartiles. Este método funciona correctamente cuando existen múltiples valores atípicos, debido a que permite detectar cada valor atípico específico.

Cuando se verifica la presencia de datos poco comunes, el Asistente muestra los siguientes indicadores de estado en la Tarjeta de informe:

Estado	Condición
	No hay puntos de datos poco comunes.

Estado	Condición
	Por lo menos un punto de dato es poco común y pudiera tener una fuerte influencia en los resultados.

Normalidad

A diferencia de la mayoría de las pruebas de igualdad de varianzas, que se obtienen a partir del supuesto de normalidad, las pruebas de Bonett y de comparación múltiple para la igualdad de desviaciones estándar no realizan supuesto alguno sobre la distribución específica de los datos.

Objetivo

Si bien las pruebas de Bonett y de comparación múltiple se basan en métodos de aproximación basados en muestras grandes, queríamos confirmar si funcionan adecuadamente para datos normales y no normales en muestras pequeñas. También queríamos informar al usuario sobre cómo la normalidad de los datos se relaciona con los resultados de las pruebas de desviaciones estándar.

Método

Para evaluar la validez de las pruebas en diferentes condiciones, realizamos simulaciones para examinar la tasa de error Tipo I de las pruebas de Bonett y comparación múltiple con datos normales y no normales con diferentes tamaños de muestra. Para obtener más detalles, véase la sección Prueba para los métodos de desviación estándar y el Apéndice B.

Resultados

Nuestras simulaciones demostraron que la distribución de los datos no tiene un importante efecto sobre las propiedades del error Tipo I de las pruebas de Bonett y comparación múltiple para muestras suficientemente grandes (tamaño de la muestra mínimo ≥ 20). Las pruebas producen tasas de error Tipo I y Tipo II que se aproximan uniformemente a la tasa de error objetivo para datos normales y no normales.

Con base en estos resultados relacionados con la tasa de error Tipo I, el Asistente muestra la información sobre la normalidad en la Tarjeta de informe.

Para los diseños de 2 muestras, el Asistente muestra el siguiente indicador:

Estado	Condición
	Este análisis utiliza la prueba de Bonett. Con muestras suficientemente grandes, la prueba funciona adecuadamente tanto para datos normales como para datos no normales.

Para los diseños de múltiples muestras, el Asistente muestra el siguiente indicador:

Estado	Condición
	Este análisis utiliza una Prueba de comparación múltiple. Con muestras suficientemente grandes, la prueba funciona adecuadamente tanto para datos normales como para datos no normales.

Validez de la prueba

En la sección Pruebas para los métodos de desviación estándar, demostramos que para las comparaciones de 2 y múltiples (k) muestras, las pruebas de Bonett y comparación múltiple producen tasas de error Tipo I que se aproximan a la tasa de error objetivo para datos normales y no normales en diseños balanceados y no balanceados cuando las muestras tienen tamaños de moderado a grande. Sin embargo, cuando todas las muestras son pequeñas, generalmente las pruebas de Bonett y comparación múltiple no funcionan adecuadamente.

Objetivo

Queríamos aplicar una regla para evaluar la validez de los resultados de las pruebas de desviaciones estándar para 2 y múltiples (k) muestras, con base en los datos del usuario.

Método

Para evaluar la validez de las pruebas en diferentes condiciones, realizamos simulaciones para examinar la tasa de error Tipo I de las pruebas de Bonett y comparación múltiple con diversas distribuciones de datos, números de muestras y tamaños de muestra, tal como se describió anteriormente en la sección Pruebas para los métodos de desviación estándar. Para mayor información, véase el Apéndice B.

Resultados

Las pruebas de Bonett y comparación múltiple funcionan adecuadamente cuando el tamaño de la muestra más pequeña es por lo menos 20. Por lo tanto, el Asistente muestra los siguientes indicadores de estado en la Tarjeta de informe para evaluar la validez de las pruebas de desviaciones estándar.

Estado	Condición
	Los tamaños de las muestras son por lo menos 20, de modo que el valor p debería ser exacto.
	Algunos de los tamaños de las muestras son menores que 20, de modo que es posible que el valor p no sea exacto. Considere aumentar los tamaños de las muestras a por lo menos 20.

Tamaño de la muestra (solo para la Prueba de desviaciones estándar de 2 muestras)

Generalmente, se realiza una prueba de hipótesis estadística para recolectar evidencia para rechazar la hipótesis nula de "no diferencia". Si la muestra es muy pequeña, la potencia de la prueba pudiera no ser adecuada para detectar una diferencia que realmente existe, 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 no proporcionaban evidencia suficiente para rechazar la hipótesis nula, queríamos 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 estos sean suficientemente grandes para detectar diferencias importantes con alta probabilidad, el tamaño tampoco debería ser excesivo, ya que ello provocaría que las diferencias despreciables se vuelvan estadísticamente significativas.

Método

El análisis de potencia y tamaño de la muestra para la prueba de desviaciones estándar de 2 muestras se basa en una aproximación de la función de potencia de la prueba de Bonett, que proporciona excelentes estimaciones de la función de potencia real de la prueba (véanse los resultados de las simulaciones resumidos en la sección Rendimiento de la función de potencia teórica).

Resultados

Cuando los datos no proporcionan evidencia suficiente contra la hipótesis nula, el Asistente utiliza la función de potencia aproximada de la prueba de Bonett para calcular diferencias prácticas que se pueden detectar con una probabilidad del 80% y 90% con el tamaño de muestra dado. Además, si el usuario proporciona una diferencia práctica particular de interés, el Asistente utiliza la función de potencia de la prueba de aproximación de normalidad para calcular tamaños de muestra que ofrezcan una probabilidad del 80% y 90% de detectar la diferencia.

Para ayudar a interpretar los resultados, la tarjeta de informe del Asistente correspondiente a la prueba de desviaciones estándar de 2 muestras exhibe los siguientes indicadores de estado cuando se verifican la potencia y el tamaño de la muestra.

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

Referencias

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

Banga, S.J. y Fox, G.D. (2013A). On Bonett's Robust Confidence Interval for a Ratio of Standard Deviations. *White paper, Minitab Inc*.

Banga, S.J. y Fox, G.D. (2013B) A graphical multiple comparison procedure for several standard deviations. *White paper, Minitab Inc*.

Bonett, D.G. (2006). Robust confidence interval for a ratio of standard deviations. *Applied Psychological Measurements*, 30, 432-439.

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

Conover, W.J., Johnson, M.E. y Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics, 23*, 351-361.

Gastwirth, J. L. (1982). Statistical properties of a measure of tax assessment uniformity. *Journal of Statistical Planning and Inference, 6*, 1-12.

Hochberg, Y., Weiss G. y Hart, S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association, 77,* 767-772.

Layard, M.W.J. (1973). Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association*, 68, 195-198.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Probability and statistics* (278-292). Stanford University Press, Palo Alto, California.

Nakayama, M.K. (2009). Asymptotically valid single-stage multiple-comparison procedures. *Journal of Statistical Planning and Inference*, 139, 1348-1356.

Pan, G. (1999) On a Levene type test for equality of two variances. *Journal of Statistical Computation and Simulation*, 63, 59-71.

Shoemaker, L. H. (2003). Fixing the F test for equal variances. *The American Statistician, 57 (2)*, 105-114.

Apéndice A: Método para la prueba de Bonett y la prueba de comparación múltiple

Los supuestos subyacentes para realizar inferencias sobre las desviaciones estándar o varianzas utilizando el método de Bonett (diseños de 2 muestras) o el procedimiento de comparación múltiple (MC) (diseños de múltiples muestras) se pueden describir de la manera siguiente. Supongamos que $X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k}$ son k ($k \geq 2$) muestras aleatorias independientes, con cada muestra extraída de una distribución con una media desconocida μ_i y una varianza σ_i^2, respectivamente, para $i = 1, \ldots, k$. Asumamos que las distribuciones principales de las muestras tienen una curtosis común finita $\gamma = E(Y - \mu)^4/\sigma^4 < \infty$. Si bien este supuesto es crucial para las derivaciones teóricas, no es crítico para la mayoría de las aplicaciones prácticas donde las muestras sean suficientemente grandes (Banga and Fox, 2013A).

Método A1: Prueba de igualdad de dos varianzas de Bonett

La prueba de Bonett solo se aplica a diseños de 2 muestras donde se comparan dos varianzas o desviaciones estándar. La prueba es una versión modificada de la prueba de igualdad de dos varianzas de Layard (1978) en diseños de dos muestras. Una prueba de igualdad de dos varianzas bilateral de Bonett con nivel de significancia α rechaza la hipótesis nula de igualdad si, y solo si,

$$\left|\ln(c\, S_1^2/S_2^2)\right| > z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

donde:

S_i es la desviación estándar de la muestra i

$g_i = (n_i - 3)/n_i, i = 1{,}2$

$z_{\alpha/2}$ se refiere al percentil $\alpha/2$ superior de la distribución normal estándar

$\hat{\gamma}_P$ es el estimador de curtosis agrupada especificado como:

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

En la expresión del estimador de curtosis agrupada, m_i es la media recortada de la muestra i, con la proporción de recorte, $1/[2(n_i - 4)^{1/2}]$.

En lo anterior, se incluye la constante c como un pequeño ajuste de las muestras con el fin de reducir el efecto de las probabilidades de error de colas desiguales en diseños no balanceados. Esta constante se especifica como $c = c_1/c_2$, donde

$$c_{\mathrm{i}} = \frac{n_{\mathrm{i}}}{n_{\mathrm{i}} - z_{\alpha/2}}, i = 1{,}2$$

Si el diseño está balanceado; es decir, si $n_1 = n_2$, entonces el valor p de la prueba se obtiene como

$$\mathrm{P} = 2\,\Pr(Z > z)$$

donde Z es una variable aleatoria distribuida como la distribución normal estándar y z el valor observado de las siguientes estadísticas con base en los datos disponibles. La estadística es

$$Z = \frac{\ln(C\ S_1^2/S_2^2)}{se}$$

donde

$$se = \sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

Sin embargo, si el diseño no estuviera balanceado, entonces el valor p de la prueba se obtiene como

$$P = 2\min(\alpha_L, \alpha_U)$$

donde $\alpha_L = \Pr(Z > z_L)$ y $\alpha_U = \Pr(Z > z_U)$. La variable z_L es la raíz más pequeña de la función

$$L(z, S_1, S_2, n_1, n_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\ se + \ln\frac{S_1^2}{S_2^2} - \ln\rho_o^2\,, z < \min(n_1, n_2)$$

y z_U es la raíz más pequeña de la función $L(z, S_2, S_1, n_2, n_1)$.

Método A2: Prueba de comparación múltiple e intervalos de comparación

Supongamos que hay k $(k \geq 2)$ grupos o muestras independientes. Nuestro objetivo era construir un sistema de k intervalos para las desviaciones estándar de la población, de modo que la prueba de igualdad de las desviaciones estándar sea significativa si, y solo si, por lo menos dos de los k intervalos no se superponen. Estos intervalos se conocen como intervalos de comparación. Este método de comparación es similar a los procedimientos de comparaciones múltiples de las medias en los modelos ANOVA de un solo factor, que en un principio desarrolló Tukey-Kramer y que posteriormente generalizaría Hochberg, et al. (1982).

Comparación de dos desviaciones estándar

Para los diseños de 2 muestras, los intervalos de confianza de la relación de las desviaciones estándar asociadas con la prueba de Bonett se puede calcular directamente para evaluar el

tamaño de la diferencia entre las desviaciones estándar (Banga y Fox, 2013A). De hecho, utilizamos este enfoque para Estadísticas > 2 varianzas en la versión 17 de Minitab. En el Asistente, sin embargo, queríamos proporcionar intervalos de comparación que fueran más fáciles de interpretar que el intervalo de confianza de la relación de las desviaciones estándar. Para ello, utilizamos el procedimiento de Bonett descrito en Método A1 para determinar los intervalos de comparación para dos muestras.

Cuando hay dos muestras, la prueba de igualdad de varianzas de Bonett es significativa si, y solo si, el siguiente intervalo de aceptación asociado con la prueba de igualdad de varianzas de Bonett no contiene 0.

$$\ln(c_1 S_1^2) - \ln(c_2 S_2^2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

donde la estimación de la curtosis agrupada $\hat{\gamma}_P$, y $g_i, i = 1{,}2$ aparecen como se especificaron anteriormente.

A partir de este intervalo, deducimos los siguientes intervalos de comparación, de modo que la prueba de igualdad de varianzas o la desviación estándar sea significativa si, y solo si, no se superponen. Estos dos intervalos son

$$\left[S_i\sqrt{\mathrm{C_i}\exp(-z_{\alpha/2}V_i)}\,, S_i\sqrt{C_i\exp(z_{\alpha/2}V_i)}\,\right], i = 1{,}2$$

donde

$$V_i = \frac{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}}}{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}} + \sqrt{\frac{\hat{\gamma}_P - g_j}{n_j - 1}}}\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1} + \frac{\hat{\gamma}_P - g_j}{n_j - 1}}, i = 1{,}2; j = 1{,}2; i \neq j$$

Utilizando estos intervalos como procedimiento para probar la igualdad de la desviación estándar es equivalente a la prueba de igualdad de desviaciones estándar de Bonett. Específicamente, los intervalos no se superponen si, y solo si, la prueba de igualdad de desviación estándar de Bonett es significativa. No obstante, tenga en cuenta que estos intervalos no son intervalos de confianza, sino que solo son apropiados para comparaciones múltiples de desviaciones estándar. Hochberg et al. hacen referencia a intervalos similares para comparar medias como intervalos de incertidumbre por la misma razón. Nos referimos a estos intervalos como intervalos de comparación.

Debido a que el procedimiento para los intervalos de comparación es equivalente a la prueba de igualdad de desviación estándar de Bonett, el valor p asociado a los intervalos de comparación es idéntico al valor p de la prueba de igualdad de dos desviaciones estándar de Bonett, descrita anteriormente.

Comparación de múltiples desviaciones estándares

Cuando hay más de dos grupos o muestras, los k intervalos de comparación se deducen de $k(k-1)/2$ pruebas de igualdad de desviaciones estándar simultáneas en pareja con un nivel de significancia por familia de α. Más específicamente, supongamos que $X_{i1}, \ldots, X_{in_i}$ y $X_{j1}, \ldots, X_{jn_j}$ son los datos de las muestras de cualquier par (i,j) de muestras. Del mismo modo que el caso de 2 muestras, la prueba de igualdad de las desviaciones estándar para el particular par (i,j) de muestras es significativo al nivel α' si, y solo si, el intervalo

$$\ln(c_i S_i^2) - \ln(c_j S_j^2) \pm z_{\alpha'/2}\sqrt{\frac{\hat{\gamma}_{ij} - g_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - g_j}{n_j - 1}}$$

no contiene 0. En el $\hat{\gamma}_{ij}$ anterior se encuentra el estimador de curtosis agrupada basado en el par (i,j) de muestras y se expresa como

$$\hat{\gamma}_{ij} = (n_i + n_j)\frac{\sum_{l=1}^{n_i}(X_{il} - m_i)^4 + \sum_{l=1}^{n_j}(X_{jl} - m_j)^4}{\left[(n_i - 1)S_i^2 + (n_j - 1)S_j^2\right]^2}$$

Además, tal como se definió anteriormente, m_i es la media recortada de la muestra i, con la proporción de recorte, $1/[2(n_i - 4)^{1/2}]$ y

$$g_i = \frac{n_i - 3}{n_i}, g_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha'/2}}, c_j = \frac{n_j}{n_j - z_{\alpha'/2}}$$

Debido a que existen $k(k-1)/2$ pruebas en parejas simultáneas, se debe elegir el nivel α' para que la tasa de error por familia se aproxime al nivel de significancia objetivo α. Uno de los posibles ajustes se basa en la aproximación de Bonferroni. Sin embargo, es bien sabido que las correcciones de Bonferroni se vuelven cada vez más conservadoras según aumente el número de muestras en el diseño. Un mejor enfoque se basa en una aproximación normal proporcionada por Nakayama (2008). Con este enfoque, simplemente reemplazamos $z_{\alpha'/2}$ con $q_{\alpha,k}/\sqrt{2}$, donde $q_{\alpha,k}$ es el punto α superior del rango de k variables aleatorias normales estándar independientes e idénticamente distribuidas; es decir,

$$\Pr\left(\max_{1 \le i < j \le k} |Z_i - Z_j| \le q_{\alpha,k}\right) = 1 - \alpha$$

donde $Z_1, \ldots, Z_k$ son variables aleatorias normales estándar independientes e idénticamente distribuidas.

Además, utilizar un método similar al de Hochberg et al. (1982), el que mejor se aproxima al procedimiento en pareja descrito anteriormente, rechaza la hipótesis nula de la igualdad de las desviaciones estándar si, y solo si, para algún par (i,j) de muestras

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > q_{\alpha,k}(V_i + V_j)/\sqrt{2}$$

donde se elige V_i para minimizar la cantidad

$$\sum_{i \ne j}\sum (V_i + V_j - b_{ij})^2$$

con

$$b_{ij} = \sqrt{\frac{\hat{Y}_{ij} - g_i}{n_i - 1} + \frac{\hat{Y}_{ij} - g_j}{n_j - 1}}$$

La solución de este problema se ilustra en Hochberg et al. (1982) es elegir

$$V_i = \frac{(k-1)\sum_{j \neq i} b_{ij} - \sum\sum_{1 \leq j < l \leq k} b_{jl}}{(k-1)(k-2)}$$

Por lo tanto, la prueba basada en el procedimiento de aproximación es significativa si, y solo si, no se superpone por lo menos un par de los siguientes k intervalos.

$$\left[S_i \sqrt{\mathrm{C_i} \exp(-q_{\alpha,k} V_i / \sqrt{2})}, S_i \sqrt{C_i \exp(q_{\alpha,k} V_i / \sqrt{2})} \right], i = 1, \ldots, k$$

Para calcular el valor p general asociado con la prueba de comparación múltiple, supongamos que P_{ij} es el valor p asociado con cualquier par (i, j) de muestras. Por lo tanto, el valor p general asociado con la prueba de comparación múltiple es

$$P = \min\{ P_{ij}, 1 \leq i < j \leq k \}$$

Para calcular P_{ij}, utilizamos el algoritmo del diseño de 2 muestras proporcionado en el Método A1 utilizando

$$se = V_i + V_j$$

donde V_i se calcula como se indicó anteriormente.

Más específicamente, si $n_i \neq n_j$

$$P_{ij} = \min(\alpha_L, \alpha_U)$$

donde $\alpha_L = \Pr(Q > z_L \sqrt{2})$, $\alpha_U = \Pr(Q > z_U \sqrt{2})$, la variable z_L es la raíz más pequeña de la función $L(z, S_i, S_j, n_i, n_j)$, la variable z_U es la raíz más pequeña de la función $L(z, S_j, S_i, n_j, n_i)$ y Q es una variable aleatoria que tiene una distribución de rangos, tal como se definió anteriormente.

Si $n_i = n_j$ entonces $P_{ij} = \Pr(Q > |z_o| \sqrt{2})$ donde

$$z_o = \frac{\ln S_i^2 - \ln S_j^2}{V_i + V_j}$$

Apéndice B: Validez de la prueba de Bonett y la prueba de comparación múltiple

Simulación B1: Validez de la prueba de Bonett (modelos de 2 muestras, diseños balanceados y no balanceados)

Generamos pares de muestras aleatorias con tamaños de pequeño a moderado a partir de distribuciones con diferentes propiedades. Las distribuciones incluyeron:

- Distribución normal estándar (N(0,1))
- Distribuciones simétricas con colas ligeras, incluida la distribución uniforme (U(0,1)) y la distribución Beta con ambos parámetros establecidos en 3 (B(3,3))
- Distribuciones simétricas con colas pesadas, incluidas las distribuciones t con 5 y 10 grados de libertad (t(5),t(10)), y la distribución de Laplace con ubicación 0 y escala 1 (Lpl))
- Distribuciones asimétricas con colas pesadas, incluida la distribución exponencial con escala 1 (Exp) y distribuciones de chi-cuadrado con 5 y 10 grados de libertad (Chi(5), Chi(10))
- Distribuciones asimétricas hacia la izquierda con colas ligeras; específicamente, la distribución Beta con los parámetros establecidos en 8 y 1, respectivamente (B(8,1))

Además, para evaluar el efecto directo de los valores atípicos, generamos pares de muestras a partir de distribuciones normales contaminadas definidas como

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

donde p es el parámetro de mezcla y $1-p$ es la proporción de contaminación (que es igual a la proporción de valores atípicos). Seleccionamos dos poblaciones normales contaminadas para el estudio: $CN(0.9,3)$, donde 10% de la población eran valores atípicos y $CN(0.8,3)$, donde 20% de la población eran valores atípicos. Ambas distribuciones son simétricas con largas colas debido a los valores atípicos.

Realizamos una prueba de Bonett bilateral con un nivel de significancia objetivo de $\alpha = 0.05$ a cada par de muestras de cada distribución. Debido a que los niveles de significancia simulados,

en cada caso, se basaron en 10,000 pares de muestras replicadas, y debido a que utilizamos un nivel de significancia de 5%, el error de simulación era $\sqrt{0.95(0.05)/10{,}000} = 0.2\%$.

Los resultados de las simulaciones se resumen abajo en la Tabla 1.

Tabla 1 Niveles de significancia simulados para una prueba de Bonett bilateral y diseños de 2 muestras balanceados y no balanceados. El nivel de significancia objetivo es 0.05.

Distribución	n_1, n_2	Nivel simulado	Distribución	n_1, n_2	Nivel simulado
N(0,1)	10, 10	0.038	Exp	10, 10	0.052
	20, 10	0.043		20, 10	0.051
	20, 20	0.045		20, 20	0.049
	30, 10	0.044		30, 10	0.044
	30, 20	0.046		30, 20	0.042
	25, 25	0.048		25, 25	0.043
	30, 30	0.048		30, 30	0.042
	40, 40	0.051		40, 40	0.042
	50, 50	0.047		50, 50	0.039
t(5)	10, 10	0.044	Chi(5)	10, 10	0.040
	20, 10	0.042		20, 10	0.043
	20, 20	0.046		20, 20	0.040
	30, 10	0.041		30, 10	0.039
	30, 20	0.046		30, 20	0.043
	25, 25	0.048		25, 25	0.042
	30, 30	0.043		30, 30	0.043
	40, 40	0.046		40, 40	0.040
	50, 50	0.050		50, 50	0.039

Distribución	n_1, n_2	Nivel simulado	Distribución	n_1, n_2	Nivel simulado
t(10)	10, 10	0.041	Chi(10)	10, 10	0.044
	20, 10	0.040		20, 10	0.042
	20, 20	0.045		20, 20	0.041
	30, 10	0.046		30, 10	0.043
	30, 20	0.045		30, 20	0.045
	25, 25	0.046		25, 25	0.046
	30, 30	0.048		30, 30	0.038
	40, 40	0.045		40, 40	0.042
	50, 50	0.051		50, 50	0.049
Lpl	10, 10	0.054	B(8,1)	10, 10	0.053
	20, 10	0.056		20, 10	0.045
	20, 20	0.055		20, 20	0.048
	30, 10	0.057		30, 10	0.042
	30, 20	0.058		30, 20	0.047
	25, 25	0.057		25, 25	0.041
	30, 30	0.053		30, 30	0.040
	40, 40	0.047		40, 40	0.042
	50, 50	0.048		50, 50	0.038

Distribución	n_1, n_2	Nivel simulado	Distribución	n_1, n_2	Nivel simulado
B(3,3)	10, 10	0.032	CN(0.9,3)	10, 10	0.024
	20, 10	0.037		20, 10	0.022
	20, 20	0.042		20, 20	0.018
	30, 10	0.039		30, 10	0.019
	30, 20	0.038		30, 20	0.020
	25, 25	0.039		25, 25	0.019
	30, 30	0.041		30, 30	0.015
	40, 40	0.044		40, 40	0.020
	50, 50	0.046		50, 50	0.017
U(0,1)	10, 10	0.030	CN(0.8,3)	10, 10	0.022
	20, 10	0.032		20, 10	0.019
	20, 20	0.031		20, 20	0.020
	30, 10	0.034		30, 10	0.017
	30, 20	0.034		30, 20	0.020
	25, 25	0.034		25, 25	0.021
	30, 30	0.037		30, 30	0.017
	40, 40	0.043		40, 40	0.023
	50, 50	0.043		50, 50	0.020

Tal como se muestra en la Tabla 1, cuando los tamaños de las muestras son más pequeños, los niveles de significancia simulados de la prueba de Bonett son menores que el nivel de significancia objetivo (0.05) para distribuciones simétricas o casi simétricas con colas de ligeras a moderadas. Sin embargo, los niveles simulados tienden a ser un poco más grandes que el nivel objetivo cuando se originan muestras pequeñas a partir de distribuciones altamente asimétricas.

Cuando las muestras tienen un tamaño moderadamente grande o grande, los niveles de significancia simulados se aproximan al nivel objetivo para todas las distribuciones. De hecho, la prueba funciona razonablemente bien, incluso para las distribuciones altamente asimétricas, tales como la distribución exponencial y la distribución Beta (8,1).

Además, los valores atípicos parecieran tener mayor impacto en muestras pequeñas que en grandes. Los niveles de significancia simulados para las poblaciones normales contaminadas se estabilizaban en aproximadamente 0.020 cuando el tamaño mínimo de las dos muestras era 20.

Cuando el tamaño mínimo de las dos muestras es 20, los niveles de significancia simulados se encuentran en el intervalo [0.038, 0.058], a excepción de la distribución uniforme plana y las distribuciones normales contaminadas. Si bien un nivel de significancia simulado de 0.040 es muy conservador para un nivel objetivo de 0.05, esta tasa de error Tipo I pudiera ser aceptable para la mayoría de los objetivos prácticos. Por lo tanto, concluimos que la prueba de Bonett es válida cuando el tamaño mínimo de dos muestras es por lo menos 20.

Simulación B2: Validez de la prueba de comparación múltiple (modelos con múltiples muestras)

Parte I: Diseños balanceados

Realizamos una simulación para examinar el rendimiento de la prueba de comparación múltiple en modelos de múltiples muestras con diseños balanceados. Generamos k muestras de igual tamaño a partir de la misma distribución, utilizando el conjunto de distribuciones anteriormente mencionadas en la simulación B1. Seleccionamos que el número de muestras en un diseño fuera $k = 3$, $k = 4$ y $k = 6$ y fijamos el tamaño de las k muestras de cada experimento en 10, 15, 20, 25, 50 y 100.

Realizamos una prueba de comparación múltiple bilateral con un nivel de significancia objetivo de $\alpha = 0.05$ a las mismas muestras de cada caso de diseño. Debido a que los niveles de significancia simulados, en cada caso, se basaron en 10,000 pares de muestras replicadas, y debido a que utilizamos un nivel de significancia de 5%, el error de simulación era $\sqrt{0.95(0.05)/10{,}000} = 0.2\%$.

Los resultados de las simulaciones se resumen abajo en las Tablas 2a y 2b.

Tabla 2a Niveles de significancia simulados para una prueba de comparación múltiple bilateral en diseños balanceados de múltiples muestras. El nivel de significancia objetivo de la prueba es 0.05.

Distribución	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nivel simulado	n_i	Nivel simulado	n_i	Nivel simulado
N(0,1)	10	0.038	10	0.038	10	0.036
	15	0.040	15	0.041	15	0.039
	20	0.039	20	0.040	20	0.041

Distribución	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nivel simulado	n_i	Nivel simulado	n_i	Nivel simulado
	25	0.045	25	0.047	25	0.047
	50	0.046	50	0.046	50	0.052
	100	0.049	100	0.049	100	0.052
t(5)	10	0.042	10	0.044	10	0.042
	15	0.041	15	0.044	15	0.046
	20	0.043	20	0.045	20	0.045
	25	0.046	25	0.048	25	0.046
	50	0.040	50	0.039	50	0.038
	100	0.038	100	0.040	100	0.040
T(10)	10	0.033	10	0.037	10	0.038
	15	0.040	15	0.042	15	0.041
	20	0.042	20	0.043	20	0.043
	25	0.041	25	0.042	25	0.045
	50	0.047	50	0.044	50	0.047
	100	0.048	100	0.046	100	0.047
Lpl	10	0.056	10	0.063	10	0.071
	15	0.056	15	0.061	15	0.063
	20	0.054	20	0.058	20	0.059
	25	0.051	25	0.056	25	0.58
	50	0.045	50	0.051	50	0.049
	100	0.044	100	0.047	100	0.050
B(3,3)	10	0.031	10	0.031	10	0.031
	15	0.037	15	0.036	15	0.034
	20	0.035	20	0.036	20	0.037

Distribución	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nivel simulado	n_i	Nivel simulado	n_i	Nivel simulado
	25	0.039	25	0.038	25	0.040
	50	0.044	50	0.044	50	0.044
	100	0.044	100	0.046	100	0.043
U(0,1)	10	0.029	10	0.025	10	0.023
	15	0.026	15	0.027	15	0.026
	20	0.028	20	0.030	20	0.028
	25	0.034	25	0.033	25	0.032
	50	0.041	50	0.036	50	0.036
	100	0.048	100	0.047	100	0.045
Exp	10	0.063	10	0.073	10	0.076
	15	0.056	15	0.058	15	0.064
	20	0.051	20	0.053	20	0.057
	25	0.043	25	0.045	25	0.050
	50	0.033	50	0.037	50	0.038
	100	0.033	100	0.035	100	0.035

Tabla 2b Niveles de significancia simulados para una prueba de comparación múltiple bilateral en diseños balanceados de múltiples muestras. El nivel de significancia objetivo de la prueba es 0.05.

Distribución	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nivel simulado	n_i	Nivel simulado	n_i	Nivel simulado
Chi(5)	10	0.040	10	0.046	10	0.048
	15	0.043	15	0.046	15	0.049
	20	0.040	20	0.040	20	0.042
	25	0.040	25	0.045	25	0.042

Distribución	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nivel simulado	n_i	Nivel simulado	n_i	Nivel simulado
	50	0.037	50	0.038	50	0.040
	100	0.036	100	0.037	100	0.038
Chi(10)	10	0.042	10	0.045	10	0.045
	15	0.038	15	0.044	15	0.047
	20	0.036	20	0.039	20	0.040
	25	0.043	25	0.044	25	0.045
	50	0.041	50	0.040	50	0.042
	100	0.038	100	0.040	100	0.042
B(8,1)	10	0.058	10	0.060	10	0.066
	15	0.057	15	0.061	15	0.064
	20	0.049	20	0.051	20	0.055
	25	0.044	25	0.046	25	0.050
	50	0.037	50	0.037	50	0.039
	100	0.037	100	0.038	100	0.039
CN(0.9,3)	10	0.020	10	0.018	10	0.016
	15	0.022	15	0.020	15	0.017
	20	0.014	20	0.012	20	0.008
	25	0.011	25	0.011	25	0.008
	50	0.009	50	0.007	50	0.006
	100	0.010	100	0.008	100	0.008
CN(0.8, 3)	10	0.017	10	0.015	10	0.011
	15	0.013	15	0.011	15	0.008
	20	0.012	20	0.012	20	0.009
	25	0.013	25	0.010	25	0.009

Distribución	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nivel simulado	n_i	Nivel simulado	n_i	Nivel simulado
	50	0.011	50	0.011	50	0.009
	100	0.014	100	0.012	100	0.010

Tal como se muestra en las Tablas 2a y 2b, cuando el tamaño de la muestra es pequeño, la prueba de comparación múltiple es generalmente conservadora para distribuciones simétricas y casi simétricas en diseños balanceados. Sin embargo, la prueba es liberal con muestras pequeñas obtenidas de poblaciones altamente asimétricas, tales como las distribuciones exponencial y Beta (8, 1). A medida que aumenta el tamaño de la muestra, sin embargo, los niveles de significancia simulados se aproximan al nivel de significancia objetivo (0.05). Además, el número de muestras no pareciera tener un fuerte efecto sobre el rendimiento de la prueba en el caso de muestras con tamaños moderados. Cuando los datos están contaminados con valores atípicos, sin embargo, existe un notable impacto sobre el rendimiento de la prueba. La prueba es constante y excesivamente conservadora cuando están presentes valores atípicos en los datos.

Parte II: Diseños no balanceados

Realizamos una simulación para examinar el rendimiento de la prueba de comparación múltiple en diseños no balanceados. Generamos 3 muestras de igual tamaño a partir de la misma distribución, utilizando el conjunto de distribuciones anteriormente descritas en la simulación B1. En el primer conjunto de experimentos, el tamaño de las dos primeras muestras era $n_1 = n_2 = 10$ y el tamaño de la tercera muestra era $n_3 = 15, 20, 25, 50, 100$. En el segundo conjunto de experimentos, el tamaño de las dos primeras muestras era $n_1 = n_2 = 15$ y el tamaño del tercer conjunto de muestras era $n_3 = 20, 25, 30, 50, 100$. En el tercer conjunto de experimentos, establecimos el tamaño de la muestra mínimo en 20, con el tamaño de las dos primeras muestras en $n_1 = n_2 = 20$ y el tamaño de la tercera muestra en $n_3 = 25, 30, 40, 50, 100$..

Realizamos una prueba de comparación múltiple bilateral con un nivel de significancia objetivo de $\alpha = 0.05$ a las tres mismas muestras de cada distribución. Debido a que los niveles de significancia simulados, en cada caso, se basaron en 10,000 pares de muestras replicadas, y debido a que utilizamos un nivel de significancia de 5%, el error de simulación era $\sqrt{0.95(0.05)/10{,}000} = 0.2\%$..

Los resultados de las simulaciones se resumen abajo en las Tablas 3a y 3b.

Tabla 3a Niveles de significancia simulados para la prueba de comparación múltiple en diseños no balanceados con múltiples muestras. El nivel de significancia objetivo de la prueba es 0.05.

Distribución	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nivel simulado	n_3	Nivel simulado	n_3	Nivel simulado
N(0,1)	15	0.032	20	0.040	25	0.045
	20	0.037	25	0.039	30	0.041
	25	0.038	30	0.037	40	0.043
	50	0.041	50	0.044	50	0.041
	100	0.042	100	0.042	100	0.044
t(5)	15	0.040	20	0.042	25	0.043
	20	0.036	25	0.040	30	0.037
	25	0.044	30	0.036	40	0.038
	50	0.033	50	0.036	50	0.035
	100	0.032	100	0.031	100	0.032
t(10)	15	0.039	20	0.042	25	0.042
	20	0.038	25	0.041	30	0.040
	25	0.040	30	0.041	40	0.041
	50	0.037	50	0.043	50	0.042
	100	0.036	100	0.039	100	0.040
Lpl	15	0.059	20	0.060	25	0.054
	20	0.057	25	0.054	30	0.051
	25	0.056	30	0.051	40	0.050
	50	0.049	50	0.051	50	0.050
	100	0.048	100	0.047	100	0.046
B(3,3)	15	0.034	20	0.033	25	0.037
	20	0.031	25	0.035	30	0.039
	25	0.031	30	0.034	40	0.039

Distribución	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nivel simulado	n_3	Nivel simulado	n_3	Nivel simulado
	50	0.036	50	0.039	50	0.038
	100	0.035	100	0.039	100	0.039
U(0,1)	15	0.027	20	0.030	25	0.032
	20	0.030	25	0.030	30	0.031
	25	0.028	30	0.032	40	0.036
	50	0.039	50	0.034	50	0.037
	100	0.042	100	0.038	100	0.042
Exp	15	0.061	20	0.053	25	0.042
	20	0.060	25	0.052	30	0.047
	25	0.054	30	0.049	40	0.043
	50	0.050	50	0.046	50	0.041
	100	0.044	100	0.040	100	0.040

Tabla 3b Niveles de significancia simulados para la prueba de comparación múltiple en diseños no balanceados con múltiples muestras. El nivel de significancia objetivo de la prueba es 0.05.

Distribución	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nivel simulado	n_3	Nivel simulado	n_3	Nivel simulado
Chi(5)	15	0.047	20	0.045	25	0.041
	20	0.043	25	0.042	30	0.039
	25	0.043	30	0.039	40	0.040
	50	0.039	50	0.037	50	0.040
	100	0.034	100	0.035	100	0.034
Chi(10)	15	0.043	20	0.042	25	0.042
	20	0.039	25	0.038	30	0.041
	25	0.040	30	0.041	40	0.038
	50	0.038	50	0.041	50	0.042

Distribución	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nivel simulado	n_3	Nivel simulado	n_3	Nivel simulado
	100	0.035	100	0.034	100	0.035
B(8,1)	15	0.056	20	0.052	25	0.048
	20	0.054	25	0.046	30	0.044
	25	0.050	30	0.047	40	0.046
	50	0.046	50	0.043	50	0.043
	100	0.043	100	0.042	100	0.044
CN(0.9,3)	15	0.017	20	0.020	25	0.017
	20	0.020	25	0.019	30	0.012
	25	0.017	30	0.016	40	0.013
	50	0.019	50	0.016	50	0.012
	100	0.014	100	0.016	100	0.010
CN(0.8, 3)	15	0.012	20	0.013	25	0.013
	20	0.016	25	0.012	30	0.012
	25	0.014	30	0.010	40	0.010
	50	0.015	50	0.010	50	0.013
	100	0.012	100	0.011	100	0.010

Los niveles de significancia simulados en las Tablas 3a y 3b concuerdan con los reportados anteriormente para múltiples muestras con diseños balanceados. Por lo tanto, los diseños no balanceados parecieran no afectar el rendimiento de la prueba de comparación múltiple. Además, cuando el tamaño de la muestra mínimo es por lo menos 20, entonces los niveles de significancia simulados se aproximan al nivel objetivo, excepto para los datos contaminados.

En conclusión, cuando la muestra más pequeña es de por lo menos 20, la prueba de comparación múltiple funciona adecuadamente con múltiples (k) muestras en diseños tanto balanceados como no balanceados. Para muestras más pequeñas, sin embargo, la prueba es conservadora cuando los datos son simétricos o casi simétricos y es liberal cuando los datos son altamente asimétricos.

Apéndice C: Función de potencia teórica

La función de potencia teórica exacta de la prueba de comparación múltiple no está disponible. Sin embargo, para los diseños de 2 muestras, se puede obtener una función de potencia aproximada basada en métodos teóricos para muestras grandes. En el caso de diseños con múltiples muestras, se requiere una investigación más profunda para derivar una aproximación similar.

Para los diseños de 2 muestras, sin embargo, se puede obtener una función de potencia teórica de la prueba de Bonett utilizando métodos teóricos basados en muestras grandes. Más específicamente, el estadístico de la prueba, T, proporcionado abajo se encuentra asintóticamente distribuido como una distribución de chi-cuadrado con 1 grado de libertad:

$$T = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

En esta expresión de T, $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$, $g_i = (n_i - 3)/n_i$ y γ se encuentra la curtosis común desconocida de las dos poblaciones.

Por lo tanto, la función de potencia teórica de una prueba de igualdad de varianzas bilateral de Bonett con un nivel de significancia aproximado de α pudiera expresarse como

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right) + \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

donde

$$se = \sqrt{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

Para las pruebas unilaterales, la función de potencia aproximada cuando se realizan pruebas con respecto a $\sigma_1 > \sigma_2$ es

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

y cuando se realizan pruebas con respecto a $\sigma_1 < \sigma_2$, la función de potencia aproximada, es

$$\pi(n_1, n_2, \rho) = \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

Tenga en cuenta que durante la fase de planificación del tamaño de la muestra del análisis de los datos, la curtosis común de las poblaciones, γ, es desconocida. Por lo tanto, el investigador debe fiarse de las opiniones de los expertos o los experimentos anteriores para obtener un valor de planificación para γ. Si no se dispone de esa información, con frecuencia es aconsejable realizar un pequeño estudio piloto con el fin de desarrollar los planes para un estudio más exhaustivo. Utilizando las muestras del estudio piloto, se obtiene un valor de planificación de γ como la curtosis agrupada expresada por

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

En el menú Asistente, la estimación de planificación de γ se obtiene retrospectivamente con base en los datos de los que dispone el usuario.

Apéndice D: Comparación de potencia teórica y simulada

Simulación D1: Potencia simulada (real) de la prueba de Bonett

Realizamos una simulación para comparar los niveles de potencia simulada de la prueba de Bonett con los niveles de potencia basados en la función de potencia aproximada derivada en el Apéndice C.

Generamos 10,000 pares de muestras para cada una de las distribuciones anteriormente descritas (véase Simulación B1). En general, los tamaños de muestra seleccionados eran suficientemente grandes para el nivel de significancia simulado de la prueba para aproximarse razonablemente al nivel de significancia objetivo, con base en nuestros resultados anteriores de la Simulación B1.

Para evaluar los niveles de potencia simulada en una relación de desviaciones estándar $\rho = \sigma_1/\sigma_2 = 1/2$, multiplicamos la segunda muestra en cada par de muestras por la constante 2. Como resultado, para una distribución específica y para los tamaños de muestra específicos n_1 y n_2, el nivel de potencia simulada se calculó como la fracción de 10,000 pares de muestras replicadas para las que la prueba de Bonett bilateral era significativa. El nivel de significancia objetivo de la prueba se fijó en $\alpha = 0.05$. Con fines comparativos, calculamos los niveles de potencia teórica correspondientes con base en la función de potencia aproximada derivada en el Apéndice C.

Los resultados se muestran abajo en la Tabla 4.

Tabla 4 Comparación de niveles de potencia simulada con niveles de potencia aproximada de una prueba de Bonett bilateral. El nivel de significancia objetivo es 0.05.

Distribución	n_1, n_2	Evaluador Potencia	Simulado Potencia	Distribución	n_1, n_2	Evaluador Potencia	Simulado Potencia
N(0,1)	20, 10	0.627	0.527	Exp	20, 10	0.222	0.227
	20, 20	0.830	0.765		20, 20	0.322	0.368
	20, 30	0.896	0.846		20, 30	0.377	0.434
	20, 40	0.925	0.886		20, 40	0.412	0.475
	30, 15	0.825	0.771		30, 15	0.320	0.307
	30, 30	0.954	0.925		30, 30	0.458	0.500

Distribución	n_1, n_2	Evaluador Potencia	Simulado Potencia	Distribución	n_1, n_2	Evaluador Potencia	Simulado Potencia
	30, 45	0.980	0.970		30, 45	0.531	0.579
	30, 60	0.989	0.984		30, 60	0.575	0.622
t(5)	20, 10	0.222	0.379	Chi(5)	20, 10	0.355	0.347
	20, 20	0.322	0.569		20, 20	0.517	0.530
	20, 30	0.377	0.637		20, 30	0.597	0.616
	20, 40	0.412	0.690		20, 40	0.644	0.661
	30, 15	0.320	0.545		30, 15	0.513	0.510
	30, 30	0.458	0.733		30, 30	0.701	0.711
	30, 45	0.531	0.795		30, 45	0.781	0.793
	30, 60	0.575	0.828		30, 60	0.823	0.833
t(10)	20, 10	0.476	0.45	Chi(10)	20, 10	0.454	0.414
	20, 20	0.673	0.673		20, 20	0.646	0.631
	20, 30	0.756	0.749		20, 30	0.730	0.717
	20, 40	0.800	0.803		20, 40	0.776	0.771
	30, 15	0.668	0.659		30, 15	0.641	0.618
	30, 30	0.850	0.852		30, 30	0.828	0.819
	30, 45	0.910	0.911		30, 45	0.892	0.882
	30, 60	0.936	0.937		30, 60	0.921	0.912
Lpl	20, 10	0.321	0.330	B(8,1)	20, 10	0.363	0.278
	20, 20	0.469	0.519		20, 20	0.528	0.463
	20, 30	0.545	0.585		20, 30	0.609	0.549
	20, 40	0.590	0.632		20, 40	0.655	0.600
	30, 15	0.466	0.475		30, 15	0.524	0.419
	30, 30	0.647	0.673		30, 30	0.713	0.634
	30, 45	0.729	0.758		30, 45	0.792	0.737

Distribución	n_1, n_2	Evaluador Potencia	Simulado Potencia	Distribución	n_1, n_2	Evaluador Potencia	Simulado Potencia
	30, 60	0.773	0.800		30, 60	0.833	0.777
B(3,3)	20, 10	0.777	0.628	CN(0.9,3)	20, 10	0.238	0.284
	20, 20	0.939	0.869		20, 20	0.346	0.452
	20, 30	0.973	0.936		20, 30	0.405	0.517
	20, 40	0.984	0.964		20, 40	0.442	0.561
	30, 15	0.935	0.871		30, 15	0.343	0.374
	30, 30	0.993	0.980		30, 30	0.491	0.598
	30, 45	0.998	0.995		30, 45	0.567	0.700
	30, 60	0.999	0.999		30, 60	0.612	0.719
U(0,1)	20, 10	0.916	0.740	CN(0.8,3)	20, 10	0.260	0.223
	20, 20	0.992	0.950		20, 20	0.379	0.396
	20, 30	0.998	0.985		20, 30	0.444	0.467
	20, 40	0.999	0.995		20, 40	0.484	0.520
	30, 15	0.991	0.941		30, 15	0.376	0.354
	30, 30	1.0	0.996		30, 30	0.535	0.549
	30, 45	1.0	1.0		30, 45	0.614	0.650
	30, 60	1.0	1.0		30, 60	0.661	0.706

Los resultados demuestran que, en general, los niveles de potencia aproximada y los niveles de potencia simulada son similares. La similitud incrementa cuando aumentan los tamaños de las muestras. Los niveles de potencia aproximada son por lo general ligeramente mayores que los niveles de potencia simulada para las distribuciones simétrica y casi simétrica con colas de ligeras a moderadas. Sin embargo, para las distribuciones simétricas con colas pesadas o para distribuciones altamente asimétricas, éstos son ligeramente menores que los niveles de potencia simulada. La diferencia entre las dos funciones de potencia no es generalmente importante, excepto en el caso donde las muestras provienen de distribuciones t con 5 grados de libertad.

En general, cuando el tamaño de la muestra mínimo es 20, los niveles de potencia aproximada y los niveles de potencia simulada son notablemente similares. Por lo tanto, la planificación de los tamaños de las muestras se puede basar en las funciones de potencia aproximadas.