Método de comparaciones múltiples

UN PROCEDIMIENTO GRÁFICO DE COMPARACIONES MÚLTIPLES PARA VARIAS DESVIACIONES ESTÁNDAR

Senin J. Banga y Gregory D. Fox
18 de junio de 2013

RESUMEN

Se proporciona un nuevo procedimiento gráfico para comparaciones múltiples de k desviaciones estándar. Como prueba de homogeneidad de las varianzas, el nuevo procedimiento tiene propiedades similares con respecto a los errores Tipo I y Tipo como la versión de Brown y Forsythe (1974) de la prueba de Levene (1960), W_{50}. Sin embargo, la representación gráfica asociada con la prueba de comparaciones múltiples proporciona una útil herramienta visual para cribar muestras con diferentes desviaciones estándar.

Términos de índice: homogeneidad de varianzas, prueba de Levene, prueba de Brown-Forsythe, prueba de Layard, comparaciones múltiples

1. Introducción

La modificación de Brown y Forsythe (1974) de la prueba de Levene (1960), conocida comúnmente como prueba W_{50}, es quizás uno de los procedimientos más utilizados para probar la homogeneidad (igualdad) de las varianzas. En parte, la prueba W_{50} es popular porque es robusta y es asintóticamente independiente de la distribución. Comparada con otras pruebas de la homogeneidad de las varianzas, la prueba W_{50} también es fácil de calcular. (Para una comparación de este tipo de pruebas, consulte Conover et al. (1981).) Además, la prueba W_{50} es muy accesible porque está disponible en muchos paquetes de software de herramientas estadísticas como SAS, Minitab, R y JMP.

Sin embargo, para algunas distribuciones, la potencia de la prueba W_{50} puede ser muy baja, particularmente en las muestras pequeñas. Por ejemplo, Pan (1999) indica que para algunas distribuciones, incluyendo la distribución normal, la prueba W_{50} podría no tener suficiente potencia para detectar diferencias entre dos desviaciones estándar, independientemente de la magnitud de las diferencias. No está claro en el análisis de Pan si la misma limitación se aplicaría a diseños de múltiples muestras. Se podría esperar que esta limitación no se aplicara a diseños con más de dos muestras, simplemente porque esos diseños suelen incluir más datos que los diseños de dos muestras. La prueba W_{50} se caracteriza por tener buenas propiedades con muestras grandes (Miller, 1968; Brown y Forsythe, 1974; Conover et al., 1981).

Se ha vuelto una práctica común realizar un procedimiento de comparación simultánea en parejas basado en una corrección de multiplicidad de Bonferroni después de una prueba W_{50} significativa. Sin embargo, como lo señala Pan (1999), es probable que este enfoque falle o produzca resultados engañosos debido a la baja potencia de la prueba W_{50} en los diseños de dos muestras. Usar la corrección de Bonferroni empeora el problema porque es conservadora, sobre todo cuando el número de comparaciones en parejas es grande. En cambio, existen muchos procedimientos de comparaciones múltiples eficaces para comparar las medias siguiendo un ANOVA de un solo factor. Para ver ejemplos, consulte Tukey (1953), Hochberg et al. (1982) y Stoline (1981). Un análisis post-hoc análogo de las comparaciones entre las varianzas de las muestras sería útil.

En este trabajo, proponemos un método gráfico para comparar las varianzas (o desviaciones estándar) de múltiples muestras. El análisis se basa en "intervalos de incertidumbre" para las varianzas que son similares a los intervalos de incertidumbre descritos por Hochberg et al. (1982) para las medias. En primer lugar, un procedimiento de comparaciones múltiples en parejas se basa en la versión modificada por Bonett (2006) de la prueba de Layard (1973) para la igualdad de las varianzas para diseños de dos muestras. La corrección de multiplicidad utilizada en las comparaciones en parejas se basa en una generalización para muestras grandes del método de Tukey-Kramer (Tukey, 1953; Kramer, 1956), propuesta por Nakayama (2009). Los intervalos de incertidumbre, a los que nos referimos como "intervalos de comparaciones múltiples" o "intervalos de CM", se derivan del procedimiento de comparaciones en parejas utilizando el mejor procedimiento aproximado descrito por Hochberg et al. (1982). La prueba de CM resultante rechaza la hipótesis nula si, y solo si, al menos un par de intervalos de CM no se superpone. Los intervalos de CM que no se superponen identifican las muestras que tienen varianzas (o desviaciones estándar) significativamente diferentes.

Realizamos estudios de simulación para evaluar las propiedades de la prueba de CM con muestras pequeñas. Para efectos de comparación, también incluimos la prueba W_{50} en los estudios de simulación.

2. Procedimiento gráfico de comparaciones múltiples

Supongamos que $Y_{i1}, \dots, Y_{in_i}, \dots, Y_{k1}, \dots, Y_{kn_k}$ son k muestras independientes, donde cada una de las muestras es independiente y está distribuida idénticamente con media $E(Y_{il}) = \mu_i$ y varianza $\mathrm{Var}(Y_{il}) = \sigma_i^2 > 0$. Además, supongamos que las muestras provienen de poblaciones con una curtosis común $\gamma = E(Y - \mu)^4/\sigma^4 < \infty$.

Además, supongamos que $\bar{Y}_i$ y S_i son la media y la desviación estándar de la muestra i, respectivamente. Supongamos que m_i es la media recortada de la muestra i con proporción de recorte $1/[2\sqrt{n_i - 4}]$ y supongamos que $\hat{\gamma}_{ij}$ es un estimador agrupado de la curtosis de las muestras (i, j) calculado como

$$\hat{\gamma}_{ij} = (n_i + n_j)\frac{\sum_{l=1}^{n_i}(Y_{il} - m_i)^4 + \sum_{l=1}^{n_j}(Y_{jl} - m_j)^4}{\left[\sum_{l=1}^{n_i}(Y_{il} - \bar{Y}_i)^2 + \sum_{l=1}^{n_j}(Y_{jl} - \bar{Y}_j)^2\right]^2}$$

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

Tenga en cuanta que $\hat{\gamma}_{ij}$ es asintóticamente equivalente al estimador agrupado de la curtosis de Layard (1973) en el que la media de la muestra $\bar{Y}_i$ ha se remplazado por la media recortada m_i. Por lo tanto, $\hat{\gamma}_{ij}$ es un estimador consistente de la curtosis común desconocida γ, siempre y cuando las varianzas de las poblaciones sean iguales. Bonett (2006) propone este estimador en lugar del estimador agrupado de la curtosis de Layard para mejorar el desempeño de la prueba de Layard con muestras pequeñas en problemas de dos muestras. En este trabajo, nos referimos a la versión modificada por Bonett (2006) de la prueba de Layard simplemente como la prueba de Bonett.

Supongamos que hay más de dos grupos o muestras independientes que comparar ($k > 2$). La procedimiento gráfico de comparaciones múltiples que proponemos se deriva de las comparaciones múltiples en parejas que se basan en la prueba de Bonett. Un enfoque alternativo consiste en basar las comparaciones en parejas en la prueba W_{50}. Sin embargo, en los diseños de dos muestras, el desempeño de potencia de la prueba W_{50} resulta problemático para algunas distribuciones, incluyendo la distribución normal (Pan, 1999). Por otra parte, Banga y Fox (2013) indican que los intervalos de confianza para la relación de las varianzas que se basan en la prueba de Bonett por lo general son superiores a los que se basan en la prueba W_{50}.

Dado un par cualquiera (i, j) de muestras, una prueba bilateral de Bonett con nivel de significancia α' rechaza la hipótesis nula de igualdad de varianzas sy, y solo si,

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > z_{\alpha'/2}\sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

donde $z_{\alpha'/2}$ es el punto percentil $\alpha'/2 \times 100$ superior de la distribución normal estándar.

$$k_i = \frac{n_i - 3}{n_i}, k_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha/2}}, c_j = \frac{n_j}{n_j - z_{\alpha/2}}$$

Puesto que hay comparaciones múltiples en parejas, exactamente $k(k-1)/2$ comparaciones, es necesario un ajuste de la multiplicidad. Por ejemplo, si se da un nivel de significancia objetivo general o por familia, α, entonces un enfoque común, conocido como la corrección de Bonferroni, es elegir el nivel de significancia de cada una de las $k(k-1)/2$ comparaciones en parejas, $\alpha' = 2\alpha/(k(k-1))$. Sin embargo, es bien sabido que la corrección de Bonferroni produce procedimientos de comparación en parejas cada vez más conservadores a medida que aumenta el número de muestras que se compara. Un enfoque alternativo y más adecuado es el propuesto por Nakayama (2009) y se basa en una aproximación para muestras grandes del método de Tukey-Kramer (Tukey, 1953; Kramer, 1956). Específicamente, la prueba general de comparaciones múltiples en parejas es significativa si, y solo si, lo siguiente es cierto para algunos pares (i, j) de muestras:

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > \frac{q_{k,\alpha}}{\sqrt{2}}\sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

donde $q_{\alpha,k}$ es el punto α superior del rango de k variables aleatorias normales estándar independientes y distribuidas idénticamente. Es decir, $q_{\alpha,k}$ satisface

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

donde $Z_1, \ldots, Z_k$ son variables aleatorias normales estándar independientes y distribuidas idénticamente. Barnard (1978) proporciona un algoritmo numérico simple basado en una cuadratura gaussiana de 16 puntos para calcular la función de distribución del rango normal.

Como lo sugiere Hochberg et al. (1982), un procedimiento gráfico de comparaciones múltiples que se aproxime al procedimiento de comparaciones múltiples en parejas descrito anteriormente rechazaría la hipótesis nula si, y solo si,

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

donde las V_i se seleccionan para minimizar lo siguiente:

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

donde

$$b_{ij} = \sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

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

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

Se deduce que una prueba de homogeneidad de varianzas basada en este procedimiento aproximado rechaza la hipótesis nula si, y solo si, al menos un par de los intervalos indicados a continuación no se superponen:

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

El procedimiento gráfico de CM consiste en mostrar estos intervalos en una gráfica para identificar visualmente las muestras con intervalos que no se superponen. Además, se puede determinar el valor p de la prueba general de homogeneidad de varianzas (o desviaciones estándar). En la siguiente sección, proporcionamos los algoritmos detallados para calcular el valor p. Pero antes queremos señalar cierta información básica acerca del procedimiento de CM.

OBSERVACIÓN

1. El estimador agrupado de la curtosis, $\hat{\gamma}_{ij}$, basado en el par (i, j) de muestras, podría haberse remplazado por el estimador agrupado de curtosis general, basado en las k muestras. Aunque este enfoque simplifica relativamente los cálculos, resultados de la simulación que no se muestran aquí indican que al usar $\hat{\gamma}_{ij}$, se obtienen mejores resultados.
2. El intervalo correspondiente a la muestra i no es un intervalo de confianza para la desviación estándar de la población original de la muestra. Hochberg et al. (1982) se refieren a dicho intervalo como un "intervalo de incertidumbre". Nosotros nos referimos al mismo como un "intervalo de comparación múltiples" o un "intervalo de CM". Los intervalos de CM solo son útiles para comparar las desviaciones estándar o varianzas de los diseños de múltiples muestras.

3. Los intervalos de CM que se describen en este documento solo pueden utilizarse para comparar más de dos desviaciones estándar. Cuando hay solamente dos muestras, los intervalos de comparación pueden construirse, pero transmiten la misma información que proporcionan los resultados de la prueba. Resulta mucho más informativo construir un intervalo de confianza para la relación de las desviaciones estándar, como el descrito por Banga y Fox (2013) y provisto con el comando Varianza de dos muestras de Minitab.

3. Valor p del método gráfico de comparaciones múltiples

Antes de describir el algoritmo para calcular el valor p del método gráfico de CM, derivaremos el valor p asociado con la modificación de Bonett (2006) de prueba de Layard en diseños de dos muestras. Posteriormente mostraremos cómo aplicar los resultados de los diseños de dos muestras al procedimiento de comparaciones múltiples.

3.1 Valor p en diseños de dos muestras

Como se mencionó anteriormente, el ajuste de Bonett (2006) de la prueba de Layard en diseños de dos muestras rechaza la hipótesis nula de homogeneidad de varianzas si, y solo si,

$$\left|\ln(c_1 S_1^2) - \ln(c_2 S_2^2)\right| > z_{\alpha/2} se$$

o de manera equivalente

$$\left|\ln(c_{\alpha/2}\, S_1^2 / S_2^2)\right| > z_{\alpha/2} se$$

donde

$$se = \sqrt{\frac{\hat{\gamma}_{12} - k_1}{n_1 - 1} + \frac{\hat{\gamma}_{12} - k_2}{n_2 - 1}}$$

$$c_{\alpha/2} = \frac{c_1}{c_2} = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Bonnet introdujo la constante $c_{\alpha/2}$ como un ajuste en muestras pequeñas para mitigar el efecto de las probabilidades de error de colas desiguales en diseños no balanceados de muestras pequeñas. Sin embargo, el efecto de la constante es insignificante en los diseños no balanceados de muestras grandes y la constante no tiene ningún efecto en los diseños balanceados.

Se deduce que, si el diseño es balanceado, entonces el valor p de la prueba bilateral para la homogeneidad de varianzas simplemente se calcula como

$$P = 2\Pr(Z > |Z_0|)$$

donde

$$Z_0 = \frac{\ln(S_1^2) - \ln(S_2^2)}{se}$$

Si el diseño no es balanceado, entonces $P = 2\min(\alpha_L, \alpha_U)$, donde α_L es la solución más pequeña para α en la ecuación,

$$\exp[\ln(c_\alpha S_1^2/S_2^2) - z_\alpha se] = 1 \qquad (1)$$

y α_U es la solución más pequeña para α en la ecuación,

$$\exp[\ln(c_\alpha S_1^2/S_2^2) + z_\alpha se] = 1 \qquad (2)$$

Los algoritmos para hallar α_L y α_U se especifican abajo. Los detalles matemáticos de los algoritmos se presentan en la sección Apéndice.

Supongamos que

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

Supongamos además que

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Las soluciones α_L y α_U se calculan en los siguientes pasos:

Caso 1: $n_1 < n_2$

- Calcular z_m como se proporciona en el resultado anterior y evaluar $L(z_m, n_1, n_2, S_1, S_2)$.
- Si $L(z_m) \leq 0$, entonces hallar la raíz, z_L, de $L(z, n_1, n_2, S_1, S_2)$ en el intervalo, $(-\infty, z_m]$ y calcular $\alpha_L = \Pr(Z > z_L)$.
- Si $L(z_m) > 0$, entonces la función $L(z, n_1, n_2, S_1, S_2)$ no tiene raíz. Establecer $\alpha_L = 0.0$.

Caso 2: $n_1 > n_2$

- Calcular $L(0, n_1, n_2, S_1, S_2) = \ln S_1^2/S_2^2$.
- Si $L(0, n_1, n_2, S_1, S_2) \geq 0$, entonces hallar la raíz, z_o, de $L(z, n_1, n_2, S_1, S_2)$ en el intervalo $[0, n_2)$; de lo contrario, hallar la raíz z_L en el intervalo $(-\infty, 0)$.
- Calcular $\alpha_L = \Pr(Z > z_L)$.

Para calcular α_U, simplemente aplicamos los pasos anteriores usando la función, $L(z, n_2, n_1, S_2, S_1)$, en lugar de la función, $L(z, n_1, n_2, S_1, S_2)$.

3.2 Valor p de las comparaciones múltiples gráficas

Partiendo del supuesto de que hay k $(k > 2)$ muestras en el diseño, supongamos que P_{ij} es el valor p de la prueba asociada con cualquier par (i, j) de muestras. Recordemos que la prueba de

comparaciones múltiples rechaza la hipótesis nula de la homogeneidad de varianzas si, y solo si, al menos un par de los k intervalos de comparación no se superpone. Se deduce que el valor p general asociado con el procedimiento de comparaciones múltiples es

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

Para calcular P_{ij}, ejecutamos el algoritmo de los diseños de dos muestras usando

$$se = V_i + V_j$$

donde V_i es como se definió anteriormente.

Si $n_i \neq n_j$, entonces

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

donde $\alpha_L = \Pr\,(Q > z_L\sqrt{2})$, $\alpha_U = \Pr\,(Q > z_U\sqrt{2})$, z_L es la raíz más pequeña de la función, $L\big(z, n_i, n_j, S_i, S_j\big)$, z_U es la raíz más pequeña de la función, $L\big(z, n_j, n_i, S_j, S_i\big)$, y Q es una variable aleatoria que se definió con anterioridad. Para hallar las cantidades z_L y z_U, se aplica el algoritmo de diseño de dos muestras descrito anteriormente al par (i, j) de muestras.

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

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

4. Estudio de simulación y resultados

Se realizan dos estudios principales de simulación para investigar el desempeño de la prueba de CM en muestras pequeñas como prueba general para la homogeneidad de las varianzas. Todas las simulaciones se realizaron utilizando la Versión 8 del paquete de software Mathematica.

Estudio 1

El primer estudio está diseñado para evaluar y comparar las propiedades de error Tipo I de la prueba de CM y la prueba W_{50}. Comparamos el desempeño de las dos pruebas con muestras provenientes de diversas distribuciones en tres diseños diferentes: un diseño de 3 muestras, un diseño de 4 muestras y un diseño de 6 muestras. En cada diseño, los tamaños de las muestras varían de 10 a 50 en incrementos de 10. Las muestras se extraen de las siguientes distribuciones originales:

- la distribución normal
- distribuciones simétricas de colas livianas, representadas por la distribución uniforme y una distribución Beta con parámetros de (3, 3)
- distribuciones simétricas de colas pesadas, representadas por una distribución t con 5 grados de libertad ($t(5)$) y la distribución de Laplace

- distribuciones asimétricas y de colas pesadas, representadas por la distribución exponencial, una distribución de chi-cuadrado con 1 grado de libertad ($\chi^2(1)$) y una distribución de chi-cuadrado con 5 grados de libertad ($\chi^2(5)$)
- una distribución contaminada CN(0.9, 3) para la cual el 90% de las observaciones se extrae de la distribución normal estándar y el 10% restante se extrae de una población normal con una media de 0 y una desviación estándar de 3.

Cada simulación consiste en 10,000 réplicas de muestra. El nivel α nominal objetivo es 0.05. El error de simulación es aproximadamente 0.002. Los niveles de significancia simulados para cada prueba se indican en la tabla 1.

Tabla 1 Comparación de los niveles de significancia simulados (($\alpha = 0.05$)

Descripción	**Distribución [Curtosis]**	n_i	$k = 3$		$k = 4$		$k = 6$	
			CM	W_{50}	**CM**	W_{50}	**CM**	W_{50}
Normal	Normal [3.0]	10	.038	.033	.038	.031	.036	.029
		20	.039	.038	.040	.038	.041	.033
		30	.043	.041	.044	.038	.046	.039
		40	.046	.043	.046	.041	.048	.041
		50	.046	.046	.046	.044	.052	.047
Simétrica con colas livianas	Uniforme [1.8]	10	.029	.029	.025	.024	.023	.020
		20	.028	.026	.030	.026	.028	.023
		30	.037	.035	.034	.032	.034	.030
		40	.038	.037	.037	.037	.035	.033
		50	.041	.041	.036	.036	.036	.036
	Beta(3, 3) [2.5]	10	.031	.032	.031	.029	.031	.025
		20	.035	.031	.036	.027	.037	.026
		30	.041	.035	.037	.034	.037	.032
		40	.040	.036	.039	.035	.040	.033
		50	.044	.039	.044	.037	.044	.035
Simétrica con colas pesadas	Laplace [6.0]	10	.056	.038	.063	.041	.071	.039
		20	.054	.044	.058	.043	.059	.041
		30	.051	.042	.053	.043	.052	.044
		40	.048	.045	.048	.045	.048	.046
		50	.045	.045	.051	.046	.049	.047

Descripción	Distribución [Curtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			CM	W_{50}	**CM**	W_{50}	**CM**	W_{50}
	$t(5)$ [9.0]	10	.042	.032	.044	.031	.042	.031
		20	.043	.039	.045	.038	.045	.040
		30	.039	.040	.040	.040	.041	.040
		40	.041	.042	.040	.041	.039	.038
		50	.040	.050	.039	.046	.038	.046
Asimétrica con colas pesadas	$\chi^2(5)$ [5.4]	10	.040	.039	.046	.040	.048	.039
		20	.040	.043	.040	.040	.042	.039
		30	.039	.047	.042	.044	.043	.042
		40	.040	.046	.041	.044	.039	.042
		50	.037	.047	.038	.047	.040	.048
	Exponencial [9.0]	10	.063	.051	.073	.049	.076	.048
		20	.051	.049	.053	.048	.057	.046
		30	.042	.048	.046	.051	.049	.049
		40	.034	.050	.038	.046	.037	.049
		50	.033	.045	.037	.047	.038	.046
	$\chi^2(1)$ [15.0]	10	.084	.048	.098	.050	.118	.050
		20	.053	.046	.060	.047	.068	.046
		30	.041	.041	.045	.045	.050	.047
		40	.044	.049	.046	.047	.045	.047
		50	.038	.050	.037	.049	.040	.049
Normal contaminada	CN(0.9, 3) [8.3]	10	.020	.016	.018	.012	.016	.010
		20	.014	.015	.012	.013	.008	.007
		30	.012	.014	.010	.011	.007	.008
		40	.009	.017	.009	.014	.006	.008
		50	.009	.016	.007	.012	.006	.009

Los resultados revelan que ambas pruebas funcionan adecuadamente para la mayoría de las distribuciones. La mayoría de los niveles de significancia simulados está cerca del objetivo de 0.05. Sin embargo, los niveles de significancia simulados de ambas pruebas tienden a ser conservadores (inferiores a 0.05) cuando se extraen muestras pequeñas de distribuciones

normales y simétricas con colas livianas. Para estas distribuciones, los niveles de significancia simulados de la prueba de CM están más cerca del nivel de significancia objetivo que los de la prueba W_{50}.

Cuando se extraen muestras pequeñas de distribuciones de colas pesadas, la prueba W_{50} tiende a ser conservadora y la prueba de CM tiende a ser liberal. La prueba de CM es aún más liberal cuando se extraen muestras pequeñas de distribuciones extremadamente asimétricas. Por ejemplo, cuando se toman muestras con un tamaño 10 de una distribución de chi-cuadrado con 1 grado de libertad, los niveles de significancia simulados para la prueba de CM son 0.084, 0.098 y 0.118 para los diseños de 3, 4 y 6 muestras, respectivamente.

Ambas pruebas están influenciadas por valores atípicos. Los niveles de significancia para la distribución normal contaminada son extremadamente conservadores, incluso cuando los tamaños de las muestras son tan grandes como 50.

Estudio 2

El segundo estudio evalúa y compara las propiedades de error Tipo II (potencia) de los dos procedimientos en un diseño de 4 muestras. Para este estudio empleamos las mismas muestras que usamos para las muestras con un tamaño de 20 y la condición $k = 4$ en el estudio 1. Las observaciones se escalan por un factor de 1, 2, 3 ó 4. Por ejemplo, en la condición denotada como 1:1:4:4, las observaciones de las muestras 1 y 2 son las mismas que se usaron en el estudio 1. Las observaciones de las muestras 3 y 4 se escalan por un factor de 4.

Incluimos la condición 1:1:1:1 para efectos de comparación. Observe que los resultados para esta condición son los mismos que se indicaron en el estudio 1 para las muestras con un tamaño de 20 y $k = 4$. Elegimos las muestras con un tamaño de 20 porque los resultados del estudio 1 sugieren que, para ambas pruebas, las muestras con un tamaño de 20 producen niveles de significancia alcanzados que están cerca del nivel objetivo para la mayoría de las distribuciones.

Los niveles de potencia simulada en estos experimentos se calculan como la proporción de réplicas de muestra que conduce a rechazos de la hipótesis nula de homogeneidad de varianzas.

Los resultados se muestran en la tabla 2.

Tabla 2 Comparación de los niveles de potencia simulados ($\alpha = 0.05$)

Descripción	**Distribución**	**Relación de desviaciones estándar**							
		1:1:1:1		**1:1:2:2**		**1:2:3:4**		**1:1:4:4**	
		CM	W_{50}	**CM**	W_{50}	**CM**	W_{50}	**CM**	W_{50}
	Normal	.040	.038	.846	.853	.998	.994	1.00	1.00
Simétrica con colas livianas	Uniforme	.030	.026	.985	.962	1.00	.999	1.00	1.00
	Beta(3, 3)	.036	.027	.938	.916	1.00	.999	1.00	1.00

Descripción	Distribución	Relación de desviaciones estándar							
		1:1:1:1		1:1:2:2		1:2:3:4		1:1:4:4	
		CM	W_{50}	CM	W_{50}	CM	W_{50}	CM	W_{50}
Simétrica con colas pesadas	Laplace	.058	.043	.597	.629	.931	.921	.996	.998
	$t(5)$	.045	.038	.657	.703	.952	.949	.997	.998
Asimétrica con colas pesadas	$\chi^2(5)$	.040	.040	.625	.704	.949	.949	.996	.999
	Exponencial	.053	.048	.431	.507	.804	.779	.963	.978
	$\chi^2(1)$	.060	.047	.298	.291	.602	.504	.838	.824
Contaminada	CN(0.9, 3)	.012	.013	.499	.612	.889	.917	.989	.998

Los resultados sugieren que las propiedades de error Tipo II (potencia) de la prueba de CM y la prueba W_{50} son similares. En general, los niveles de potencia simulada logrados con ambas pruebas son del mismo orden de magnitud. Solamente en un caso la potencia de las dos pruebas difiere por más de 0.1.

Los niveles de potencia simulada para la prueba de CM son ligeramente mejores que los de la prueba W_{50} cuando las muestras provienen de distribuciones simétricas con colas de livianas a moderadas. Por otro lado, la prueba W_{50} parece ser ligeramente más potente que la prueba de CM cuando las muestras provienen de distribuciones con colas pesadas.

5. Ejemplo

En esta sección, aplicamos el procedimiento gráfico de CM y la prueba W_{50} a un conjunto de datos tomado de Ott et al. (2010), página 397. Los datos se describen de la siguiente manera:

> *Una empresa de fundición tiene varios hornos en los que calienta las materias primas antes de vaciarlas en un molde de cera. Es muy importante que estos metales se calienten a una temperatura precisa con muy poca variación. Se seleccionan tres hornos de forma aleatoria y se registran sus temperaturas (°C) con mucha exactitud en 10 calentamientos sucesivos. Los datos recolectados son los siguientes:*

Horno 1	1670.87	1670.88	1671.51	1672.01	1669.63	1670.95	1668.70	1671.86	1669.12	1672.52
Horno 2	1669.16	1669.60	1669.76	1669.18	1671.92	1669.69	1669.45	1669.35	1671.89	1673.45
Horno 3	1673.08	1672.75	1675.14	1674.94	1671.33	1660.38	1679.94	1660.51	1668.78	1664.32

La figura 1 muestra gráficas de caja de las temperaturas para cada horno. Las gráficas de caja sugieren que no existen valores atípicos en las temperaturas registradas y que la variabilidad de la temperatura para el horno 3 es diferente de la del horno 1 o el horno 2.

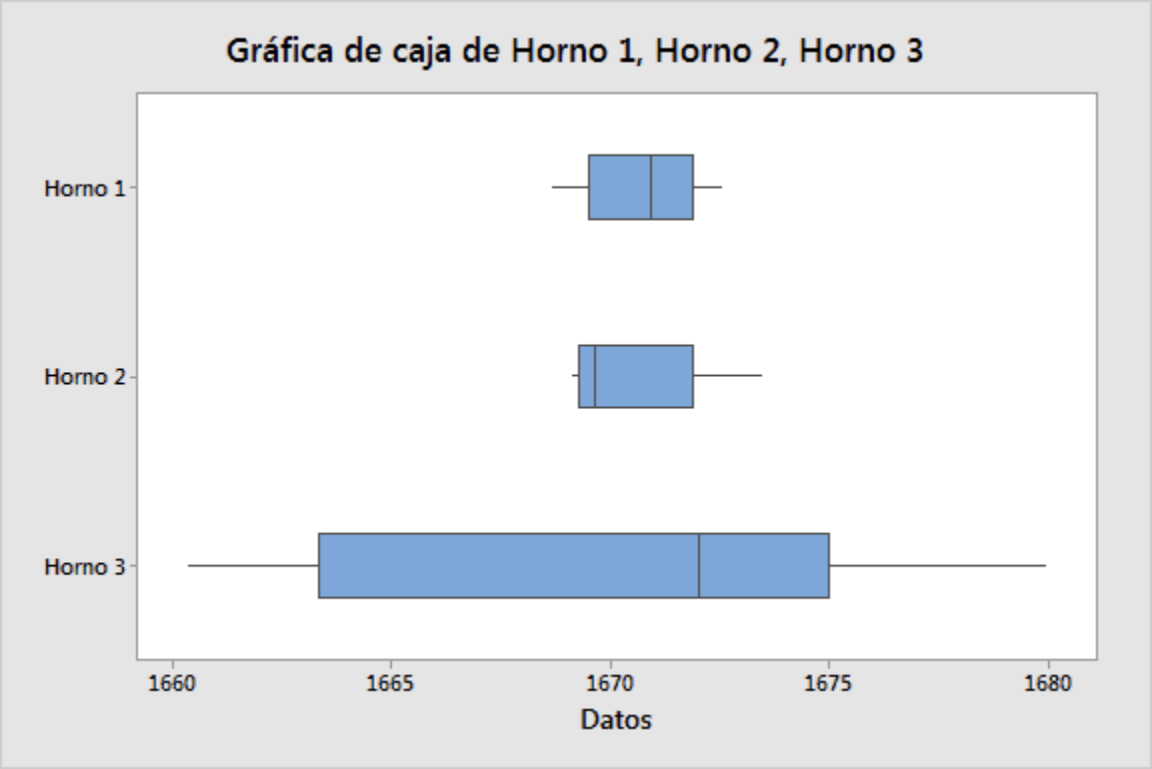

Figura 1 Gráficas de caja de la temperatura del horno (°C)

La figura 2 muestra los intervalos de CM para los mismos datos, así como los resultados de la prueba general de CM y la prueba W_{50}, que se menciona en la leyenda como la prueba de Levene. Los valores p significativos para ambas pruebas indican que la variabilidad de las temperaturas es diferente en los tres hornos. Los intervalos de CM que no se superponen confirman que la variabilidad del horno 3 es diferente de la del horno 2 o el horno 1. Los intervalos de CM son $(0.896, 2.378)$, $(1.072, 2.760)$ y $(4.366, 12.787)$ para los hornos 1, 2 y 3, respectivamente.

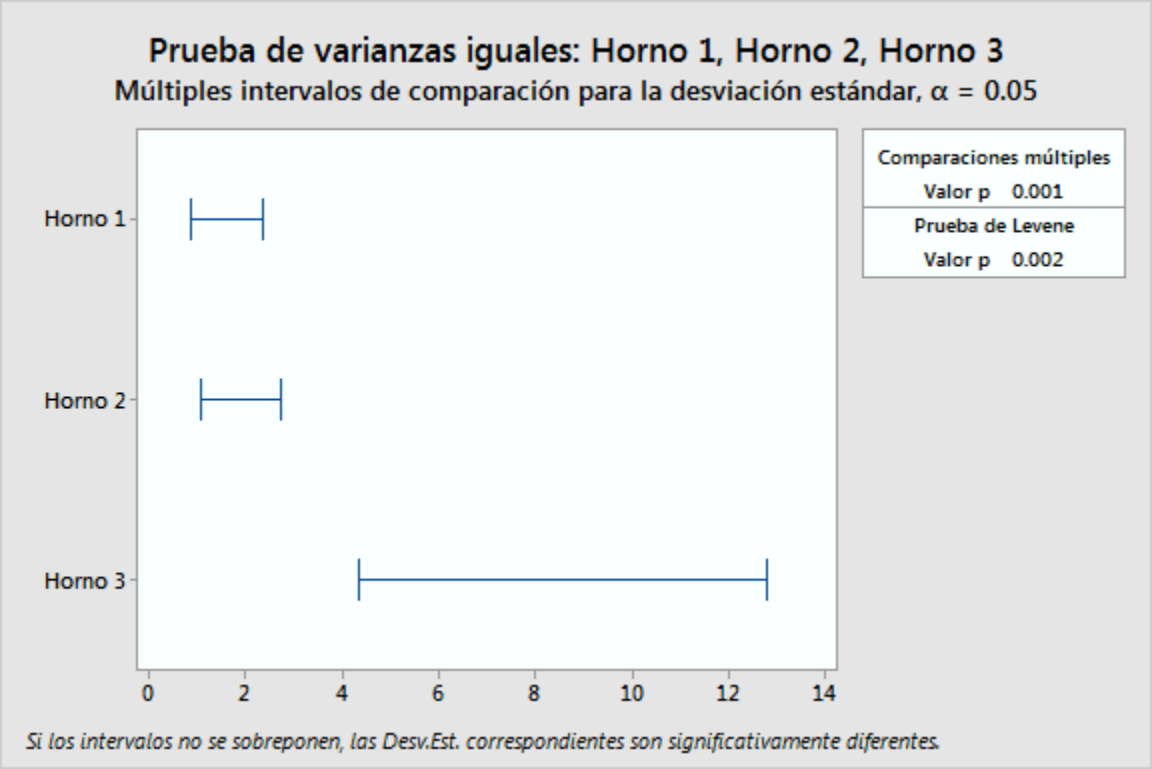

Figura 2 Intervalos de CM y valores p para la prueba de CM y la prueba W_{50} (prueba de Levene)

6. Conclusión

En general, los resultados de la simulación indican que, para los diseños con múltiples muestras pequeñas, el desempeño de la prueba de CM es similar al de la prueba W_{50}. La prueba de CM es ligeramente más adecuada para las distribuciones simétricas o casi simétricas con colas de livianas a moderadas, mientras que la prueba W_{50} podría ser preferible cuando los datos se extraen de distribuciones muy asimétricas y distribuciones con colas pesadas. Una clara ventaja del procedimiento de CM es que proporciona una potente herramienta visual para cribar las muestras con diferentes desviaciones estándar o varianzas cuando la prueba general de la homogeneidad de las desviaciones estándar es significativa. El procedimiento gráfico de CM está disponible en la versión 17 de Minitab.

7. Apéndice

El ajuste de Bonett (2006) de la prueba de Layard en diseños de dos muestras rechaza la hipótesis nula de homogeneidad de varianzas si, y solo si,

$$\left|\ln(c_1 S_1^2) - \ln(c_2 S_2^2)\right| > z_{\alpha/2} se$$

o de manera equivalente

$$\left|\ln(c_{\alpha/2}\, S_1^2 / S_2^2)\right| > z_{\alpha/2} se$$

donde

$$se = \sqrt{\frac{\hat{\gamma}_{12} - k_1}{n_1 - 1} + \frac{\hat{\gamma}_{12} - k_2}{n_2 - 1}}$$

$$c_{\alpha/2} = \frac{c_1}{c_2} = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Por lo tanto, si el diseño es balanceado, entonces $c_{\alpha/2} = 1$, de modo que el valor p de la prueba es simplemente

$$P = 2\Pr(Z > |Z_0|)$$

donde

$$Z_0 = \frac{\ln(S_1^2) - \ln(S_2^2)}{se}$$

Si el diseño no es balanceado, entonces $P = 2\min(\alpha_L, \alpha_U)$ donde

α_L es la solución más pequeña para α en la ecuación,

$\exp[\ln(c_\alpha S_1^2 / S_2^2) - z_\alpha se] = 1$ (1)

y α_U es la solución más pequeña α de la ecuación,

$\exp[\ln(c_\alpha S_1^2 / S_2^2) + z_\alpha se] = 1$ (2)

El enfoque para resolver estas ecuaciones para α es resolver primero las ecuaciones para $z \equiv z_\alpha$ y luego obtener $\alpha = \Pr(Z > z)$ donde la variable aleatoria Z tiene la distribución normal estándar. Antes de describir cómo resolver estas ecuaciones, señalamos que la ecuación (1) puede re-expresarse como la ecuación $L(z) = 0$ donde

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

Del mismo modo, la ecuación (2) es equivalente a la ecuación $U(z) = 0$, donde

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

Observamos que $L(z, n_2, n_1, S_2, S_1) = -U(z, n_1, n_2, S_1, S_2)$. En consecuencia, solo se deben hallar las raíces de una de las dos funciones.

El algoritmo para resolver la ecuación (1) o (2), se deriva del siguiente resultado:

Resultado

Supongamos que n_1, n_2, S_1 y S_2 se especifican y son fijos. Para diseños no balanceados, la función, $L(z, n_1, n_2, S_1, S_2)$, tiene, a lo sumo, dos raíces.

4. *Si $n_1 < n_2$ entonces $L(z, n_1, n_2, S_1, S_2)$ es convexa: satisface $L(-\infty, n_1, n_2, S_1, S_2) = L(n_1, n_1, n_2, S_1, S_2) = +\infty$ y alcanza su mínimo en*

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

   Por lo tanto, if $L(z_m, n_1, n_2, S_1, S_2) \leq 0$, entonces hay dos raíces: una en el intervalo $(-\infty, z_m]$ y otra en el intervalo $[z_m, n_1)$. Por otro lado, if $L(z_m, n_1, n_2, S_1, S_2) > 0$, entonces la función $L(z, n_1, n_2, S_1, S_2)$ no tiene ninguna raíz.

5. Si $n_1 > n_2$, entonces $L(z, n_1, n_2, S_1, S_2)$ decrece monotónicamente de $+\infty$ a $-\infty$ y, por lo tanto, tiene una raíz única. Si $L(0, n_1, n_2, S_1, S_2) = ln\, S_1^2/S_2^2 \geq 0$, entonces la raíz están en el intervalo $[0, n_2)$; de lo contrario, la raíz se encuentra en el intervalo $(-\infty, 0)$.

Comprobación

En lo siguiente, supongamos que $L(z) \equiv L(z, n_1, n_2, S_1, S_2)$.

En primer lugar, queremos demostrar que si $n_1 < n_2$, entonces es convexa y alcanza su mínimo en

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Según se definió anteriormente

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

Entonces, tenemos $\lim_{z \to -\infty} L(z) = +\infty$ y

$$\lim_{z \to \min(n_1, n_2)} L(z) = \begin{cases} +\infty \text{ si } n_1 < n_2 \\ -\infty \text{ si } n_2 < n_1 \end{cases}$$

Además, observe que la derivada de $L(z)$ satisface

$$-\frac{(n_1 - z)(n_2 - z)}{se} L'(z) = z^2 - (n_1 + n_2)z + n_1 n_2 + \frac{n_1 - n_2}{se}$$

Supongamos que

$$Q(z) = -\frac{(n_1 - z)(n_2 - z)}{se} L'(z)$$

Si $n_1 < n_2$, entonces $Q(z)$ cuadrático tiene dos raíces calculadas como

$$z_1 = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

y

$$z_2 = \frac{n_1 + n_2 + \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Puesto que $Q(n_1) = \frac{n_1 - n_2}{se} < 0$, tenemos $z_1 < n_1 = \min(n_1, n_2) < z_2$ de modo que $Q(z) > 0$ para z en $(-\infty, z_1)$ y de modo que $Q(z) < 0$ para z en (z_1, n_1). Se deduce que $L'(z) < 0$ para z en $(-\infty, z_1)$ y que $L'(z) > 0$ para z en (z_1, n_1). Por lo tanto, $L(z)$ es convexa en el dominio $(-\infty, \min(n_1, n_2))$ y alcanza su valor mínimo en $z_1 \equiv z_m$.

Si $n_1 > n_2$, entonces hay dos casos: el caso donde $n_1 - n_2 > 4/se$ y el caso donde

$0 < n_1 - n_2 < 4/se$. En el primer caso, z_1 y z_2 son las raíces de $Q(z)$ de modo que $n_2 = \min(n_1, n_2) < z_1 < z_2$. (Esto es porque $n_2 - \frac{z_1 + z_2}{2} = \frac{n_2 - n_1}{2} < 0$). Por lo tanto, $Q(z) > 0$ para z en el dominio $(-\infty, \min(n_1, n_2))$. En el segundo caso, $Q(z)$ no tiene raíces de modo que $Q(z) > 0$ en el dominio.

Se deduce que si $n_1 > n_2$, entonces $L'(z) < 0$ de modo que $L(z)$ decrece monotónicamente de $+\infty$ a $-\infty$.

8. Referencias

Banga, S. J. y Fox, G. D. (2013). On Bonett's Robust Confidence Interval for a Ratio of Standard Deviations. En proceso de impresión.

Barnard, J.Barnard, J. (1978). Probability Integral of the Normal Range. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 27, 197–198.

Bonett, D. G.G. (2006). Robust Confidence Interval for a Ratio of Standard Deviations. *Applied Psychological Measurements*, 30, 432–439.

Brown, M. B. y Forsythe A. B.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.

Hochberg, Y., Weiss, G. y Hart S. (1982). On Graphical Procedures for Multiple Comparisons. *Journal of the American Statistical Association*, 77, 767–772.

Kramer, C. Y. (1956). Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*, 12, 307–310.

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., *Contributions to Probability and Statistics*, Palo Alto, CA: Stanford University Press, 278–292.

Miller, R. G.G. (1968). Jackknifing Variances. *Annals of Mathematical Statistics*, 39, 567–582.

Nakayama, M. K. (2009). Asymptotically Valid Single-Stage Multiple-Comparison Procedures. *Journal of Statistical Planning and Inference*, 139, 1348–1356.

Ott, R. L. y Longnecker, M. (2010). *An introduction to Statistical Methods and Data Analysis, sixth edition*, Brooks/Cole, Cengage Learning.

Pan, G. (1999). On a Levene Type Test for Equality of Two Variances. *Journal of Statistical Computation and Simulation*, 63, 59–71.

Stoline, M. R. (1981). The Status of Multiple of Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs. *The American Statistician*, 35, 134–141.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons*. Mimeographed monograph.

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