ANOVA de un solo factor

Revisión general

ANOVA de un solo factor se utiliza para comparar las medias de tres o más grupos con el fin de determinar si difieren entre sí de manera significativa. Otra función importante es la estimación de las diferencias entre grupos específicos.

El método más común para detectar diferencias entre grupos en ANOVA de un solo factor es la Prueba F, que se basa en el supuesto de que las poblaciones de todas las muestras comparten una desviación estándar común, aunque desconocida. Reconocimos, en la práctica, que con frecuencia las muestras tienen desviaciones estándar diferentes. Por lo tanto, queríamos investigar el método de Welch, una alternativa a la Prueba F, que puede utilizar variaciones estándar desiguales. También queríamos desarrollar un método para calcular múltiples comparaciones que justificaran las muestras con desviaciones estándar desiguales. Con este método, podemos graficar intervalos individuales, los cuales proporcionan una fácil manera de identificar los grupos que difieren entre sí.

En este informe, describimos cómo desarrollamos los métodos utilizados en el procedimiento ANOVA de un solo factor del Asistente de Minitab para:

- Prueba de Welch
- Intervalos de múltiples comparaciones

Adicionalmente, examinamos las condiciones que pueden afectar la validez de los resultados del ANOVA de un solo factor, incluida la presencia de datos poco comunes, el tamaño de la muestra y la potencia de la prueba, así como la normalidad de los datos. Con base en en estas condiciones, el Asistente realiza automáticamente las siguientes verificaciones en sus datos y notifica los resultados en la Tarjeta de informe:

- Datos poco comunes
- Tamaño de la muestra

- Normalidad de los datos

En este informe, investigamos cómo se relacionan estas condiciones con el ANOVA de un solo factor en la práctica y describimos cómo establecimos las directrices para comprobar estas condiciones en el Asistente.

Métodos para ANOVA de un solo factor

La Prueba F contra la Prueba de Welch

La Prueba F comúnmente utilizada en el ANOVA de un solo factor se basa en el supuesto de que todos los grupos comparten una desviación estándar (σ) común, aunque desconocida. En la práctica, este supuesto rara vez se cumple, lo cual produce problemas para controlar la tasa de error Tipo I. Error Tipo I es la probabilidad de rechazar incorrectamente la hipótesis nula (concluir que las muestras son significativamente diferentes entre sí cuando no es así). Cuando las muestras tienen desviaciones estándar diferentes, existe mayor probabilidad de que la prueba produzca una conclusión incorrecta. Para solucionar este problema, la prueba de Welch se desarrolló como una alternativa a la Prueba F (Welch, 1951).

Objetivo

Queríamos determinar si utilizaríamos la Prueba F o la Prueba de Welch para el procedimiento ANOVA de un solo factor en el Asistente. Para este fin, necesitábamos evaluar en qué grado coincidían los resultados reales de las pruebas F y de Welch con el nivel de significancia objetivo (alfa o tasa de error Tipo I) de la prueba; es decir, si la prueba rechazará incorrectamente la hipótesis nula con mayor o menor frecuencia que la prevista considerando diferentes tamaños de muestras y desviaciones estándar de la muestra.

Método

Para comparar la Prueba F con la Prueba de Welch, realizamos múltiples simulaciones con diferentes números de muestras, tamaños de muestras y desviaciones estándar de las muestras. Para cada condición, realizamos 10,000 pruebas ANOVA utilizando la Prueba F y el método de Welch. Generamos datos aleatorios para que las medias de las muestras fueran iguales y por lo tanto, para cada prueba, la hipótesis nula fuera verdadera. A continuación, realizamos las pruebas utilizando los niveles de significancia objetivo de 0.05 y 0.01. Contamos el número de veces, en 10,000 pruebas, que las pruebas F y de Welch efectivamente rechazaron la hipótesis nula y comparamos esta proporción con el nivel de significancia objetivo. Si la prueba funciona adecuadamente, el error Tipo I estimado debería estar muy cerca del nivel de significancia objetivo.

Resultados

Encontramos que el método de Welch resultó tan bien como, o quizás mejor que, la Prueba F bajo todas las condiciones que probamos. Por ejemplo, cuando comparamos 5 muestras utilizando la prueba de Welch, las tasas de error Tipo I se encontraron entre 0.0460 y 0.0540, muy cerca del nivel de significancia objetivo de 0.05. Ello indica que la tasa de error Tipo I del

método de Welch coincide con el valor objetivo, incluso cuando el tamaño de la muestra y la desviación estándar varían entre las muestras.

No obstante, las tasas de error Tipo I de la prueba F se encontraron entre 0.0273 y 0.2277. En particular, la prueba F tuvo un bajo rendimiento bajo las condiciones siguientes:

- Las tasas de error Tipo I eran inferiores a 0.05 cuando la muestra de mayor tamaño también tenía la desviación estándar más elevada. Esta condición da lugar a una prueba más conservadora y demuestra que el simple hecho de aumentar el tamaño de la muestra no constituye una solución viable cuando las desviaciones estándar de las muestras no son iguales.
- Las tasas de error Tipo I eran superiores a 0.05 cuando los tamaños de las muestras eran iguales, pero las desviaciones estándar eran diferentes. Las tasas también eran mayores que 0.05 cuando la muestra con una desviación estándar considerable era de menor tamaño que las demás muestras. En particular, cuando muestras de menor tamaño tienen desviaciones estándar de mayor tamaño, aumenta sustancialmente el riesgo de que esta prueba rechace incorrectamente la hipótesis nula.

Para obtener más información sobre la metodología y los resultados de la simulación, consulte el Apéndice A.

Debido a que el método de Welch funcionó correctamente cuando las desviaciones estándar y los tamaños de las muestras no eran iguales, utilizamos el método de Welch para el procedimiento ANOVA de un solo factor en el Asistente.

Intervalos de comparación

Cuando una prueba de ANOVA es estadísticamente significativa, lo cual indica que por lo menos una de las medias de la muestra 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 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. Si bien existe un método publicado para las comparaciones múltiples con desviaciones estándar iguales, necesitábamos extender este método para entender las muestras con desviaciones estándar desiguales.

Objetivo

Nuestra intención era desarrollar un método para calcular intervalos de comparación individuales que se pudieran utilizar para realizar comparaciones entre las muestras que, además, coincidieran en la mayor medida posible con los resultados de las pruebas. También queríamos proporcionar un método visual para determinar cuáles muestras eran estadísticamente diferentes de las demás.

Método

Los métodos de múltiples comparaciones convencionales (Hsu 1996) proporcionan un intervalo para la diferencia entre cada par de medias mientras se controla el error aumentado que ocurre cuando se realizan múltiples comparaciones. En el especial caso de iguales tamaños de muestra y bajo el supuesto de desviaciones estándar iguales, es posible mostrar intervalos individuales para cada media de modo tal que corresponda exactamente con los intervalos de las diferencias de todos los pares. En el caso de tamaños de muestra desiguales, bajo el supuesto de desviaciones estándar iguales, Hochberg, Weiss y Hart (1982) desarrollaron intervalos individuales que son aproximadamente equivalentes a los intervalos de las diferencias entre pares, con base en el método de múltiples comparaciones de Tukey-Kramer. En el Asistente, aplicamos la misma estrategia del método de múltiples comparaciones de Games-Howell, el cual no parte del supuesto de desviaciones estándar iguales. La estrategia utilizada en el Asistente de la versión 16 de Minitab era similar en concepto, pero no se basaba directamente en la estrategia de Games-Howell. Para mayor información, véase el Apéndice B.

Resultados

El Asistente muestra los intervalos de comparación de la Gráfica de comparación de medias en el informe de resumen de ANOVA de un solo factor. Cuando la prueba de ANOVA es estadísticamente significativa, cualquier intervalo de comparación que no se superponga a por lo menos otro intervalo se marca en rojo. Es posible que la prueba y los intervalos de comparación no concuerden; sin embargo, este resultado es poco común, debido a que ambos métodos tienen la misma probabilidad de rechazar la hipótesis nula cuando es verdadera. Si la prueba de ANOVA es significativa y todos los intervalos se superponen, entonces el par con la menor cantidad de superposición se marca en rojo. Si la prueba de ANOVA no es estadísticamente significativa, entonces ninguno de los intervalos se marca en rojo, incluso si algunos no se superponen.

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.

Objetivo

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

Método

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

Resultados

El Asistente identifica un punto de dato como poco común si supera en 1.5 el rango intercuartil posterior a los cuartiles inferior o superior de la distribución. Los cuartiles inferior y superior son los percentiles 25 y 75 de los datos. El rango intercuartil es la diferencia entre 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.
	Por lo menos un punto de dato es poco común y pudiera tener una fuerte influencia en los resultados.

Tamaño de la muestra

La potencia es una propiedad importante de cualquier prueba de hipótesis, debido a que indica la probabilidad de que usted encuentre un efecto o diferencia significativos cuando uno de tales efectos existe verdaderamente. La potencia es la probabilidad que tiene de rechazar la hipótesis nula a favor de la hipótesis alternativa. Frecuentemente, la manera más fácil de aumentar la potencia de una prueba es aumentar el tamaño de la muestra. En el Asistente, para las pruebas con baja potencia, indicamos qué tan grande necesita ser su muestra para poder hallar la diferencia especificada. Si no se especifica diferencia alguna, notificamos la diferencia que podría detectar con la potencia adecuada. Para proporcionar esta información, era necesario desarrollar un método para calcular la potencia, debido a que el Asistente utiliza el método de Welch, el cual no tiene una fórmula de potencia exacta.

Objetivo

Para desarrollar una metodología para calcular la potencia, era necesario considerar dos preguntas. En primer lugar, el Asistente no requiere que los usuarios ingresen un conjunto completo de medias; solo requiere que ingresen una diferencia entre las medias que tenga implicaciones prácticas. Para cualquier diferencia, existe un número infinito de posibles configuraciones de medias que podrían producir una diferencia. Por lo tanto, era necesario desarrollar un enfoque razonable para determinar cuáles medias se deben utilizar para calcular la potencia, considerando que no podríamos calcular la potencia para todas las posibles configuraciones de medias. En segundo lugar, era necesario desarrollar un método para calcular la potencia, debido a que el Asistente utiliza el método de Welch, el cual no requiere tamaños de muestras ni desviaciones estándar iguales.

Método

Para cubrir el infinito número de posibles configuraciones de medias, desarrollamos un método basado en el enfoque utilizado en el procedimiento ANOVA de un solo factor estándar en Minitab (**Estadísticas > ANOVA > Un solo factor**). Nos concentramos en los casos donde difieren solo dos de las medias en la cantidad especificada y las otras medias son iguales (establecidas en el promedio ponderado de las medias). Debido a que asumimos que difieren solo dos medias con respecto a la media general (y no más de dos), el enfoque proporciona una estimación de potencia conservador. Sin embargo, debido a que es posible que las muestras tengan diferentes tamaños o desviaciones estándar, el cálculo de la potencia aún depende de las medias que se supone que son diferentes.

Para resolver este problema, identificamos los dos pares de medias que representan los casos mejor y peor. El peor caso ocurre cuando el tamaño de la muestra es pequeño en relación con la varianza de la muestra y se minimiza la potencia; el mejor caso ocurre cuando el tamaño de la muestra es grande en relación con la varianza de la muestra y se maximiza la potencia. Todos los cálculos de potencia consideran estos dos casos extremos, que minimizan y maximizan la potencia bajo el supuesto de que exactamente dos medias difieren del promedio de medias ponderado general.

Para desarrollar el cálculo de la potencia, utilizamos un método demostrado en Kulinskaya et al. (2003). Comparamos los cálculos de potencia de nuestra simulación, el método que desarrollamos para cubrir la configuración de medias y el método demostrado en Kulinskaya et al. (2003). También examinamos otra aproximación de potencia que muestra más claramente cómo la potencia depende de la configuración de medias. Para más información sobre el cálculo de la potencia, véase el Apéndice C.

Resultados

La comparación de estos métodos demostró que el método de Kulinskaya proporciona una buena aproximación de potencia y que nuestro método para trabajar la configuración de medias es apropiado.

Cuando los datos no proporcionan evidencia suficiente contra la hipótesis nula, el Asistente calcula diferencias prácticas que se pueden detectar con una probabilidad del 80% y 90% para los tamaños de las muestras dados. Adicionalmente, si especifica una diferencia práctica, el Asistente calcula los valores de potencia mínimo y máximo para esta diferencia. Cuando los valores de potencia son inferiores a 90%, el Asistente calcula un tamaño de la muestra con base en la diferencia especificada y las desviaciones estándar observadas de las muestras. Para asegurar que el tamaño de la muestra ofrezca los valores de potencia mínimo y máximo de 90% o más, asumimos que la diferencia especificada se encuentra entre las dos medias que tienen la mayor variabilidad.

Si el usuario no especifica una diferencia, el Asistente busca la diferencia más elevada en la que el máximo del rango de valores de potencia es 60%. Este valor se etiqueta en el límite entre las barras roja y amarilla del Informe de potencia, que corresponde a la potencia de 60%. También buscamos la menor diferencia en la cual el mínimo del rango de valores de potencia es 90%. Este valor se etiqueta en el límite entre las barras amarilla y verde en el Informe de potencia, que corresponde a la potencia de 90%.

Cuando se verifican la potencia y el tamaño de la muestra, el Asistente muestra los siguientes indicadores de estado en la Tarjeta de informe.

Estado	Condición
	Los datos no proporcionan evidencia suficiente para concluir que existen diferencias significativas entre las medias. No se especificó diferencia alguna.
	La prueba halla una diferencia entre las medias, de modo que la potencia no representa problema alguno. O La potencia es suficiente. La prueba no halló una diferencia entre las medias, pero la muestra es lo suficientemente grande como para proporcionar por lo menos una probabilidad del 90% de detectar la diferencia especificada.

Estado	Condición
	La potencia pudiera ser suficiente. La prueba no halló una diferencia entre las medias, pero la muestra es lo suficientemente grande para proporcionar un probabilidad entre el 80% y el 90% de detectar la diferencia especificada. Se informa el tamaño de la muestra que se requiere para alcanzar una potencia del 90%.
	La potencia pudiera no ser suficiente. La prueba no halló una diferencia entre las medias, y la muestra es lo suficientemente grande para proporcionar una probabilidad entre el 60% y el 80% de detectar la diferencia especificada. Se informan los tamaños de las muestras que se requieren para alcanzar una potencia del 80% y una potencia del 90%.
	La potencia no es suficiente. La prueba no halló una diferencia entre las medias, y la muestra no es lo suficientemente grande para proporcionar una probabilidad de por lo menos el 60% de detectar la diferencia especificada. Se informan los tamaños de las muestras que se requieren para alcanzar una potencia del 80% y una potencia del 90%.

Normalidad

Un supuesto común de muchos métodos estadísticos es que los datos están normalmente distribuidos. Afortunadamente, incluso cuando los datos no están normalmente distribuidos, los métodos basados en el supuesto de normalidad funcionan correctamente. De cierto modo, esto lo explica el teorema del límite central, el cual sostiene que la distribución de cualquier media de una muestra tiene una distribución aproximadamente normal y que la aproximación se vuelve casi normal a medida que aumenta el tamaño de la muestra.

Objetivo

Nuestro objetivo era determinar qué tan grande necesita ser la muestra para ofrecer una aproximación razonablemente buena de la distribución normal. Lo que queríamos era examinar la prueba de Welch y los intervalos de comparación con muestras de tamaño de pequeño a moderado con diversas distribuciones no normales. Nuestra intención era determinar en qué grado coincidían los resultados reales de las pruebas del método de Welch y los intervalos de comparación con el nivel de significacnia elegido (alfa o tasa de error Tipo I) para la prueba; es decir, si la prueba rechazaba la hipótesis nula incorrectamente con mayor o menor frecuencia que la prevista con diferentes tamaños de muestra, números de niveles y distribuciones no normales.

Método

Para estimar el error Tipo I, realizamos múltiples simulaciones, variando el número de muestras, el tamaño de la muestra y la distribución de los datos. Las simulaciones incluían distribuciones asimétricas y con colas pesadas que se desviaban sustancialmente de la distribución normal. El tamaño y la desviación estándar eran constantes en todas las muestras dentro de cada prueba.

Para cada condición, realizamos 10,000 pruebas de ANOVA utilizando el método de Welch y los intervalos de comparación. Generamos datos aleatorios para que las medias de las muestras fueran iguales y por lo tanto, para cada prueba, la hipótesis nula fuera verdadera. A

continuación, realizamos las pruebas utilizando un nivel de significancia objetivo de 0.05. Contamos el número de veces sobre 10,000 que las pruebas en efecto rechazaron la hipótesis nula y comparamos esta proporción con el nivel de significancia objetivo. Para los intervalos de comparación, contamos el número de veces sobre 10,000 que los intervalos indicaron una o más diferencias. Si la prueba funciona correctamente, el error Tipo I debería encontrarse muy cerca del nivel de significancia objetivo.

Resultados

En general, las pruebas y los intervalos de comparación funcionaron correctamente en todas las condiciones con tamaños de muestra pequeños como 10 o 15. Para pruebas con 9 o menos niveles, en casi todos los casos, los resultados se encuentran todos dentro de 3 puntos porcentuales del nivel de significancia para un tamaño de la muestra de 10 y dentro de 2 puntos porcentuales para un tamaño de la muestra de 15. Para pruebas con 10 o más niveles, en la mayoría de los casos, los resultados se encuentran todos dentro de 3 puntos porcentuales con un tamaño de la muestra de 15 y dentro de 2 puntos porcentuales con un tamaño de la muestra de 20. Para más información, véase Apéndice D.

Debido a que las pruebas funcionaron correctamente con muestras relativamente pequeñas, el Asistente no prueba la normalidad de los datos. En cambio, el Asistente verifica el tamaño de las muestras e indica cuándo las muestras son menores que 15 para niveles entre 2 y 9 y menores que 20 para niveles entre 10 y 12. Con base en estos resultados, el Asistente muestra los siguientes indicadores de estado en la Tarjeta de informe:

Estado	Condición
	Los tamaños de muestra son por lo menos 15 o 20, de modo que la normalidad no representa problema alguno.
	Debido que algunos tamaños de muestra son menores que 15 o 20, la normalidad pudiera ser un problema.

Referencias

Dunnet, C. W. (1980). Pairwise Multiple Comparisons in the Unequal Variance Case. *Journal of the American Statistical Association, 75, 796-800*.

Hoaglin, D. C., Iglewicz, B. y Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association, 81, 991-999*.

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

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. Boca Raton, FL: Chapman & Hall.

Kulinskaya, E., Staudte, R. G., y Gao, H. (2003). Power approximations in testing for unequal means in a One-Way ANOVA weighted for unequal variances, *Communication in Statistics, 32 (12), 2353-2371*.

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

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

Apéndice A: La Prueba F versus la Prueba de Welch

La Prueba F puede aumentar la tasa de error Tipo I cuando se viola el supuesto de igualdad de las desviaciones estándar; la prueba de Welch está diseñada para evitar estos problemas.

Prueba de Welch

Se observan muestras aleatorias de tamaños n_1, ..., n_k de k poblaciones. Supongamos que $\mu_1,\ldots,\mu_k$ denota las medias de la población y que $\sigma_1^2, \ldots, \sigma_k^2$ denota las varianzas de la población. Supongamos que $\bar{x}_1, \ldots, \bar{x}_k$ denota las medias de la muestra y que $s_1^2, \ldots, s_k^2$ denota las varianzas de la muestra. Nos interesaba probar las hipótesis:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

H_1: $\mu_i \neq \mu_j$ para i, j.

La prueba de Welch para la igualdad de k medias compara el estadístico

$$W^* = \frac{\sum_{j=1}^{k} w_j (\bar{x}_j - \hat{\mu})^2/(k-1)}{1+[2(k-2)/(k^2-1)]\sum_{j=1}^{k} h_j}$$

con la distribución de F(k – 1, f), donde

$w_j = \frac{n_j}{s_j^2}$,

$W = \sum_{j=1}^{k} w_j$,

$\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{W}$,

$h_j = \frac{(1- w_j/W)^2}{n_j-1}$ y

$f = \frac{k^2-1}{3\sum_{j=1}^{k} h_j}$.

La prueba de Welch rechaza la hipótesis nula si $W^* \geq F_{k-1,f,1-\alpha}$, el percentil de la distribución de F que se excede con la probabilidad α.

Desviaciones estándar desiguales

En esta sección demostramos la sensibilidad de la prueba F a las violaciones del supuesto de desviaciones estándar iguales y la comparamos con la prueba de Welch.

Los siguientes resultados corresponden a las pruebas de ANOVA de un solo factor utilizando 5 muestras de N(0, σ^2). Cada fila está basada en 10,000 simulaciones utilizando la prueba F y la prueba de Welch. Probamos dos condiciones para la desviación estándar aumentando la

desviación estándar de la quinta muestra, duplicándola y cuadruplicándola en comparación con las otras muestras. Probamos tres diferentes condiciones para el tamaño de la muestra: los tamaños de las muestras son iguales, la quinta muestra es mayor que las otras y la quinta muestra es menor que las otras.

Tabla 1 Las tasas de error Tipo I para las pruebas F simuladas y las pruebas de Welch con 5 muestras y un nivel de significancia objetivo $\alpha = 0.05$

Desviación estándar (σ1, σ2, σ3, σ4, σ5)	Tamaño de la muestra (n1, n2, n3, n4, n5)	Prueba F	Prueba de Welch
1, 1, 1, 1, 2	10, 10, 10, 10, 20	.0273	.0524
1, 1, 1, 1, 2	20, 20, 20, 20, 20	.0678	.0462
1, 1, 1, 1, 2	20, 20, 20, 20, 10	.1258	.0540
1, 1, 1, 1, 4	10, 10, 10, 10, 20	.0312	.0460
1, 1, 1, 1, 4	20, 20, 20, 20, 20	.1065	.0533
1, 1, 1, 1, 4	20, 20, 20, 20, 10	.2277	.0503

Cuando los tamaños de las muestras son iguales (filas 2 y 5), la probabilidad de que la prueba F rechace incorrectamente la hipótesis nula es mayor que el 0.05 objetivo y la probabilidad aumenta a medida que aumenta la desigualdad entre las desviaciones estándar. El problema se acentúa al disminuir el tamaño de la muestra con la mayor desviación estándar. Por otro lado, aumentar el tamaño de la muestra con la mayor desviación estándar reduce la problabilidad de rechazo. Sin embargo, aumentar el tamaño de la muestra en demasía reduce la probabilidad de rechazo en exceso, lo cual no sólo hace que la prueba sea más conservadora de lo necesario cuando se utiliza la hipótesis nula, sino que también afecta negativamente la potencia de la prueba cuando se usa la hipótesis alternativa. Compare estos resultados con la prueba de Welch, la cual coincide en gran medida con el nivel de significancia objetivo de 0.05 en cada caso.

A continuación, realizamos una simulación para casos con k = 7 muestras. Cada fila de la tabla resume las 10,000 pruebas F simuladas. Variamos las desviaciones estándar y los tamaños de las muestras. Los niveles de significancia objetivo son $\alpha = 0.05$ y $\alpha = 0.01$. Tal como se evidencia arriba, observamos desviaciones con respecto a los valores objetivo que pueden ser considerablemente graves. Utilizar un menor tamaño de la muestra cuando la variabilidad es mayor conduce a propabilidades de error Tipo I considerablemente grandes, mientras que utilizar una muestra más grande puede conducir a una prueba extremadamente conservadora. Los resultados se muestran en la Tabla 2, a continuación.

Tabla 2 Tasas de error Tipo I para pruebas F simuladas con 7 muestras

Desviación estándar (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	Tamaños de las muestras (n1, n2, n3, n4, n5, n6, n7)	Objetivo α = 0.05	Objetivo α = 0.01
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	21, 21, 21, 21, 22, 22, 12	0.0795	0.0233
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 21, 21, 21, 21, 24, 12	0.0785	0.0226
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 21, 21, 21, 21, 21, 15	0.0712	0.0199
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 20, 20, 21, 21, 23, 15	0.0719	0.0172
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 20, 20, 20, 21, 21, 18	0.0632	0.0166
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 20, 20, 20, 20, 20, 20	0.0576	0.0138
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	18, 19, 19, 20, 20, 20, 24	0.0474	0.0133
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	18, 18, 18, 18, 18, 18, 32	0.0314	0.0057
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	15, 18, 18, 19, 20, 20, 30	0.0400	0.0085
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	12, 18, 18, 18, 19, 19, 36	0.0288	0.0064
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	15, 15, 15, 15, 15, 15, 50	0.0163	0.0025
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	12, 12, 12, 12, 12, 12, 68	0.0052	0.0002
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	21, 21, 21, 21, 22, 22, 12	0.1097	0.0436
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 21, 21, 21, 21, 24, 12	0.1119	0.0452
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 21, 21, 21, 21, 21, 15	0.0996	0.0376
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 20, 20, 21, 21, 23, 15	0.0657	0.0345
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 20, 20, 20, 21, 21, 18	0.0779	0.0283
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 20, 20, 20, 20, 20, 20	0.0737	0.0264
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	18, 19, 19, 20, 20, 20, 24	0.0604	0.0204
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	18, 18, 18, 18, 18, 18, 32	0.0368	0.0122
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	15, 18, 18, 19, 20, 20, 30	0.0390	0.0117
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	12, 18, 18, 18, 19, 19, 36	0.0232	0.0046
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	15, 15, 15, 15, 15, 15, 50	0.0124	0.0026
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	12, 12, 12, 12, 12, 12, 68	0.0027	0.0004

Desviación estándar (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	Tamaños de las muestras (n1, n2, n3, n4, n5, n6, n7)	Objetivo α = 0.05	Objetivo α = 0.01
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	21, 21, 21, 21, 22, 22, 12	0.134	0.0630
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 21, 21, 21, 21, 24, 12	0.1329	0.0654
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 21, 21, 21, 21, 21, 15	0.1101	0.0484
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 20, 20, 21, 21, 23, 15	0.1121	0.0495
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 20, 20, 20, 21, 21, 18	0.0876	0.0374
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 20, 20, 20, 20, 20, 20	0.0808	0.0317
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	18, 19, 19, 20, 20, 20, 24	0.0606	0.0243
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	18, 18, 18, 18, 18, 18, 32	0.0356	0.0119
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	15, 18, 18, 19, 20, 20, 30	0.0412	0.0134
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	12, 18, 18, 18, 19, 19, 36	0.0261	0.0068
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	15, 15, 15, 15, 15, 15, 50	0.0100	0.0023
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	12, 12, 12, 12, 12, 12, 68	0.0017	0.0003
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	21, 21, 21, 21, 22, 22, 12	0.1773	0.1006
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 21, 21, 21, 21, 24, 12	0.1811	0.1040
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 21, 21, 21, 21, 21, 15	0.1445	0.0760
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 20, 20, 21, 21, 23, 15	0.1448	0.0786
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 20, 20, 20, 21, 21, 18	0.1164	0.0572
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 20, 20, 20, 20, 20, 20	0.1020	0.0503
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	18, 19, 19, 20, 20, 20, 24	0.0834	0.0369
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	18, 18, 18, 18, 18, 18, 32	0.0425	0.0159

Desviación estándar (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	Tamaños de las muestras (n1, n2, n3, n4, n5, n6, n7)	Objetivo α = 0.05	Objetivo α = 0.01
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	15, 18, 18, 19, 20, 20, 30	0.0463	0.0168
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	12, 18, 18, 18, 19, 19, 36	0.0305	0.0103
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	15, 15, 15, 15, 15, 15, 50	0.0082	0.0021
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	12, 12, 12, 12, 12, 12, 68	0.0013	0.0001

Apéndice B: Intervalos de comparación

La gráfica de compración de medias le permite evaluar la significancia estadística de las diferencias entre las medias de las poblaciones.

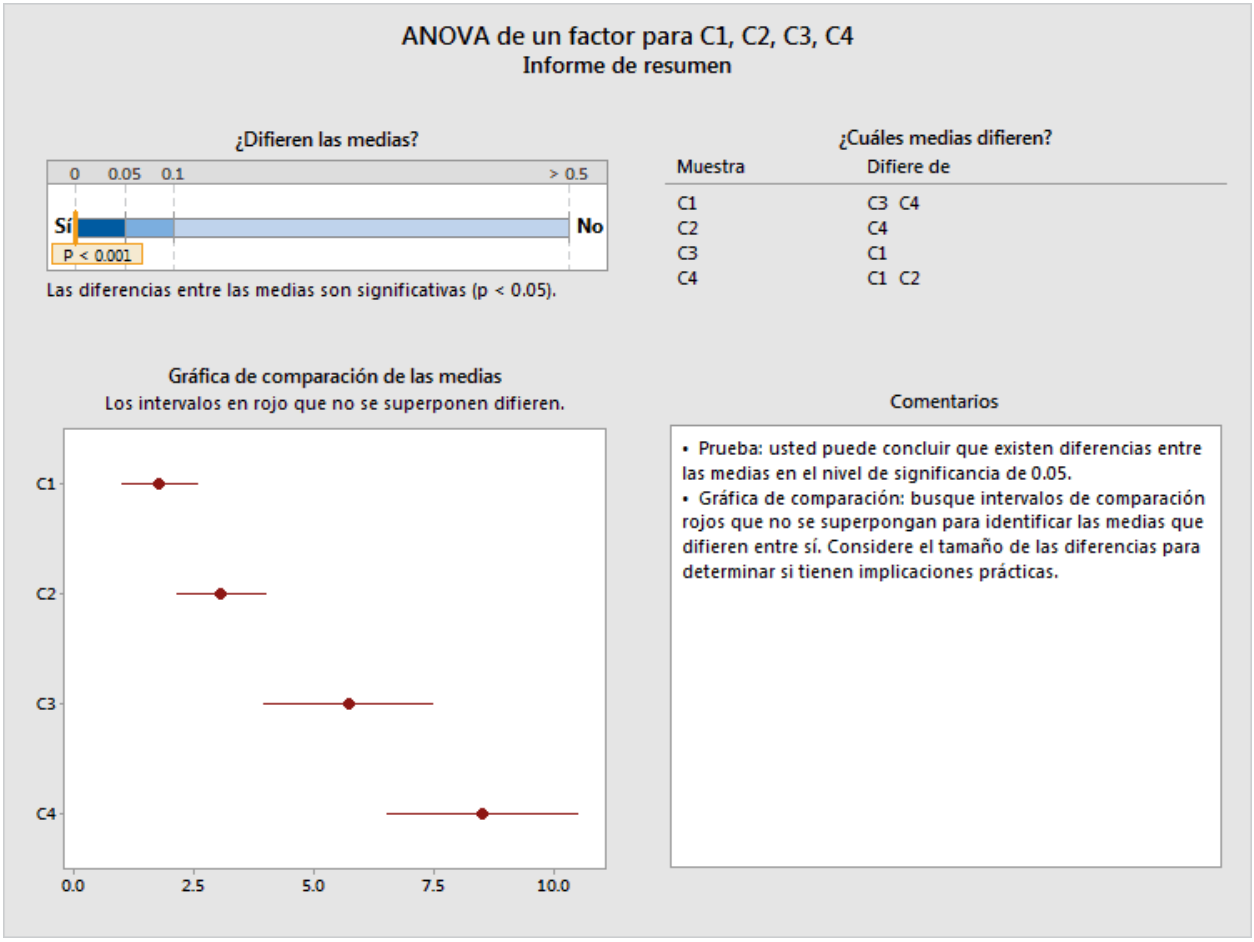
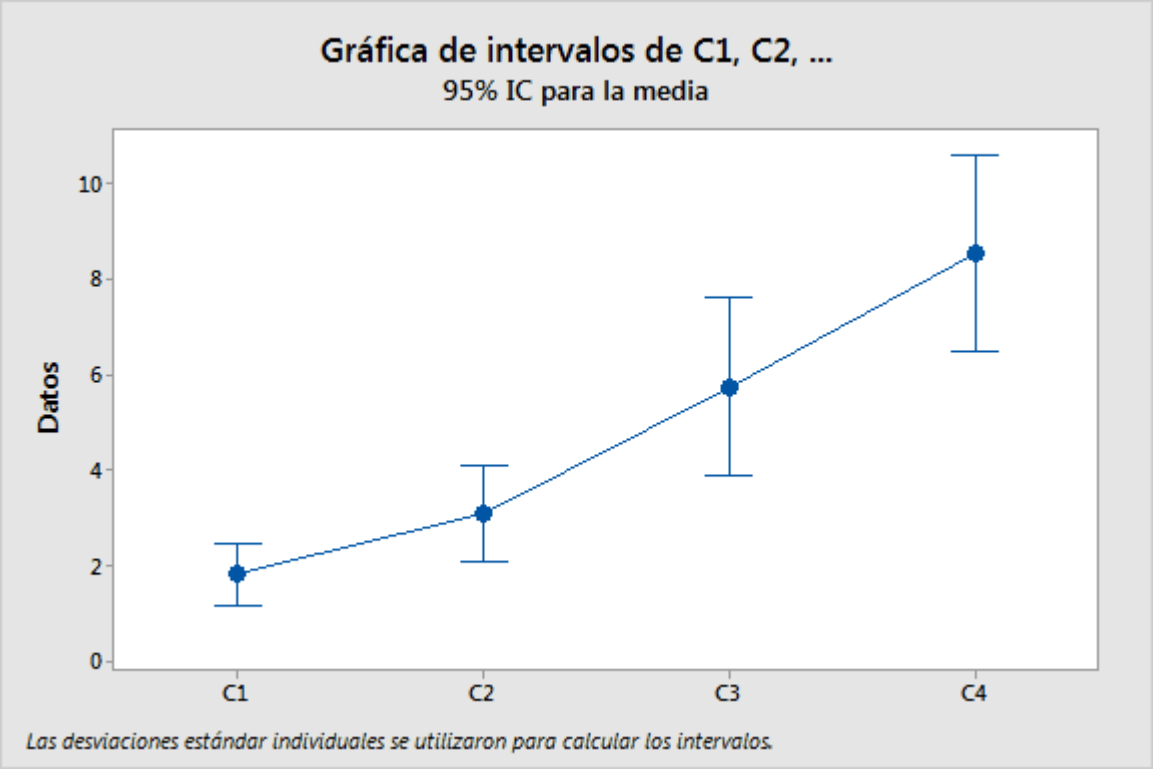


Figura 1 La Gráfica de comparación de medias en el informe de resumen de ANOVA de un solo factor

Un conjunto de intervalos similar aparece en la salida del procedimiento ANOVA de un solo factor estándar en (**Estadísticas** > **ANOVA** > **Un solo factor**):



Sin embargo, tenga en cuenta que los intervalos anteriores son solo intervalos de confianza individuales para las medias. Cuando la prueba de ANOVA (tanto F como de Welch) concluye que algunas medias son diferentes, existe una tendencia natural a observar los intervalos que no se superponen y a sacar conclusiones sobre cuáles medias difieren. Este análisis informal de los intervalos de confianza individuales con frecuencia conducirá a conclusiones razonables; sin embargo, no controla la probabilidad de error de la misma manera que la prueba de ANOVA. Dependiendo del número de poblaciones, los intervalos tienen sustancialmente, en menor o mayor grado, la probabilidad de concluir que existen diferencias. Como resultado, los dos métodos pueden fácilmente llegar a conclusiones discordes. La gráfica de comparación está diseñada para coincidir más uniformemente con los resultados de la prueba de Welch cuando se realizan múltiples comparaciones, aunque no siempre es posible lograr una uniformidad absoluta.

Los métodos de múltiples comparaciones, tales como las comparaciones de Tukey-Kramer y Games-Howell en Minitab (**Estadísticas** > **ANOVA** > **Un solo factor**), le permiten obtener conclusiones estadísticamente válidas sobre las diferencias entre las medias individuales. Estos dos métodos comparan pares, lo cual proporciona un intervalo para la diferencia entre cada par de medias. La probabilidad de que todos los intervalos contengan de manera simultánea las diferencias que estiman es de por lo menos $1 - \alpha$. El método de Tukey-Kramer depende del supuesto de varianzas iguales, mientras que el método de Games-Howell no requiere varianzas

iguales. Si la hipótesis nula de medias iguales es verdadera, entonces todas las diferencias son cero y la probabilidad de que cualquiera de los intervalos de Games-Howell no contendrá cero es cuando mucho α. De modo que podemos utilizar los intervalos para realizar una prueba de hipótesis con un nivel de significancia α. Utilizamos los intervalos de Games-Howell como el punto de partida para obtener los intervalos de la gráfica de comparación en el Asistente.

Con un conjunto de intervalos $[L_{ij}, U_{ij}]$ para todas las diferencias $\mu_i - \mu_j$, $1 \le i < j \le k$, deseamos hallar un conjunto de intervalos $[L_i, U_i]$ para las medias individuales μ_i, $1 \le i \le k$, que transmita la misma información. Esto requiere que cualquier diferencia d se encuentre en el intervalo $[L_{ij}, U_{ij}]$ si y solo si existen $\mu_i \in [L_i, U_i]$ y $\mu_j \in [L_j, U_j]$ de modo tal que $\mu_i - \mu_j = d$. Las cotas de los intervalos deben estar relacionadas con las ecuaciones

$$U_i - L_j = U_{ij} \text{ y}$$
$$L_i - U_j = L_{ij}.$$

Para k = 2, solo tenemos una diferencia, pero dos intervalos individuales, de modo que es posible obtener intervalos de comparación exactos. De hecho, existe cierta flexibilidad en la anchura de los intervalos que satisface esta condición. Para k = 3, existen tres diferencias y tres intervalos individuales, de modo que, una vez más, es posible satisfacer la condición; sin embargo, esta vez sin la flexibilidad de establecer la anchura de los intervalos. Para k = 4, existen seis diferencias, pero solo cuatro intervalos individuales. Los intervalos de comparación deben intentar transmitir la misma información utilizando menos intervalos. En general, para k ≥ 4, existen más diferencias que medias individuales, de modo que no existe una solución exacta a menos que se impongan condiciones adicionales a los intervalos de las diferencias, tales como anchuras iguales.

Los intervalos de Tukey-Kramer tienen anchuras iguales solo si todos los tamaños de las muestras son iguales. Las anchuras iguales también son una consecuencia de partir del supuesto de que las varianzas son iguales. Los intervalos de Games-Howell no parten del supuesto de varianzas iguales, por lo que no tienen anchuras iguales. En el Asistente, tendremos que depender de métodos aproximados para definir los intervalos de comparación.

El intervalo de Games-Howell para $\mu_i - \mu_j$ es

$$\bar{x}_i - \bar{x}_j \pm \left|q^*(k, \hat{v}_{ij})\right| \sqrt{s_i^2/n_i + s_j^2/n_j}$$

donde $q^*(k, \hat{v}_{ij})$ es el percentil apropiado de la distribución del rango studentizado, que depende de k, el número de medias que se están comparando, y así sucesivamente

ν_{ij}, los grados de libertad asociados con el par (i, j):

$$\hat{v}_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\left(\frac{s_i^2}{n_i}\right)^2 \frac{1}{n_i - 1} + \left(\frac{s_j^2}{n_j}\right)^2 \frac{1}{n_j - 1}}.$$

Hochberg, Weiss y Hart (1982) obtuvieron intervalos individuales que son aproximadamente equivalentes a estas comparaciones en pareja utilizando:

$$\bar{x}_i \pm |q^*(k, v)| s_p X_i.$$

Se seleccionan los valores X_i para minimizar

$$\sum \sum_{i \neq j} \left(X_i + X_j - a_{ij}\right)^2,$$

Donde:

$$a_{ij} = \sqrt{1/n_i + 1/n_j}\,.$$

Adaptamos este enfoque al caso de las varianzas desiguales al obtener los intervalos de las comparaciones de Games-Howell con la forma

$$\bar{x}_i \pm d_i.$$

Se seleccionan los valores d_i para minimizar

$$\sum \sum_{i \neq j} \left(d_i + d_j - b_{ij}\right)^2,$$

Donde:

$$b_{ij} = \left|q^*(k, \hat{v}_{ij})\right| \sqrt{s_i^2/n_i + s_j^2/n_j}\,.$$

La solución es

$$d_i = \frac{1}{k-1} \sum_{j \neq i} b_{ij} - \frac{1}{(k-1)(k-2)} \sum_{j \neq i, l \neq i, j < l} b_{jl}.$$

Las gráficas de abajo comparan los resultados de las simulaciones correspondientes a la prueba de Welch con los resultados de los intervalos de comparación utilizando dos métodos: el método basado en Games-Howell que utilizamos ahora y el método utilizado en la versión 16 de Minitab basado en el promedio de los grados de libertad. El eje vertical es la proporción de veces sobre 10,000 simulaciones que la prueba de Welch rechaza incorrectamente la hipótesis nula o que no se superpusieron todos los intervalos de comparación. El alfa objetivo es $\alpha = 0.05$ en estos ejemplos. Estas simulaciones cubren diversos casos de desviaciones estándar y tamaños de muestra desiguales; cada posición sobre el eje horizontal representa un caso diferente.

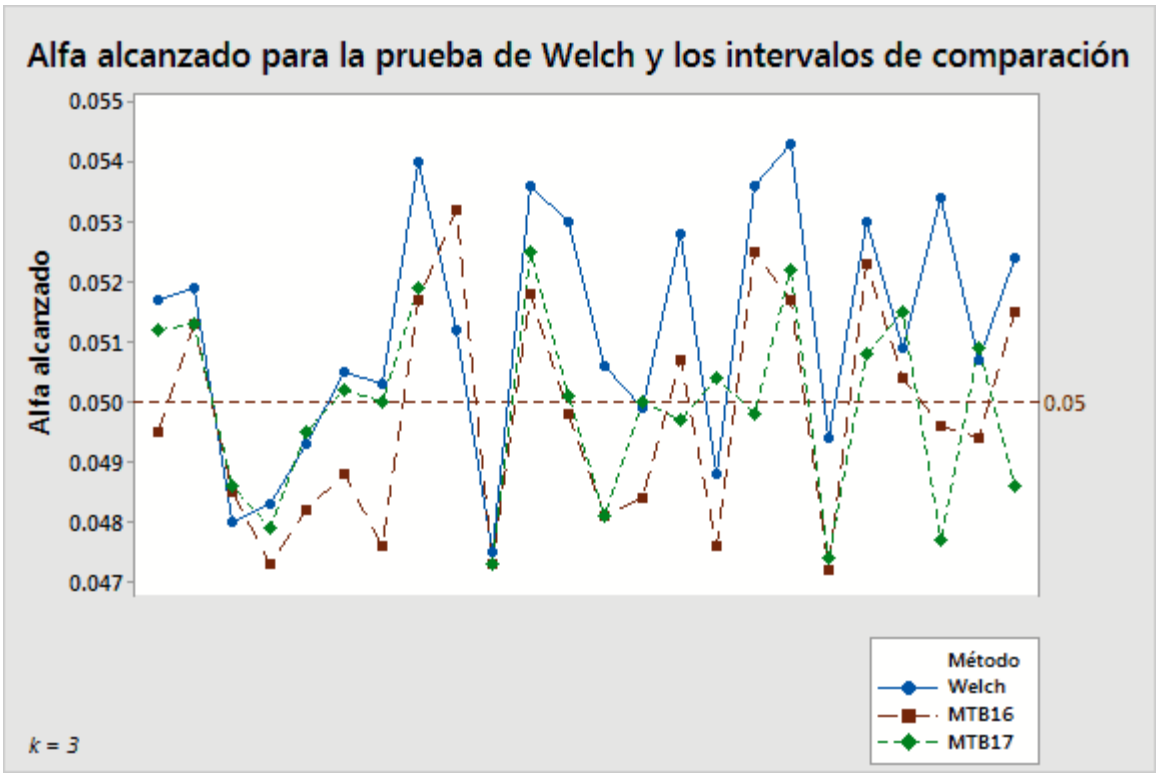


Figura 2 La prueba de Welch comparada con dos métodos utilizados para calcular intervalos de comparación para 3 muestras

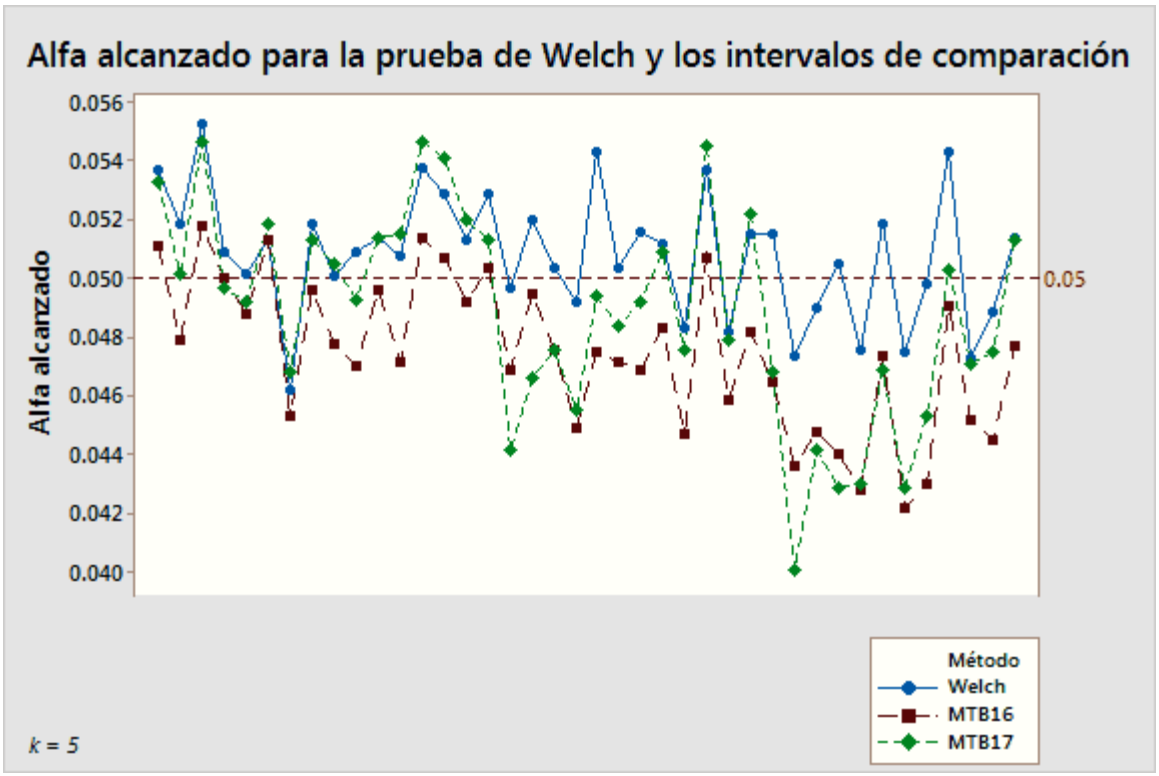


Figura 3 La prueba de Welch comparada con dos métodos utilizados para calcular intervalos de comparación para 5 muestras

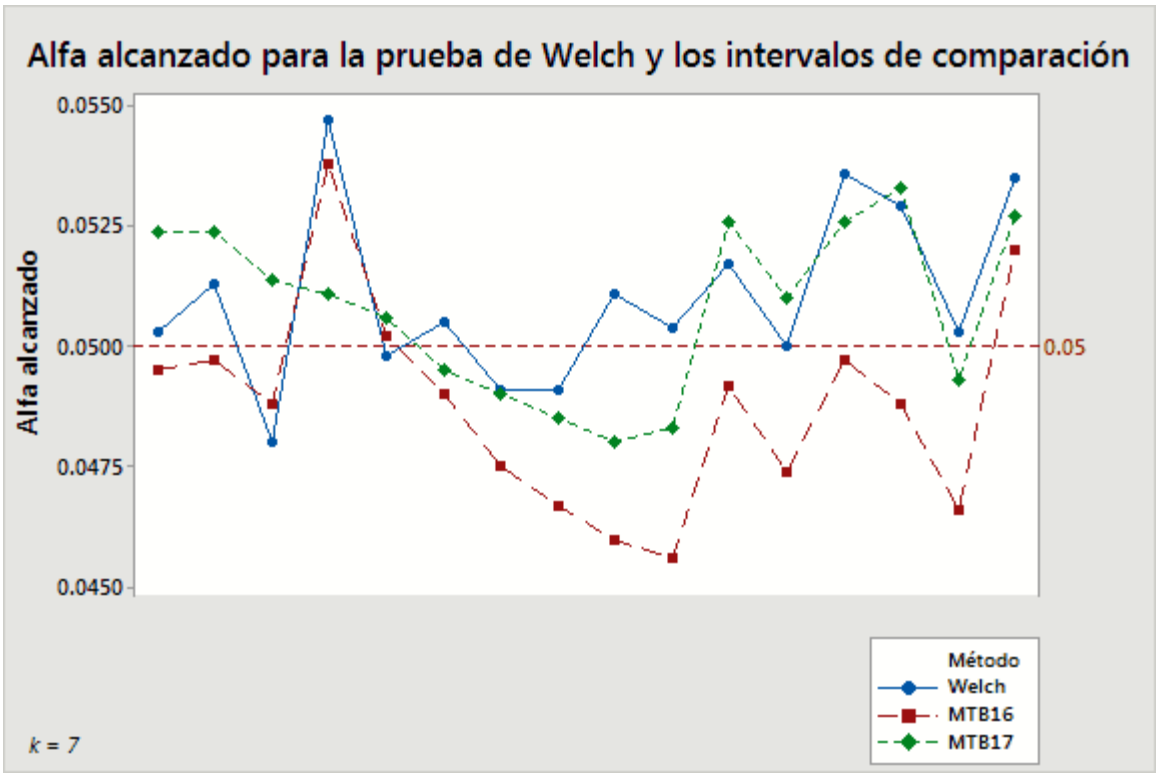


Figura 4 La prueba de Welch comparada con dos métodos utilizados para calcular intervalos de comparación para 7 muestras

Estos resultados muestran valores de alfa simulados en un estrecho rango alrededor del valor objetivo de 0.05. Además, los resultados obtenidos utilizando el método basado en Games-Howell, implementado en la versión 17 de Minitab, se encuentran dudosamente más alineados en mayor medida con los resultados de la prueba de Welch que el método utilizado en la versión 16 de Minitab.

Existe evidencia de que la probabilidad de cobertura de los intervalos pueda ser susceptible de desviaciones estándar desiguales. Sin embargo, la susceptibilidad no es ni siquiera tan extrema como la de la prueba F. La gráfica de abajo ilustra esta dependencia en el caso de k = 5.

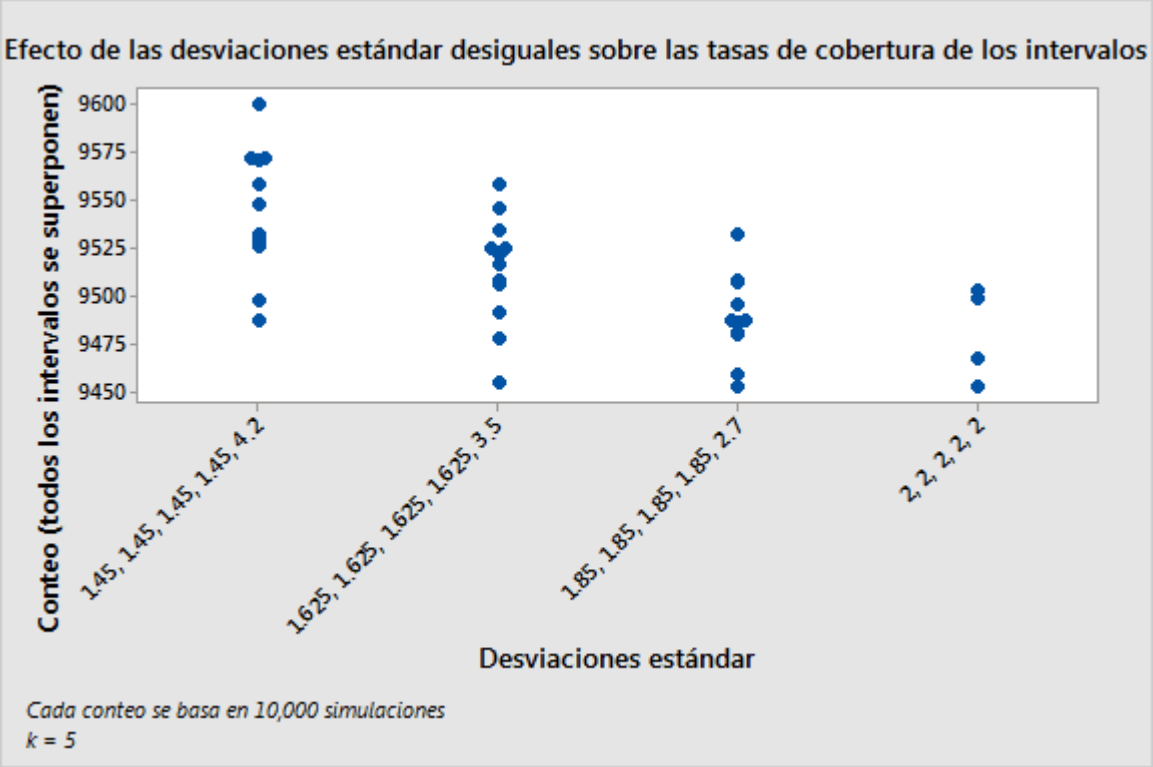


Figura 5 Resultados de la simulación con desviaciones estándar desiguales

Uso de la prueba de hipótesis y los intervalos de comparación en conjunto

En pocos casos, es posible que la prueba de hipótesis y la comparación no coincidan en rechazar la hipótesis nula. La prueba puede rechazar la hipótesis nula mientras todos los intervalos de comparación permanezcan superpuestos. Sin embargo, la prueba pudiera no rechazar la hipótesis nula si hubiera intervalos que no se superpongan. Es posible que la prueba y los intervalos de comparación no concuerden; sin embargo, este resultado es poco común, debido a que ambos métodos tienen la misma probabilidad de rechazar la hipótesis nula cuando es verdadera.

Cuando esto ocurre, primero consideramos los resultados de la prueba y utilizamos las comparaciones para investigar con mayor profundidad en caso de que hubiera una prueba significativa. Si la prueba rechaza la hipótesis nula en el nivel de significancia α, entonces cualquier intervalo de comparación que no se superponga con por lo menos otro se marcará en rojo. Este se utiliza como evidencia visual de que la media del grupo correspondiente difiere de por lo menos otra. Incluso si se superponen todos los intervalos, el par con la menor cantidad

de superposición se marca en rojo si la prueba es lo suficiente significativa para indicar la diferencia "más probable" (véase la Figura 6, abajo). Esta es hasta cierto punto una elección arbitraria, especialmente si hay otros pares con poca superposición. Sin embargo, ningún otro par tiene un límite en su diferencia que se aproxime a cero.

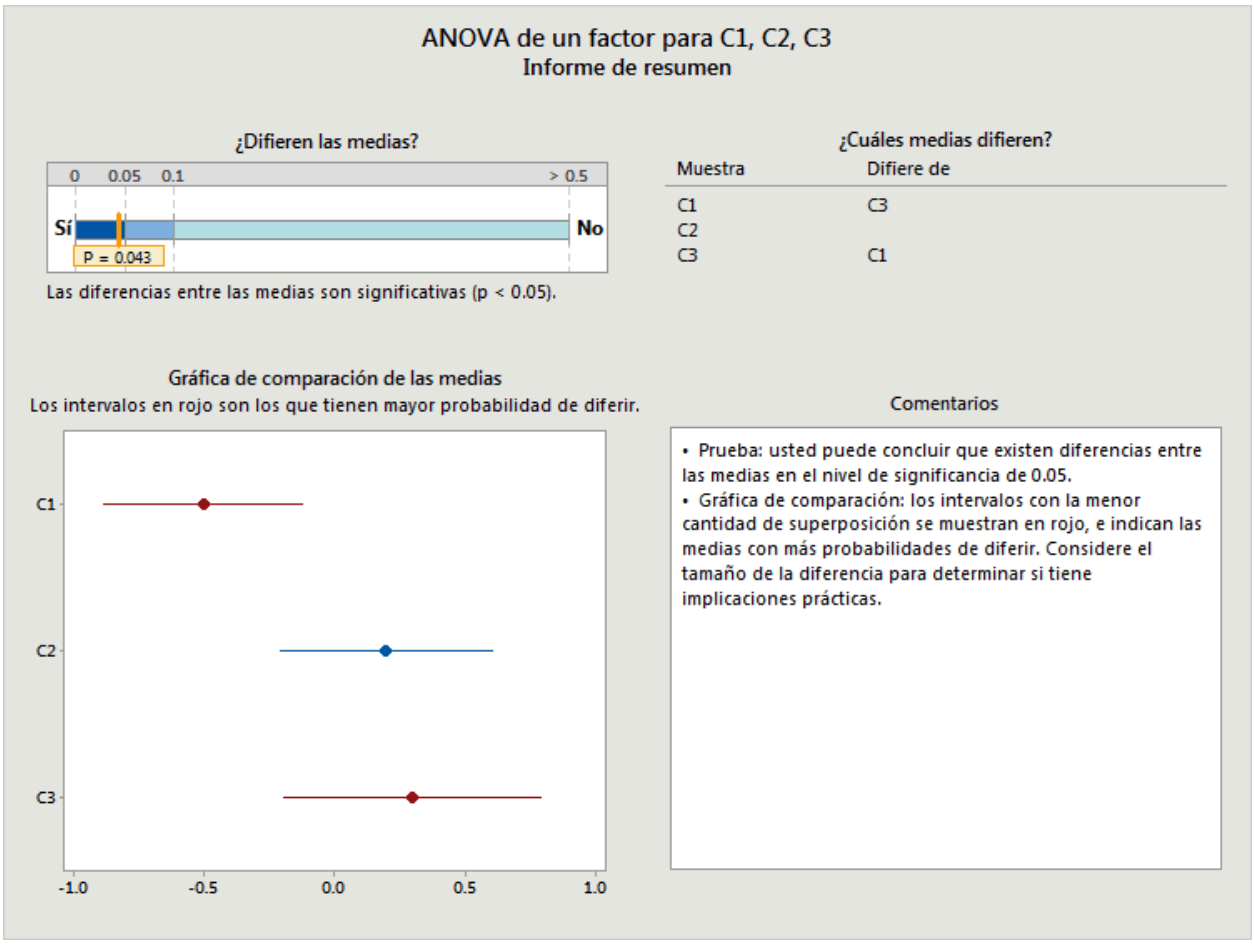


Figura 6 Prueba significativa, intervalos marcados en rojo incluso cuando se superponen entre otras muestras

Si la prueba no rechaza la hipótesis nula, entonces ninguno de los intervalos se marca en rojo, incluso si hay intervalos que no se superponen (véase la Figura 7, abajo). Si bien los intervalos implican que hay diferencias entre las medias, recuerde que no logar rechazar la hipótesis nula no es lo mismo que concluir que la hipótesis nula es verdadera. Solo indica que las diferencias observadas no son lo suficientemente grandes como para descartar la casualidad como la causa. También es importante resaltar que la brecha entre los intervalos que no se superponen será por lo general muy pequeña en esta situación, de modo que las diferencias muy pequeñas siguen concordando con los intervalos, lo que no necesariamente indica que existe una diferencia con implicaciones prácticas.

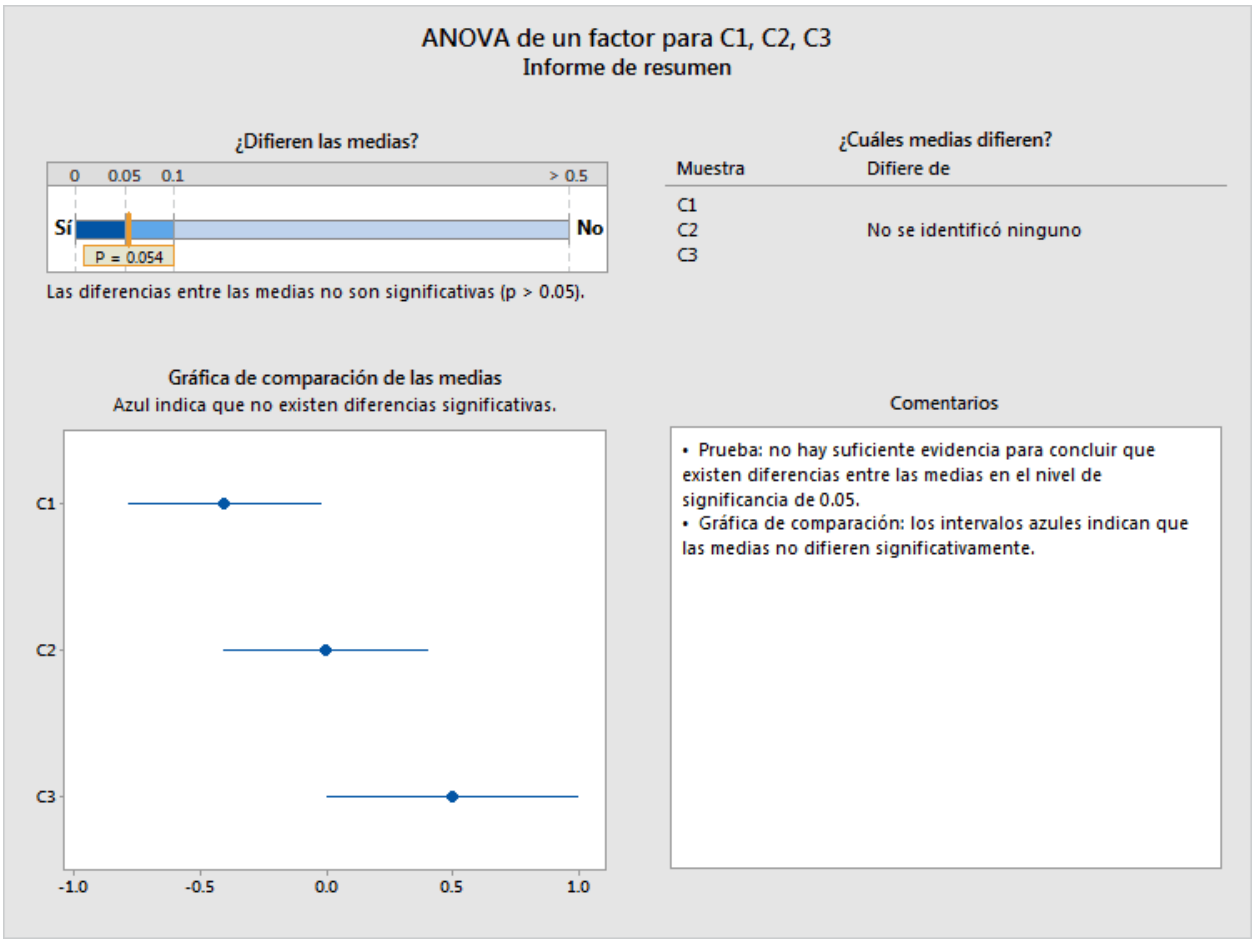


Figura 7 La prueba no pasa, ningún intervalo está marcado en rojo, incluso cuando no existe superposicióp alguna entre las muestras

Apéndice C: Tamaño de la muestra

En ANOVA de un solo factor, los parámetros sometidos a prueba son las medias de población μ_1, μ_2, … μ_k de diferentes grupos o poblaciones. Los parámetros satisfacen la hipótesis nula si son todos iguales. Si hay diferencias entre las medias, estas satisfacen la hipótesis alternativa. La probabilidad de rechazar la hipótesis nula no debe ser mayor que α en el caso de las medias que satisfacen la hipótesis nula. Las probabilidades reales dependen de la desviación estándar de las distribuciones y del tamaño de las muestras. La potencia para detectar cualquier desviación con respecto a la hipótesis nula aumenta con desviaciones estándar más pequeñas o muestras más grandes.

Podemos calcular la potencia de la prueba F bajo el supuesto de distribuciones normales con desviaciones estándar iguales utilizando una distribución F no central. El parámetro de no centralidad es:

$$\theta_F = \sum_{i=1}^{k} n_i \, (\mu_i - \mu)^2 / \sigma^2$$

donde μ es el promedio ponderado de las medias:

$$\mu \; = \; \sum_{i=1}^{k} n_i \mu_i \; / \sum_{i=1}^{k} n_i \; ,$$

y σ es la desviación estándar, que se supone que es constante. Si el resto permanece igual, aumenta la potencia con θF. Este es el sentido preciso en el cual aumenta la potencia a medida que las medias se desvían de la hipótesis nula.

A diferencia de la prueba F, la prueba de Welch no tiene una única fórmula exacta para la potencia. Sin embargo, revisaremos dos fórmulas aproximadas razonablemente acertadas. La primera utiliza una distribución F no central de una manera similar a la potencia de la prueba F. El parámetro de no centralidad que se utilizará mantendrá la forma:

$$\theta_W \; = \; \sum_{i=1}^{k} w_i (\mu_i - \; \mu)^2$$

donde μ es el promedio ponderado:

$$\mu \; = \sum_{i=1}^{k} w_i \mu_i / \sum_{j=1}^{k} w_j$$

sin embargo, las ponderaciones dependerán de las desviaciones estándar y de los tamaños de las muestras; es decir, $w_i = \; n_i / \sigma_i^2$ o $w_i = \; n_i / s_i^2$, dependiendo de si estamos simulando los resultados de las desviaciones estándar conocidas σ_i^2 o estimando la potencia con base en las desviaciones estándar de las muestras s_i^2. A continuación, se calcula la potencia aproximada:

$$P\left(F_{k \, - \, 1, f, \theta_w} \geq \; F_{k \, - \, 1, f, 1 \, - \, \alpha}\right)$$

donde los grados de libertad del denominador son

$$f \; = \; \frac{k^2 - 1}{3 \sum_{i=1}^{k} \left(1 - w_i / \sum_{j=1}^{k} w_j\right) / (n_i - 1)}.$$

Tal como se evidencia abajo, esto proporciona aproximaciones razonablemente acertadas a la potencia observada en las simulaciones. Si bien nosotros utilizamos una aproximación diferente para calcular la potencia en el menú Asistente, esta proporciona información relevante, y es la base para seleccionar la configuración de las medias con la cual configuramos la potencia en el menú Asistente.

Configuración de medias

Con el fin de conservar el enfoque utilizado para la potencia y el tamaño de la muestra en Minitab (**Estadísticas > ANOVA > Un solo factor**), el Asistente no solicita al usuario un conjunto de medias completo con el cual evaluar la potencia. En su lugar, solicita al usuario una diferencia entre las medias que tenga implicaciones prácticas. Para una diferencia dada, existe un número infinito de posibles configuraciones de medias en las que las medias más grandes y pequeñas podrían diferir en esa cantidad. Por ejemplo, todos los siguientes tienen una diferencia máxima de 10 entre un conjunto de cinco medias:

$\mu_1 = 0, \mu_2 = 5, \mu_3 = 5, \mu_4 = 5, \mu_5 = 10;$

$\mu_1 = 5, \mu_2 = 0, \mu_3 = 10, \mu_4 = 10, \mu_5 = 0;$

$\mu_1 = 0, \mu_2 = 10, \mu_3 = 0, \mu_4 = 0, \mu_5 = 0;$

y existen infinitamente muchos más.

Seguimos el enfoque utilizado para la potencia y el tamaño de la muestra de Minitab (Estadísticas > Potencia y tamaño de la muestra > ANOVA de un solo factor), particularmente elegimos un caso donde todas las medias, a excepción de dos, son el promedio (ponderado) de las medias y las dos restantes difieren en la cantidad especificada. Sin embargo, debido a la posibilidad de varianzas y tamaños de muestra desiguales, el parámetro de no centralidad y, por lo tanto, la potencia, aun dependen de cuáles dos medias se supone que son diferentes.

Considere la configuración de las medias $\mu_1, \ldots, \mu_k$ en la cual todas las medias, a excepción de dos, son iguales a la media ponderada general μ, y dos medias; es decir $\mu_i > \mu_j$, difieren entre sí y de la media general. Supongamos que $\Delta = \mu_i - \mu_j$ denotan la diferencia entre las dos medias. Supongamos que $\Delta_i = \mu_i - \mu$ y $\Delta_j = \mu - \mu_j$. Por lo tanto, $\Delta = \Delta_i + \Delta_j$. Además, debido a que μ representa la media ponderada de todas las medias k, y se parte del supuesto de que (k – 2) menos las medias son iguales a μ, tenemos que:

$$\mu = \left[\sum_{l \neq i,j} w_l\mu_l + w_i(\mu + \Delta_i) + w_j(\mu - \Delta_j)\right] \Big/ \sum_{l=1}^{k} w_l = \mu + (w_i\Delta_i - w_j\Delta_j) \Big/ \sum_{l=1}^{k} w_l .$$

Por lo tanto:

$$w_i\Delta_i = w_j\Delta_j = w_j(\Delta - \Delta_i) ,$$

y por lo tanto,

$$\Delta_i = \frac{w_j}{w_i + w_j}\Delta$$

$$\Delta_j = \frac{w_i}{w_i + w_j}\Delta$$

Para esta particular configuración de las medias, podemos calcular el parámetro de no centralidad relacionado con la prueba de Welch:

$$\theta_W = w_i(\mu_i - \mu)^2 + w_j(\mu_j - \mu)^2$$
$$= \frac{w_i w_j^2 \Delta^2 + w_j w_i^2 \Delta^2}{(w_i + w_j)^2} = \frac{w_i w_j \Delta^2}{w_i + w_j}$$

Esta cantidad es creciente en w_i para w_j fijo y viceversa. Por lo tanto, se maximiza en el par (i, j) con las dos ponderaciones mayores y se minimiza en el par con las dos ponderaciones menores. Todos los cálculos de potencia consideran estos dos casos extremos, que minimizan y maximizan la potencia bajo el supuesto de que exactamente dos medias difieren del promedio de medias ponderado general.

Si especifica una diferencia para la prueba, se evalúan los valores de potencia mínimo y máximo para esta diferencia. El rango de estas potencias se indica en los informes relacionados con una barra codificada con color en la que las potencias en o por debajo de 60% están en el rojo, las potencias en o por encima de 90% están en el verde y las potencias entre 60% y 90% están en el amarillo. Los resultados de la Tarjeta de informe dependen de dónde se encuentre el rango de potencias en relación con esta escala codificada con color. Si el rango completo se encuentra en el rojo, entonces la potencia de cualquier par de grupos es menor que o igual a 60% y el icono rojo aparece en la tarjeta de informe para indicar un problema de potencia insuficiente. Si el rango completo se encuentra en el verde, la potencia de cualquier grupo es por lo menos 90% y el icono verde en la Tarjeta de informe indica la condición de potencia suficiente. El resto de las demás condiciones se trata como situaciones intermedias, que se indican con un icono amarillo en la Tarjeta de informe.

En los casos donde no se cumpla la condición de verde, el Asistente calcula un tamaño de la muestra que provocaría la condición de verde considerando la diferencia que especifique el usuario y las desviaciones estándar observadas de las muestras. La potencia estimada depende de los tamaños de las muestras a través de las ponderaciones $w_i = n_i/s_i^2$. Si se parte del supuesto de que todas las muestras tienen el mismo tamaño, entonces las ponderaciones más pequeñas corresponden a los dos grupos con las desviaciones estándar más grandes de las muestras. El Asistente busca un tamaño de la muestra que ofrezca una potencia de por lo menos 90% si la diferencia especificada se encuentra entre los dos grupos con la mayor variabilidad. Por lo tanto, tomar un tamaño de la muestra de por lo menos esta magnitud para todos los grupos provocaría que todo el rango de valores de potencia sea de por lo menos 90%, lo cual satisface la condición de verde.

Si el usuario no especifica una diferencia para el cálculo de potencia, entonces el Asistente busca la diferencia más elevada en la que el máximo del rango de potencias calculadas es 60%. Este valor se etiqueta en el límite entre las secciones roja y amarilla de la barra, que corresponde a la potencia de 60%. También busca la menor diferencia en la cual el mínimo del rango de

potencias calculadas es 90%. Este valor se etiqueta en el límite entre las secciones amarilla y verde de la barra, que corresponde a la potencia de 90%.

Cálculo de potencia

La potencia se calcula utilizando la aproximación de Kulinskaya et al. (2003):

Defina:

$\lambda = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2$,

$A = \sum_{i=1}^{k} h_i$,

$B = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2(1 - w_i/W)/(n_i - 1)$,

$D = \sum_{i=1}^{k} w_i^2(\mu_i - \mu)^4/(n_i - 1)$,

$E = \sum_{i=1}^{k} w_i^3(\mu_i - \mu)^6/(n_i - 1)^2$.

Los tres primeros cumulantes del numerador $\sum_{i=1}^{k} w_i(\bar{x}_i - \hat{\mu})^2$ del estadístico de Welch se pueden estimar como:

$\kappa_1 = k - 1 + \lambda + 2A + 2B$,

$\kappa_2 = 2(k - 1 + 2\lambda + 7A + 14B + D)$,

$\kappa_3 = 8(k - 1 + 3\lambda + 15A + 45B + 6D + 2E)$.

Supongamos que $F_{k-1,\,f,\,1-\alpha}$ denota el cuantil $(1 - \alpha)$ de la distribución $F(k - 1, f)$. Recuerde que todo $W^* \geq F_{k-1,\,f,\,1-\alpha}$ es el criterio para rechazar la hipótesis nula en una prueba de Welch con un tamaño de α.

Supongamos

$q = (k - 1)\left[1 + \frac{2(k-2)A}{k^2 - 1}\right] F_{k-1,f,1-\alpha}$,

$b = \kappa_1 - 2\kappa_2^2/\kappa_3$,

$c = \kappa_3/(4\kappa_2)$ [Nota: la expresión para c se muestra en Kulinskaya et al. (2003) sin los paréntesis.]

$\nu = 8\kappa_2^3/\kappa_3^2$.

Entonces, la potencia aproximada estimada de la prueba de Welch es:

$$P(\chi_\nu^2 \geq \frac{q - b}{c})$$

donde χ_ν^2 es una variable aleatoria de chi-cuadrado con ν grados de libertad.

Los siguientes resultados comparan la potencia de los dos métodos de aproximación y la potencia simulada de diferentes ejemplos, con base en 10,000 simulaciones.

Tabla 3 Cálculos de potencia para los dos métodos de aproximación en comparación con la potencia simulada

Ejemplo	Alfa	Potencia simulada	F no central	Kulinskaya et al.
μ's: 0, 0, 0, -0.1724, 0.8276	0.10	0.1372	0.135702	0.135795
σ's: 2, 2, 2, 2, 4	0.05	0.0739	0.072563	0.069512
n's: 12, 12, 12, 12, 10	0.01	0.0195	0.016587	0.012538
μ's: 0, 0, 0, -0.3448, 1.6552	0.10	0.2498	0.251064	0.257455
σ's: 2, 2, 2, 2, 4	0.05	0.1574	0.153128	0.156215
n's: 12, 12, 12, 12, 10	0.01	0.0541	0.045211	0.042195
μ's: 0, 0, 0, -0.5172, 2.4828	0.10	0.4534	0.44557	0.453506
σ's: 2, 2, 2, 2, 4	0.05	0.3211	0.311994	0.321575
n's: 12, 12, 12, 12, 10	0.01	0.1273	0.121225	0.125065
μ's: 0, 0, 0, -0.6896, 3.3104	0.10	0.662	0.671317	0.670296
σ's: 2, 2, 2, 2, 4	0.05	0.5219	0.533819	0.538617
n's: 12, 12, 12, 12, 10	0.01	0.2842	0.271316	0.282759
μ's: 0, 0, 0, -0.8620, 4.1380	0.10	0.8417	0.852589	0.846697
σ's: 2, 2, 2, 2, 4	0.05	0.7382	0.752173	0.746121
n's: 12, 12, 12, 12, 10	0.01	0.4883	0.487601	0.49323
μ's: 0, 0, 0, -1.0344, 4.9656	0.10	0.9429	0.952077	0.954929
σ's: 2, 2, 2, 2, 4	0.05	0.8866	0.901485	0.897937
n's: 12, 12, 12, 12, 10	0.01	0.691	0.711055	0.703379
μ's: 0, 0, 0, 0, 0, -0.148148, 1.85185	0.10	0.2011	0.189392	0.200114
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.1201	0.108986	0.11742
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.0385	0.028986	0.031456
μ's: 0, 0, 0, 0, 0, -0.296296, 3.70370	0.10	0.4942	0.485917	0.500143
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.3677	0.351593	0.375296
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.177	0.149041	0.177189
μ's: 0, 0, 0, 0, 0, -0.444444, 5.55556	0.10	0.8125	0.829702	0.819542
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.7131	0.727384	0.720807
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.4876	0.474291	0.49469

Ejemplo	Alfa	Potencia simulada	F no central	Kulinskaya et al.
μ's: 0, 0, 0, 0, 0, -0.592593, 7.40741	0.10	0.9645	0.977211	0.984213
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.9286	0.949997	0.949239
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.7938	0.831174	0.814067
μ's: 0, 0, 0, 0, 0, -0.740741, 9.25926	0.10	0.9961	0.998947	1.00000
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.9895	0.996653	1.00000
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.9528	0.977536	0.98705
μ's: 0, 0, 0, 0, 0, -0.888889, 11.1111	0.10	0.9999	0.999985	1.00000
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.9995	0.999926	1.00000
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.9943	0.99891	1.00000
μ's: 0, 0, 0, 0, 0, -0.518519, 6.48148	0.10	0.9059	0.929392	0.924696
σ's: 2, 2, 2, 2, 2, 2, 5	0.05	0.8403	0.868721	0.85672
n's: 20, 20, 20, 20, 20, 20, 10	0.01	0.6511	0.67121	0.66652
μ's: 0, 0, 0, 0, 0, -.5, .5	0.10	0.187	0.186658	0.18329
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.1098	0.106600	0.100189
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.0315	0.027773	0.021332
μ's: 0, 0, 0, 0, 0, -1, 1	0.10	0.4734	0.474736	0.472469
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.3394	0.338655	0.334430
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.1378	0.137788	0.128693
μ's: 0, 0, 0, 0, 0, -1.5, 1.5	0.10	0.8228	0.817355	0.810181
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.7112	0.707319	0.698461
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.4391	0.441154	0.431868
μ's: 0, 0, 0, 0, 0, -2, 2	0.10	0.9691	0.973246	0.973319
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.9312	0.940585	0.936546
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.78170.7817	0.799339	0.785099
μ's: 0, 0, 0, 0, 0, -2.5, 2.5	0.10	0.9984	0.998579	0.999763
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.9936	0.99533	0.997481
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.9587	0.967674	0.966249
μ's: 0, 0, 0, 0, 0, -3, 3	0.10	1.0000	0.999975	1.00000
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.9997	0.99987	1.00000
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.9959	0.997927	0.99961

Ejemplo	Alfa	Potencia simulada	F no central	Kulinskaya et al.
μ's: 0, 0, 0, 0, 0, -3.5, 3.5	0.10	1.00000	1.00000	1.00000
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	1.00000	1.00000	1.00000
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.99998	0.99995	1.00000
μ's: 0, 0, 0, 0, 0, -1.75, 1.75	0.10	0.914	0.921225	0.916652
σ's: 2, 2, 2, 2, 2, 2, 2	0.05	0.8418	0.852755	0.843856
n's: 12, 12, 12, 12, 12, 12, 12	0.01	0.619	0.633815	0.620704
μ's: 0, -0.5, 0.5	0.10	0.2548	0.259249	0.257149
σ's: 2, 2, 2	0.05	0.1549	0.160861	0.156251
n's: 12, 12, 12	0.01	0.0470	0.049045	0.042292
μ's: 0, -1, 1	0.10	0.654	0.659073	0.654105
σ's: 2, 2, 2	0.05	0.5205	0.522885	0.515816
n's: 12, 12, 12	0.01	0.2612	0.26355	0.252469
μ's: 0, -1.5, 1.5	0.10	0.9364	0.935939	0.937768
σ's: 2, 2, 2	0.05	0.8747	0.875620	0.872608
n's: 12, 12, 12	0.01	0.6614	0.664478	0.652563
μ's: 0, -1.75, 1.75	0.10	0.981	0.981434	0.986815
σ's: 2, 2, 2	0.05	0.9522	0.956100	0.959796
n's: 12, 12, 12	0.01	0.8251	0.830726	0.823624
μ's: 0, -2, 2	0.10	0.9953	0.995969	0.999332
σ's: 2, 2, 2	0.05	0.9878	0.988175	0.993705
n's: 12, 12, 12	0.01	0.9308	0.931922	0.933446
μ's: 0, -2.5, 2.5	0.10	0.9999	0.999923	1.00000
σ's: 2, 2, 2	0.05	0.9997	0.999634	1.00000
n's: 12, 12, 12	0.01	0.9949	0.994725	0.99909
μ's: 0, -3, 3	0.10	1.0000	1.00000	1.00000
σ's: 2, 2, 2	0.05	1.0000	1.00000	1.00000
n's: 12, 12, 12	0.01	0.9999	0.99985	1.00000
μ's: 0, -3.5, 3.5	0.10	1.0000	1.00000	1.00000
σ's: 2, 2, 2	0.05	1.0000	1.00000	1.00000
n's: 12, 12, 12	0.01	0.9999	1.00000	1.00000

Ejemplo	Alfa	Potencia simulada	F no central	Kulinskaya et al.
μ's: 0, -0.142857, 0.857143	0.10	0.1452	0.143156	0.146824
σ's: 2, 2, 4	0.05	0.0790	0.077699	0.077538
n's: 14, 12, 8	0.01	0.0223	0.018200	0.014338
μ's: 0, -0.285714, 1.71429	0.10	0.2765	0.274240	0.286222
σ's: 2, 2, 4	0.05	0.1787	0.170628	0.179469
n's: 14, 12, 8	0.01	0.0624	0.051588	0.050335
μ's: 0, -0.428571, 2.57143	0.10	0.4861	0.476925	0.490018
σ's: 2, 2, 4	0.05	0.3487	0.338626	0.355743
n's: 14, 12, 8	0.01	0.1467	0.132405	0.141352
μ's: 0, -0.50000, 3	0.10	0.5846	0.588533	0.596795
σ's: 2, 2, 4	0.05	0.4425	0.444491	0.460707
n's: 14, 12, 8	0.01	0.2107	0.19729	0.212798
μ's: 0, -0.571429, 3.42857	0.10	0.6933	0.694684	0.696773
σ's: 2, 2, 4	0.05	0.5631	0.555731	0.567129
n's: 14, 12, 8	0.01	0.3052	0.279131	0.299302
μ's: 0, -0.714286, 4.28571	0.10	0.8480	0.861469	0.859329
σ's: 2, 2, 4	0.05	0.7402	0.759703	0.759762
n's: 14, 12, 8	0.01	0.4871	0.480052	0.497421
μ's: 0, -0.857143, 5.14286	0.10	0.9434	0.952562	0.961913
σ's: 2, 2, 4	0.05	0.8869	0.898817	0.902716
n's: 14, 12, 8	0.01	0.6649	0.687058	0.692591
μ's: 0, -1, 6	0.10	0.9849	0.987981	0.999989
σ's: 2, 2, 4	0.05	0.9609	0.967589	0.985049
n's: 14, 12, 8	0.01	0.8294	0.847436	0.853787
μ's: 0, -1.14286, 6.85714	0.10	0.9976	0.997776	1.00000
σ's: 2, 2, 4	0.05	0.9890	0.992220	1.00000
n's: 14, 12, 8	0.01	0.9222	0.940972	0.96383

Ejemplo	Alfa	Potencia simulada	F no central	Kulinskaya et al.
μ's: 1, 2, 3	0.10	0.8838	0.882194	0.884649
σ's: 0.3, 2.4, 3.6	0.05	0.7995	0.797869	0.802137
n's: 13, 19, 25	0.01	0.5632	0.556486	0.563208
μ's: 1, 2, 3	0.10	0.5649	0.566831	0.565141
σ's: 2.77489, 2.77489, 2.77489	0.05	0.4305	0.431302	0.428126
n's: 13, 19, 25	0.01	0.1994	0.201329	0.195734

Los resultados anteriores se resumen en la gráfica de abajo, que muestra las discrepancias entre cada aproximación y el valor de la potencia estimada por simulación.

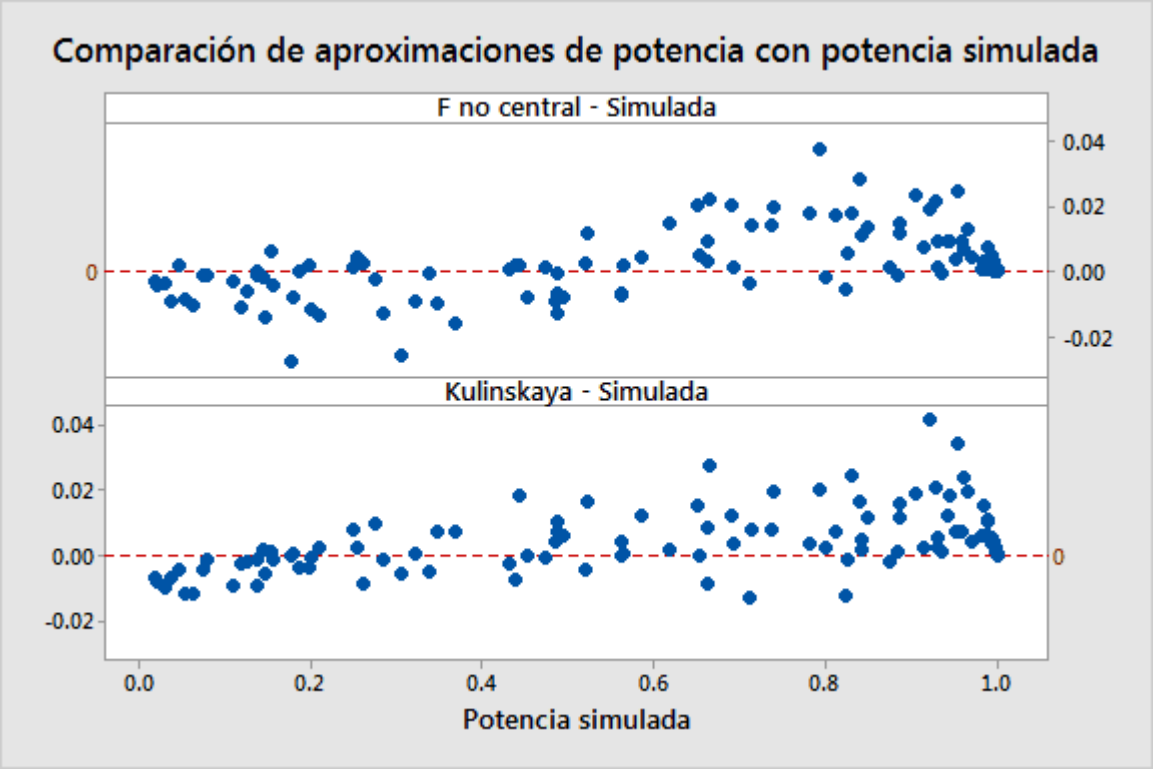


Figura 8 Comparación de dos aproximaciones de potencia con la potencia estimada por la simulación

Apéndice D: Normalidad

En esta sección, presentamos las simulaciones que examinaron el desempeño de la prueba de Welch y los intervalos de comparación con muestras de tamaño de pequeño a moderado de diversas distribuciones no normales.

Las tablas de abajo resumen los resultados de las simulaciones de los diferentes tipos de distribuciones en función de las hipótesis nulas de medias iguales. Para estos ejemplos, todas las desviaciones estándar son también iguales y todas las muestras tienen el mismo tamaño. El número de muestras es k = 3, 5, o 7.

Cada celda muestra la estimación del error Tipo I con base en 10,000 simulaciones. El nivel de significancia objetivo (α objetivo) es 0.05.

Tabla 4 Resultados de la simulación de la prueba de Welch con igual media para diferentes distribuciones

	Tamaño de la muestra n = 10			Tamaño de la muestra n = 15		
Distribución	**k = 3**	**k = 5**	**k = 7**	**k = 3**	**k = 5**	**k = 7**
N(0,1)	0.0490	0.0486	0.0512	0.0534	0.0522	0.0550
T(3)	0.0371	0.0361	0.0348	0.0353	0.0385	0.0365
T(5)	0.0440	0.0425	0.0439	0.0435	0.0428	0.0428
Laplace(0,1)	0.0433	0.0354	0.0345	0.0445	0.0397	0.0407
Uniforme(-1, 1)	0.0544	0.0640	0.0718	0.0517	0.0573	0.0585
Beta(3, 3)	0.0504	0.0577	0.0622	0.0501	0.0538	0.0564
Exponencial	0.0508	0.0621	0.0748	0.0483	0.0633	0.0779
Chi-cuadrado(3)	0.0473	0.0579	0.0753	0.0499	0.0588	0.0703
Chi-cuadrado(5)	0.0458	0.0594	0.0643	0.0504	0.0606	0.0679
Chi-cuadrado(10)	0.0463	0.0510	0.0585	0.0463	0.0552	0.0567
Beta(8, 1)	0.0500	0.0622	0.0775	0.0549	0.0653	0.0760

Todas las tasas de error Tipo I se encuentran dentro de 3 puntos porcentuales con respecto al α, incluso con muestras con un tamaño de 10. Las desviaciones grandes tienden a ocurrir con más grupos y con distribuciones que se alejan de lo normal. Con muestras con un tamaño de 10, los únicos casos en los que la probabilidad de aceptación se desvía en más de 2 puntos porcentuales son para k = 7. Estos ocurren para la distribución uniforme, que tiene colas mucho más cortas que la normal, así como para las altamente asimétricas distribuciones exponencial,

chi-cuadrado(3) y beta(8, 1). Aumentar los tamaños de las muestras a 15 mejora notablemente los resultados de la distribución uniforme, pero no así los de las dos distribuciones altamente asimétricas.

Realizamos una simulación similar para los intervalos de comparación. El α simulado en este caso es el número de simulaciones sobre 10,000 en el que algunos intervalos no se superponen. El $\alpha = 0.05$ objetivo.

Tabla 5 Resultados de las simulaciones de los intervalos de comparación con medias iguales para diferentes distribuciones

	Tamaño de la muestra n = 10			Tamaño de la muestra n = 15		
Distribución	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
N(0,1)	0.0493	0.0494	0.0469	0.0538	0.0518	0.0561
t(3)	0.0378	0.0321	0.0254	0.0347	0.0343	0.0289
t(5)	0.0449	0.0399	0.0361	0.0447	0.0444	0.0412
Laplace(0,1)	0.0438	0.0305	0.0246	0.0456	0.0366	0.0348
Uniforme(-1, 1)	0.0559	0.0605	0.0699	0.0534	0.0607	0.0590
Beta(3, 3)	0.0515	0.0569	0.0615	0.0510	0.0553	0.0568
Exponencial	0.0353	0.0254	0.0207	0.0346	0.0310	0.0275
Chi-cuadrado(3)	0.0375	0.0305	0.0296	0.0384	0.0359	0.0339
Chi-cuadrado(5)	0.0405	0.0390	0.0353	0.0417	0.0433	0.0416
Chi-cuadrado(10)	0.0425	0.0428	0.0447	0.0435	0.0476	0.0464
Beta(8, 1)	0.0381	0.0352	0.0287	0.0459	0.0428	0.0403

Al igual que con la prueba de Welch, las tasas de error Tipo I se encuentran todas dentro de 3 puntos porcentuales con respecto al α objetivo, incluso con muestras de un tamaño de 10. Las desviaciones más grandes tienden a ocurrir con más muestras y con distribuciones que se alejan de lo normal. Con muestras con un tamaño de 10, algunas veces las tasas de error se encuentran a más de 2 puntos porcentuales para k = 7 (y en un solo caso, para k = 5). Estos casos ocurren para la distribución t, la cual tiene colas extremadamente pesadas, con 3 grados de libertad, para la distribución de Laplace y para las altamente asimétricas distribuciones exponencial y Chi-cuadrado (3). Aumentar el tamaño de la muestra a 15 mejora los resultados, lo cual solo deja a las distribuciones t(3) y exponencial con valores de α simulados que se encuentran fuera del objetivo en más de 2 puntos porcentuales. Tenga en cuenta que a diferencia de los resultados de la prueba de Welch, las desviaciones más grandes correspondientes a los intervalos de comparación se encuentran del lado conservador.

ANOVA de un solo factor en el Asistente permite hasta k = 12 muestras, por lo que ahora consideramos resultados para más de 7 muestras. La tabla de abajo muestra las tasas de error Tipo I usando la prueba de Welch para datos no normales en k = 9 grupos. De nuevo, el $\alpha = 0.05$ objetivo.

Tabla 6 Resultados de la simulación de la prueba de Welch para diferentes distribuciones con 9 muestras

Distribución	k = 9
t(3)	0.0362
t(5)	0.0426
Laplace(0,1)	0.0402
Uniforme(-1, 1)	0.0625
Beta(3, 3)	0.0584
Exponencial	0.0885
Chi-cuadrado(3)	0.0774
Chi-cuadrado(5)	0.0686
Chi-cuadrado(10)	0.0581
Beta(8, 1)	0.0863

Como puede esperarse, las distribuciones altamente asimétricas muestran las mayores desviaciones con respecto al α objetivo. Aun así, ninguna de las tasas de error se desvía del objetivo en más de 4 puntos porcentuales, aunque la desviación de la distribución exponencial se encuentra cerca. La Tarjeta de informe trata las muestras de 15 lo suficiente como para no etiquetar un problema de datos no normales debido a que todos los resultados se encuentran por lo menos razonablemente cerca del α objetivo.

Las muestras de tamaño n = 15 no se comportan del mismo modo que cuando tenemos k = 12 muestras. A continuación, consideramos los resultados simulados de la prueba de Welch para diferentes tamaños de muestras utilizando distribuciones extremadamente no normales, lo cual nos permitirá desarrollar un criterio razonable para el tamaño de la muestra.

Tabla 7 Resultados de la simulación de la prueba de Welch para diferentes distribuciones con 12 muestras

n	T(3)	Uniforme	Chi-cuadrado(5)
10	0.0397	0.0918	0.0792
15	0.0351	0.0695	0.0717
20	0.0362	0.0622	0.0671
30	0.0408	0.0573	0.0657

Para estas distribuciones, n = 15 es aceptable si estamos dispuestos a aceptar una desviación ligeramente superior a 2 puntos porcentuales con respecto al α objetivo. Para mantener la desviación por debajo de 2 puntos porcentuales, el tamaño de la muestra debe ser 20. Ahora, consideramos los resultados de las distribuciones más asimétricas: chi-cuadrado (3) y exponencial.

Tabla 8 Resultados de la simulación de la prueba de Welch para las distribuciones chi-cuadrado y exponencial con 12 muestras

n	Chi-cuadrado(3)	Exponencial
10	0.1013	0.1064
15	0.0854	0.1079
20	0.0850	0.0951
30	0.0746	0.0829
40	0.0727	0.0735
50	0.0675	0.0694

Estas distribuciones altamente asimétricas presentan más de un desafío. Si está dispuesto a aceptar una desviación a más de 3 puntos porcentuales con respecto al $\alpha = 0.05$ objetivo, entonces n = 15 se podría considerar lo suficiente uniforme para la distribución de chi-cuadrado (3); sin embargo, la distribución exponencial requeriría algo más cercano a n = 30. Si bien el criterio de un tamaño de la muestra específico es quizás arbitrario, y que n = 20 funciona bastante bien para un amplio rango de distribuciones y marginalmente bien para distribuciones extremadamente asimétricas, utilizamos n = 20 como el tamaño de muestra mínimo recomendado para de 10 a 12 muestras. Claramente, si existiera la necesidad de mantener poca desviación, incluso para las distribuciones extremadamente asimétricas, entonces se recomiendan muestras más grandes.