RESUMO DO ASSISTENTE DO MINITAB

Este artigo é parte de uma série de artigos que explicam a pesquisa conduzida pelos estatísticos do Minitab para desenvolver os métodos e verificações de dados usados no Assistente no Minitab Statistical Software.

ANOVA para um fator

Visão geral

A ANOVA para um fator é usada para comparar as médias de três ou mais grupos para determinar se elas diferem significativamente uma da outra. Outra função importante é estimar as diferenças entre grupos específicos.

O método mais comum para detectar diferenças entre grupos na ANOVA para um fator é o teste F, que é baseado na suposição de que populações de todas as amostras compartilham um desvio padrão comum, porém desconhecido. Reconhecemos, na prática, que as amostras frequentemente têm diferentes desvios padrão. Portanto, queríamos investigar o método de Welch, alternativa ao teste F, que pode lidar com desvios padrão diferentes. Também queríamos desenvolver um método para calcular múltiplas comparações que correspondem a amostras com desvios padrão diferentes. Com este método, podemos representar graficamente intervalos individuais, que fornecem uma maneira fácil de identificar grupos que diferem uns dos outros.

Neste documento, descrevemos como nós desenvolvemos os métodos usados no procedimento ANOVA para um fator do Assistente do Minitab.

- Teste de Welch
- Intervalos de comparação múltiplos

Adicionalmente, examinamos as condições que podem afetar a validade dos resultados do ANOVA para um fator, incluindo a presença de dados atípicos, o tamanho amostral e o poder do teste e a normalidade dos dados. Com base nessas condições, o assistente realiza automaticamente as seguintes verificações em seus dados e relata os resultados no Cartão de Relatórios:

- Dados atípicos
- Tamanho amostral
- Normalidade dos dados

Neste artigo, investigamos como essas condições se relacionam à ANOVA para um fator na prática e descrevemos como estabelecemos as diretrizes para verificar essas condições no Assistente.

Métodos ANOVA para um fator

O teste F versus o teste de Welch

O teste F comumente usado na ANOVA para um fator é baseado na suposição de que todos os grupos compartilham um desvio padrão (σ) comum, porém desconhecido. Na prática, essa suposição raramente é real, o que leva a problemas de controle de taxa de erros do Tipo I. O erro do Tipo I é a probabilidade de rejeitar incorretamente a hipótese nula (concluindo que as amostras sejam significativamente diferentes quando elas não são). Quando as amostras têm diferentes desvios padrão, há uma maior verossimilhança de que o teste irá alcançar uma conclusão incorreta. Para abordar este problema, o teste de Welch foi desenvolvido como uma alternativa ao teste F (Welch, 1951).

Objetivo

Queríamos determinar se usamos o teste F ou o de Welch para o procedimento ANOVA para um fator no Assistente. Para fazer isso, precisávamos avaliar quão próximo os resultados do teste real do teste F e de Welch correspondiam ao nível alvo de significância (alfa ou taxa de erros Tipo I) do teste, isto é, se o teste rejeitou incorretamente a hipótese nula com mais ou menos frequência do que o pretendido, dados os diferentes tamanhos e os desvios padrão amostrais.

Método

Para comparar o teste F e o de Welch, realizamos múltiplas simulações, variando o número de amostras, o tamanho amostral e o desvio padrão amostral. Para cada condição, realizamos 10.000 testes ANOVA usando o teste F e o método de Welch. Geramos dados aleatórios para que as médias das amostras fossem as mesmas e desta forma, para cada teste, a hipótese nula foi real. Depois disso, realizamos os testes usando níveis de significância alvo de 0,05 e 0,01. Contamos o número de vezes entre os 10.000 testes em que os testes F e Welch realmente rejeitaram a hipótese nula e comparamos essa proporção ao nível de significância alvo. Se o teste tiver sido corretamente realizado, as taxas de erro tipo I estimadas deverão estar muito próximas do nível de significância alvo.

Resultados

Descobrimos que o método de Welch apresentou desempenho tão bom ou melhor do que o teste F, sob todas as condições que testamos. Por exemplo, ao comparar 5 amostras usando o teste de Welch, as taxas de erros Tipo I estavam entre 0,0460 e 0,0540, muito próximas do nível de significância alvo de 0,05. Indica que a taxa de erros Tipo I do método de Welch corresponde ao valor alvo mesmo quando o tamanho amostral e o desvio padrão variam entre amostras.

Por outro lado, as taxas de errso Tipo I do teste F ficaram entre 0,0273 e 0,2277. Em particular, o teste F teve um desempenho ruim sob as seguintes condições:

- As taxas de erros Tipo I caíram abaixo de 0,05 quando a maior amostra também apresentou o maior desvio padrão. Esta condição resulta em um teste mais conservador e demonstra que simplesmente aumentar o tamanho amostral não é uma solução viável quando os desvios padrão das amostras não são iguais.
- As taxas de erros Tipo I ficaram acima de 0,05 quando os tamanhos amostrais eram iguais, mas os desvios padrão eram diferentes. As taxas também eram maiores do que 0,05 quando a amostra com um desvio padrão maior era de um tamanho menor do que as outras amostras. Em particular, quando amostras menores têm desvios padrão maiores, há um aumento substancial no risco de que este teste rejeite incorretamente a hipótese nula.

Para obter mais informações sobre a metodologia e os resultados da simulação, consulte o Apêndice A.

Devido ao bom desempenho do método de Welch quando os desvios padrão e tamanhos amostrais eram diferentes, usamos o método de Welch para o procedimento ANOVA para um fator no Assistente.

Intervalos de comparação

Quando um teste ANOVA é estatisticamente significativo, indicando que, no mínimo, uma das médias amostrais é diferente das outras, o próximo passo na análise é determinar quais amostras são estatisticamente diferentes. Uma maneira intuitiva para fazer esta comparação é representar graficamente os intervalos de confiança e identificar as amostras cujos intervalos não se sobrepõem. Contudo, as conclusões tiradas do gráfico podem não corresponder aos resultados do teste porque os intervalos de confiança individuais não são criados para comparações. Apesar de existir um método publicado para múltiplas comparações para amostras com desvios padrão iguais, precisávamos estender este método para considerar amostras com desvios padrão diferentes.

Objetivo

Queríamos desenvolver um método para calcular intervalos de comparação individuais que podem ser usados para fazer comparações entre amostras e que também correspondessem os resultados do teste o mais proximamente possível. Também queríamos fornecer um método visual para determinar quais amostras são estatisticamente diferentes das outras.

Método

Métodos de múltiplas comparações padrão (Hsu 1996) fornecem um intervalo para a diferença entre cada par de médias, ao mesmo tempo em que controlam os erros crescentes que ocorrem ao fazermos múltiplas comparações. No caso especial de tamanhos amostrais iguais e sob a suposição de desvio padrão iguais, é possível exibir intervalos individuais para cada média de uma maneira que corresponda exatamente aos intervalos para as diferenças de todos os pares. Para o caso de tamanhos amostrais diferentes, com a suposição de desvios padrão iguais, Hochberg, Weiss, and Hart (1982) desenvolveram intervalos

individuais que são aproximadamente equivalentes aos intervalos para diferenças entre pares, com base no método Tukey-Kramer de múltiplas comparações. No Assistente, aplicamos a mesma abordagem ao método Games-Howell de múltiplas comparações, que não assume desvios padrão iguais. A abordagem usada no Assistente, na versão 16 do Minitab foi similar em conceito, mas não foi baseada diretamente na abordagem Games-Howell. Para obter mais detalhes, consulte o Apêndice B.

Resultados

O Assistente exibe os intervalos de comparação no Gráfico de comparações de médias no Relatório de resumo ANOVA para um fator. Quando o teste ANOVA é estatisticamente significativo, qualquer intervalo de comparação que não se sobrepõe com, no mínimo, um outro intervalo, é marcado em vermelho. É possível que o teste e os intervalos de comparação discordem, apesar deste resultado ser raro porque ambos os métodos têm a mesma probabilidade de rejeitar a nipótese nula quando esta é real. Se o teste ANOVA for significativo e, contudo, todos os intervalos se sobrepuserem, o par com a menor quantidade de sobreposição será marcado em vermelho. Se o teste ANOVA não for estatisticamente significativo, nenhum dos intervalos será marcado em vermelho, ainda que alguns dos intervalos não se sobreponha.

Verificações dos dados

Dados atípicos

Dados atípicos são valores de dados extremamente grandes ou pequenos, também conhecidos como outliers. Os dados atípicos podem ter uma forte influência nos resultados da análise e podem afetar as chances de encontrar resultados estatisticamente significativos, especialmente quando a amostra é pequena. Dados atípicos podem indicar problemas com coleta de dados ou um comportamento atípico do processo que você está estudando. Portanto, muitas vezes, vale a pena investigar esses pontos de dados e eles devem ser corrigidos quando possível.

Objetivo

Queríamos desenvolver um método para verificar valores de dados que fossem muito grandes ou muito pequenos na amostra geral, e que podem afetar o resultado da análise.

Método

Desenvolvemos um método de verificação de dados atípicos que se baseia no método descrito por Hoaglin, Iglewicz e Tukey (1986) para identificar outliers nos boxplots.

Resultados

O Assistente identifica um ponto de dados como atípico quando sua amplitude interquartílica ultrapassa em 1,5 vez o quartil inferior ou superior da distribuição. Os quartis inferior e super são os percentis 25° e 75° dos dados. O intervalo interquartílico é a diferença entre os dois quartis. Esse método funciona bem mesmo quando há vários outliers, porque ele possibilita a detecção de cada outlier específico.

Ao verificar dados atípicos, o Assistente exibe os seguintes indicadores de status no Cartão de Relatório:

Status	Condição
	Não há pontos de dados incomuns.
	No mínimo um ponto de dados é atípico e pode ter forte influência sobre os resultados.

Tamanho amostral

O poder é uma importante propriedade de qualquer teste de hipótese porque ele indica a verossimilhança de que você irá encontrar um efeito significativo ou diferença quando existe um realmente. O poder é a probabilidade de que você irá rejeitar a hipótese nula a favor da hipótese alternativa. Frequentemente, a maneira mais fácil de aumentar o poder de um teste é aumentar o tamanho amostral. No Assistente, para os testes com poder baixo, indicamos quão grande sua amostra precisa ser para encontrar a diferença que você especificou. Se nenhuma diferença for especificada, relatamos a diferença que você poderia detectar com poder adequado. Para fornecer as informações, precisamos desenvolver um método para calcular o poder porque o Assistente usa o método de Welch, que não tem uma fórmula exata para poder.

Objetivo

Para desenvolver uma metodologia para calcular o poder, precisamos abordar duas questões. Primeiro, o Assistente não exige que os usuários insiram um conjunto completo de médias; ele só requer que eles insiram uma diferença entre as médias que tenham implicações práticas. Para qualquer diferença fornecida, existe um número infinito de possíveis configurações de médias que poderiam produzir aquela diferença. Portanto, precisávamos desenvolver uma abordagem razoável para determinar quais médias usar ao calcular o poder, dado que poderíamos não calcular o poder para todas as configurações possíveis de médias. Segundo, precisávamos desenvolver um método para calcular o poder, porque o Assistente usa o método de Welch, que não exige tamanhos amostrais iguais ou desvios padrão.

Método

Para abordar o número infinito de possíveis configurações de médias, desenvolvemos um método baseado na abordagem usada no procedimento padrão ANOVA para um fator no Minitab (**Stat** > **ANOVA** > **Um fator**). Focamos nos casos onde somente duas das médias diferem pelo valor declarado e a outras médias são iguais (definidas para a média ponderada das médias). Como supomos que somente duas médias diferem da média geral (e não mais de duas), a abordagem fornece uma estimativa conservadora de poder. Contudo, como as amostras podem ter tamanhos ou desvios padrão diferentes, o cálculo de poder ainda depende de quais duas médias supõe-se que diferem.

Para solucionar este problema, identificamos os dois pares de médias que representam o melhor e o pior casos. O pior caso ocorre quando o tamanho amostral é pequeno em relação à variância amostral, e o poder é minimizado; o melhor caso ocorre quando o tamanho amostral é grande em relação à variância amostral e o poder é maximizado. Todos os cálculos de poder consideram esses dois casos extremos, que minimizam e maximizam o poder sob a suposição de que exatamente duas médias diferem da média ponderada geral das médias.

Para desenvolver o cálculo de poder, usamos um método mostrado em Kulinskaya et al. (2003). Comparamos os cálculos de poder da nossa simulação, o método que desenvolvemos para abordar a configuração das médias e o método mostrado em Kulinskaya et al. (2003). Também examinamos outra aproximação de poder que mostra mais

claramente como o poder depende da configuração das médias. Para obter mais informações sobre o cálculo do poder, consulte o Apêndice C.

Resultados

Nossa comparação desses métodos mostraram que o método Kulinskaya fornece uma boa aproximação de poder e que nosso método para lidar com a configuração das médias é apropriado.

Quando os dados não fornecem evidência suficiente contra a hipótese nula, o Assistente calcula diferenças práticas que podem ser detectadas com uma probabilidade de 80% e de 90% para os tamanhos amostrais dados. Além disso, se você especificar uma diferença prática, o Assistente calcula os valores de poder mínimo e máximo para esta diferença. Quando os valores de poder estão abaixo de 90%, o Assistente calcula um tamanho amostral com base na diferença especificada e nos desvios padrão amostrais observados. Para garantir que o tamanho amostral resulta em ambos valores de poder mínimo e máximo sendo 90% ou maiores, supomos que a diferença especificada esteja entre as duas médias com a maior variabilidade.

Se o usuário não especificar uma diferença, o Assistente encontra a maior diferença na qual o máximo do intervalo de valores de poder seja 60%. Este valor é rotulado no limite entre as barras vermelha e amarela no Relatório de Poder, correspondendo a 60% do poder. Também encontramos a menor diferença na qual o mínimo do intervalo de valores de poder é 90%. Este valor é rotulado no limite entre as barras amarela e verde no Relatório de Poder, correspondendo a 90% do poder.

Ao verificar o poder e o tamanho amostral, o Assistente exibe os seguintes indicadores de status no Cartão de Relatórios:

Status	Condição
	Os dados não fornecem evidências suficientes para concluir que há diferenças entre as médias. Nenhuma diferença foi especificada.
	O teste descobre uma diferença entre as médias, portanto, o poder não é um problema. OU O poder é suficiente. O teste não encontrou uma diferença entre as médias, mas a amostra é grande o suficiente para fornecer pelo menos uma chance de 90% de detecção da diferença dada.
	O poder pode ser suficiente. O teste não encontrou uma diferença entre as médias, mas a amostra é grande o suficiente para fornecer uma chance de 80% a 90% de detecção da diferença dada. O tamanho amostral necessário para atingir 90% de poder é informado.
	Talvez o poder não seja suficiente. O teste não encontrou uma diferença entre as médias, e a amostra é grande o suficiente para fornecer uma chance de 60% a 80% de detecção da diferença dada. Os tamanhos amostrais necessários para atingir 80% e 90% de poder são informados.
	O poder não é suficiente. O teste não encontrou uma diferença entre as médias, e a amostra não é grande o suficiente para fornecer pelo menos uma chance de 60% de detecção da diferença dada. Os tamanhos amostrais necessários para atingir 80% e 90% de poder são informados.

Normalidade

Uma suposição comum em diversos métodos estatísticos é que os dados são normalmente distribuídos. Felizmente, mesmo quando os dados não são normalmente distribuídos, os métodos baseados na suposição de normalidade podem funcionar bem. Isso é explicado, em parte, pelo teorema do limite central, que diz que a distribuição de qualquer média amostral tem uma distribuição normal aproximada, e que a aproximação torna-se quase normal conforme o tamanho amostral torna-se maior.

Objetivo

Nosso objetivo foi determinar quão grande a amostra precisa ser para dar uma aproximação razoavelmetne boa da distribuição normal. Queríamos examinar o teste de Welch e intervalos de comparação com amostras de tamanho pequeno a moderado com diversas distribuições não-normais. Queríamos determinar o quão próximo os resultados do teste real do método de Welch e os intervalos de comparação corresponderam ao nível escolhido de significância (alfa ou taxa de erros Tipo I) para o teste, ou seja, se o teste rejeitou incorretamente a hipótese nula com mais ou menos frequência do que era esperado dado os tamanhos amostrais diferentes, números de níveis e distribuições não-normais.

Método

Para estimar o erro Tipo I, realizamos múltiplas simulações, variando o número de amostras, tamanho amostral e a distribuição dos dados. As simulações incluíram distribuições assimétricas e de cauda pesada que se desviam substancialmente da distribuição normal. O desvio padrão e de tamanho foram constantes entre amostras dentro de cada teste.

Para cada condição, realizamos 10.000 testes ANOVA usando o método de Welch e os intervalos de comparação. Geramos dados aleatórios para que as médias das amostras fossem as mesmas e desta forma, para cada teste, a hipótese nula foi real. Depois disso, realizamos os testes usando um nível de significância de destino de 0,05. Contamos o número de vezes entre os 10.000 quando os testes realmente rejeitaram a hipótese nula e comparamos essa proporção ao nível de significância alvo. Para os intervalos de comparação, contamos o número de vezes entre 10.000 quando os intervalos indicam uma ou mais diferenças. Se o teste apresentar bom desempenho, as taxas de erro Tipo I deverão estar muito próximas do nível de significância alvo.

Resultados

No geral, os testes e os intervalos de comparação apresentam desempenho muito bom entre todas as condições com tamanhos amostrais tão pequenos como 10 ou 15. Para testes com níveis 9 ou menores, em quase todos os casos, os resultados estão todos dentro de 3 pontos percentuais do nível de significância alvo para um tamanho amostral de 10 e dentro de 2 pontos percentuais para um tamanho amostral de 15. Para testes que têm 10 ou mais níveis, na maioria dos casos, os resultados estão dentro de 3 pontos percentuais com um tamanho amostral de 15 e dentro de 2 pontos percentuais com um tamanho amostral de 20. Para obter mais informações, consulte o Apêndice D.

Como os testes foram corretamente realizados com amostras relativamente pequenas, o Assistente não testa os dados quanto à normalidade. Em vez disso, o Assistente verifica os tamanhos amostrais e indica quando as amostras têm menos de 15 para níveis de 2 a 9 e menos de 20 para níveis de 10 a 12. Com base nestes resultados, o Assistente exibe os indicadores de status a seguir no Cartão de Relatórios:

Status	Condição
	Os tamanhos amostrais são, no mínimo, de 15 ou 20, portanto, a normalidade não é um problema.
	Como alguns tamanhos amostrais são menores do que 15 ou 20, a normalidade pode ser um problema.

Referências

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

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

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

Apêndice A: O teste F versus o teste de Welch

O teste F pode resultar em um aumento da taxa de erros do Tipo I quando a suposição de desvios padrão iguais é violada, o teste de Welch foi criado para evitar esses problemas.

Teste de Welch

Amostras aleatórias de tamanhos n_1, ..., n_k de k populações são observadas. Permita que $\mu_1,\ldots,\mu_k$ denote as médias populacionais e permita que $\sigma_1^2, \ldots, \sigma_k^2$ denote as variâncias populacionais. Permita que $\bar{x}_1, \ldots, \bar{x}_k$ denote as médias amostrais e permita que $s_1^2, \ldots, s_k^2$ denote as variâncias amostrais. Estamos interessados em testar as hipóteses:

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

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

O teste de Welch para testar a igualdade de k médias compara a estatística

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

à distribuição F(k – 1, f), em que

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

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

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

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

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

O teste de Welch rejeita a hipótese nula se $W^* \geq F_{k-1,f,1-\alpha}$, o percentil da distribuição F que é excedido com probabilidade α.

Desvios padrão diferentes

Nesta seção demonstramos a sensibilidade do teste F a violações da suposição de desvios padrão iguais e o comparamos ao teste de Welch.

Os resultados a seguir são para testes ANOVA para um fator usando 5 amostras de N(0, σ2). Cada linha é baseada em 10.000 simulações usando-se o teste F e o teste de Welch. Testamos duas condições para o desvio padrão aumentando o desvio padrão da quinta amostra, dobrando-a e quadruplicando-a comparada a outras amostras. Testamos as três diferentes condições para o tamanho amostral: tamanhos amostrais são iguais, a quinta amostra é maior do que as outras, e a quinta amostra é menor do que as outras.

Tabela 1 Taxas de erros do Tipo I para testes F simulados e testes de Welch com 5 amostras com nível de significância alvo α = 0,05

Desvio padrão (σ1, σ2, σ3, σ4, σ5)	Tamanho amostral (n1, n2, n3, n4, n5)	Teste F	Teste de Welch
1, 1, 1, 1, 2	10, 10, 10, 10, 20	0,0273	0,0524
1, 1, 1, 1, 2	20, 20, 20, 20, 20	0,0678	0,0462
1, 1, 1, 1, 2	20, 20, 20, 20, 10	0,1258	0,0540
1, 1, 1, 1, 4	10, 10, 10, 10, 20	0,0312	0,0460
1, 1, 1, 1, 4	20, 20, 20, 20, 20	0,1065	0,0533
1, 1, 1, 1, 4	20, 20, 20, 20, 10	0,2277	0,0503

Quando os tamanhos amostrais são iguais (linhas 2 e 5), a probabilidade de que o teste F rejeite incorretamente a hipótese nula é maior do que o alvo 0,05, e a probabilidade aumenta quando a diferença entre desvios padrão é maior. O problema fica ainda pior ao dimunuirmos o tamanho da amostra com o maior desvio padrão. Por outro lado, aumentar o tamanho da amostra com o maior desvio padrão reduz a probabilidade de rejeição. Contudo, aumentar o tamanho amostral demasiadamente torna a probabilidade de rejeição muito pequena, que não somente torna o teste mais conservador do que o necessário sob a hipótese nula, mas também afeta adversamente o poder do teste sob a hipótese alternativa. Compare esses resultados com o teste de Welch, que concorda bem com o nível de significância alvo de 0,05 em cada caso.

Em seguida, conduzimos uma simulação para os casos com k = 7 amostras. Cada linha da tabela resume 10.000 testes F simulados. Variamos os desvios padrão e tamanhos das amostras. Os níveis de significância alvo são α = 0,05 e α = 0,01. Como acima, vemos desvios dos valores alvo que podem ser bem grandes. Usando um tamanho amostral menor quando a variabilidade é maior leva a probabilidades de erros do Tipo I muito grandes, ao passo que usar uma amostra maior pode levar a um teste extremamente conservador. Os resultados são apresentados na Tabela 2 abaixo.

Tabela 2 Taxas de erros Tipo I para testes F simulados com 7 amostras

Desvio padrão (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	Tamanhos amostrais (n1, n2, n3, n4, n5, n6, n7)	Alvo α = 0,05	Alvo 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

Desvio padrão (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	Tamanhos amostrais (n1, n2, n3, n4, n5, n6, n7)	Alvo α = 0,05	Alvo α = 0,01
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
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

Desvio padrão (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	Tamanhos amostrais (n1, n2, n3, n4, n5, n6, n7)	Alvo α = 0,05	Alvo α = 0,01
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
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

Apêndice B: Intervalos de comparação

O gráfico de comparação de médioas permite avaliar a significância estatística de diferenças entre as médias populacionais.

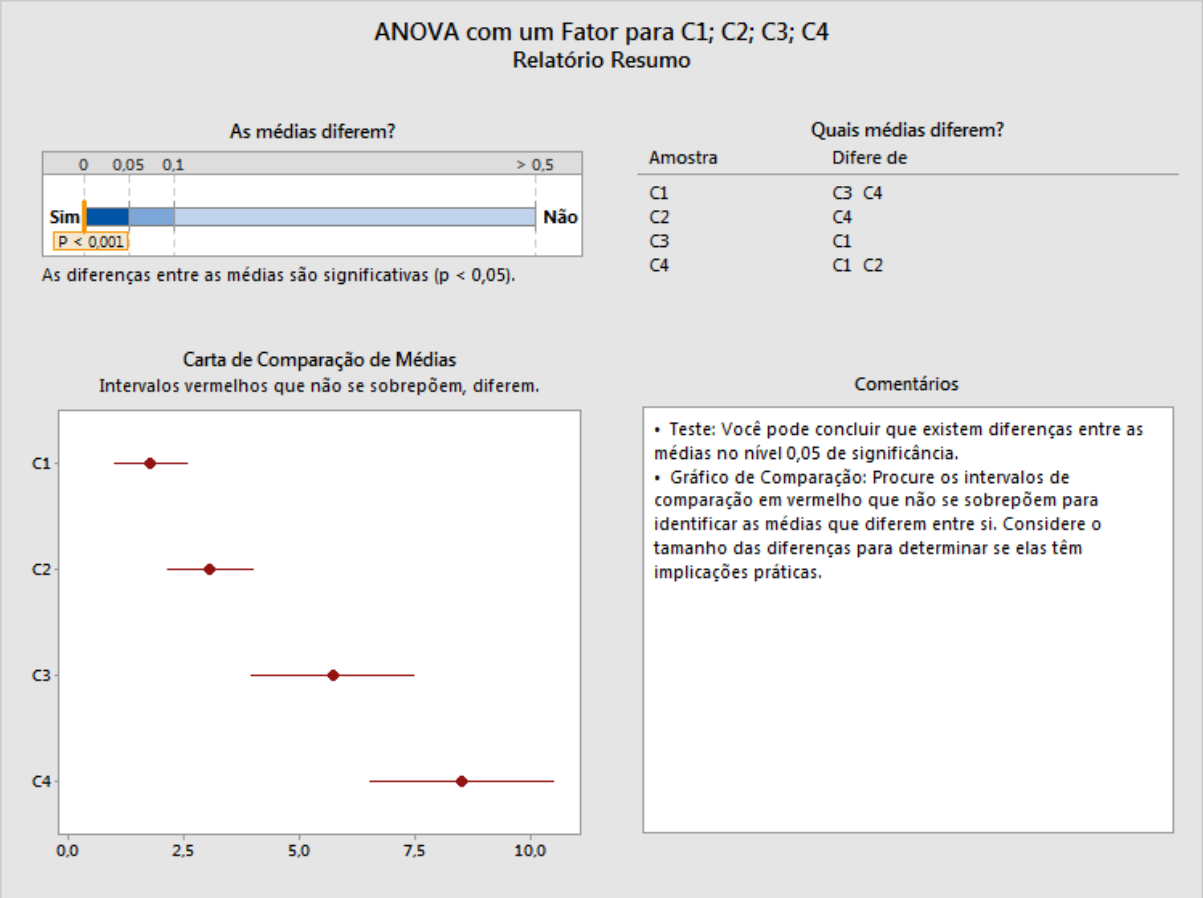
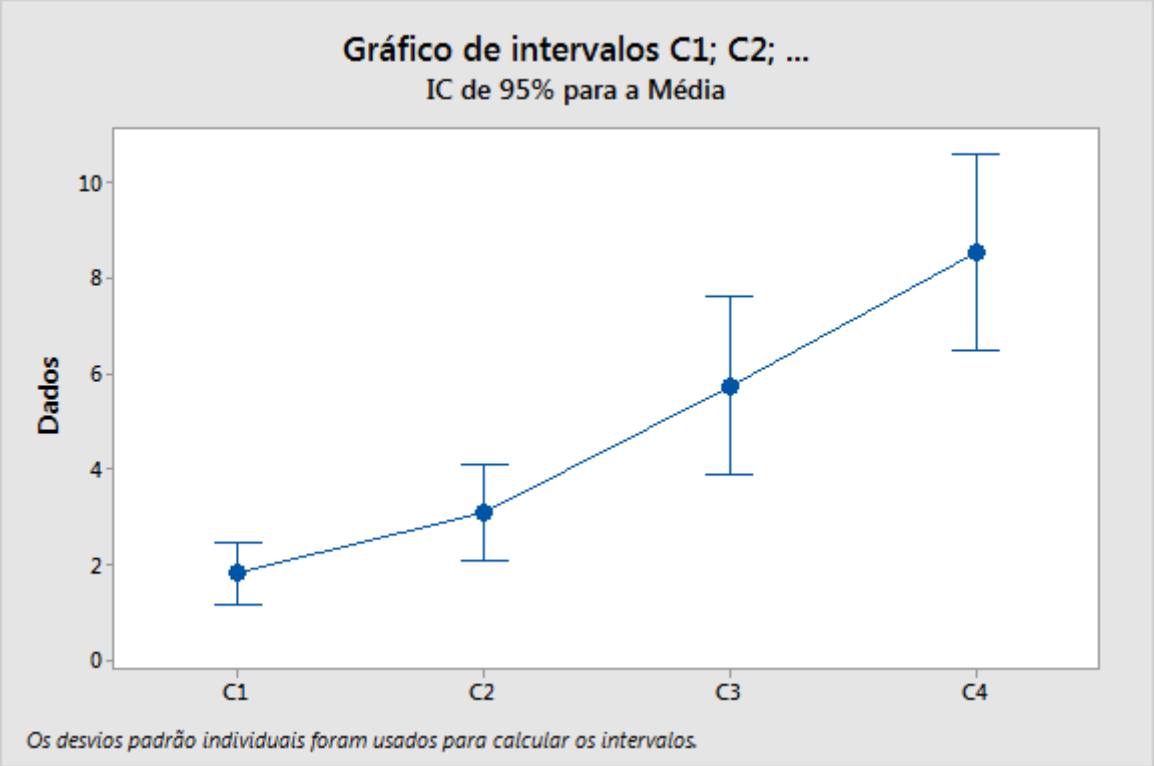


Figura 1 O Gráfico de Comparação de Médias no Relatório de resumo ANOVA para um fator do Assistente

Um conjunto similar de intervalos aparece na saída do procedimento padrão ANOVA para um fator no Minitab (**Stat** > **ANOVA** > **Um fator**):



Contudo, observe que os intervalos acima são simplemente intervalos de confiança individuais para as médias. Quando o teste ANOVA (F ou Welch) conclui que algumas médias são diferentes, há uma tendência natural para procurar intervalos que não se sobrepõem e tirar conclusões sobre quais médias diferem. Essa análise informal dos intervalos de confiança individuais irá frequentemente levar a conclusões razoáveis, mas ela não controla a probabilidade de erro da mesma forma que o teste ANOVA faz. Dependendo do número de populações, os intervalos podem ter substancialmente mais ou menos verossimilhança do que o teste para concluir que há diferenças. Como um resultado, os dois métodos podem facilmente alcançar conclusões inconsistentes. O gráfico de comparação é criado para corresponder mais consistentemente aos resultados do teste de Welch fazendo múltiplas comparações, apesar de nem sempre ser possível alcançar consistência completa.

Os métodos de múltiplas comparações, como as comparações de Tukey-Kramer e Games-Howell no Minitab (**Stat** > **ANOVA** > **Um fator**), permitem tirar conclusões estatisticamente válidas sobre diferenças entre as médias individuais. Esses dois métodos são métodos de comparação pareada, que fornecem um intervalo para a diferença entre cada par de médias. A probabilidade de que todos os intervalos simultaneamente contenham as diferenças que eles estão estimando é, no mínimo $1 - \alpha$. O método Tukey-Kramer depende da suposição de variâncias iguais, enquanto o método Games-Howell não requer variâncias iguais. Se a hipótese nula de médias iguais for real, todas as diferenças são zero, e a probabilidade de que quaisquer dos intervalos Games-Howell irão falhar em conter zero é de, no máximo α. Portanto, podemos usar os intervalos para realizar um teste de hipótese com nível de significância α. Usamos intervalos Games-Howell como ponto inicial para derivar os intervalos do gráfico de comparação no Assistente.

Dado como um conjunto de intervalos [L_{ij}, U_{ij}] para todas as diferenças $\mu_i - \mu_j$, $1 \le i < j \le k$, queremos encontrar um conjunto de intervalos [L_i, U_i] para as médias individuais μ_i, $1 \le i \le k$, que transmite as mesmas informações. Isso requer que quaisquer diferenças d estejam no intervalo [L_{ij}, U_{ij}] se, e somente se, existir $\mu_i \in [L_i, U_i]$ e $\mu_j \in [L_j, U_j]$ de tal forma que $\mu_i - \mu_j = d$.. Os pontos extremos dos intervalos devem ser relacionados pelas equações

$$U_i - L_j = U_{ij} \text{ e}$$

$$L_i - U_j = L_{ij}.$$

Para k = 2, temos somente uma diferença, mas dois intervalos individuais, portanto, é possível obter intervalos de comparação exata. Na realidade, existe bastante flexibilidade na largura dos intervalos que satisfazem esta condição. Para k = 3, há três diferenças e três intervalos individuais, portanto, novamente, é possível satisfazer a condição, mas agora sem a flexibilidade na configuração da largura dos intervalos. Para k = 4, há seis diferenças, mas somente quatro intervalos individuais. Os intervalos de comparação devem tentar transmitir as mesmas informações usando menos intervalos. Em geral, para k ≥ 4, há mais diferenças do que médias individuais, portanto, não há uma solução exata, exceto se condições adicionais forem impostas nos intervalos para diferenças, como larguras iguais.

Os intervalos de Tukey-Kramer têm larguras iguais somente se todos os tamanhos amostrais forem iguais. As larguras iguais também são uma consequência de supor variâncias iguais. Os intervalos Games-Howell não supõem variâncias iguais e, portanto, não têm larguras iguais. No Assistente, teremos que confiar em métodos aproximados para definir intervalos de comparação.

O intervalo Games-Howell para $\mu_i - \mu_j$ é

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

em que $q^*(k, \hat{v}_{ij})$ é o percentil apropriado da distribuição da amplitude estudentizada, que depende de k,o número de médias que estão sendo comparadas, e em

ν_{ij}, os graus de liberdade associados com o 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, and Hart (1982) obtiveram intervalos individuais que são aproximadamente equivalentes a essas comparações pareadas usando:

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

Os valores X_i são selecionados para minimizar

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

Em que:

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

Adaptamos esta abordagem para o caso de variâncias diferentes ao derivar intervalos de comparações de Games-Howell da forma

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

Os valores d_i são selecionados para minimizar

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

Em que:

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

A solução é

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

Os gráficos a seguir comparam resultados de simulações do teste de Welch com os resultados dos intervalos de comparação usando dois métodos: o método com base em Games-Howell que usamos agora e o método usado na versão 16 do Minitab com base em uma média aritmética de graus de liberdade. O eixo vertical é a proporção de vezes em 10.000 simulações que o teste de Welch rejeita incorretamente a hipótese nula ou que nem todos os intervalos de comparação se sobrepõem. O alvo alfa é $\alpha = 0{,}05$ nestes exemplos. Essas simulações cobrem diversos casos de desvios padrão e tamanhos amostrais diferentes; cada posição ao longo do eixo horizontal representa um caso diferente.

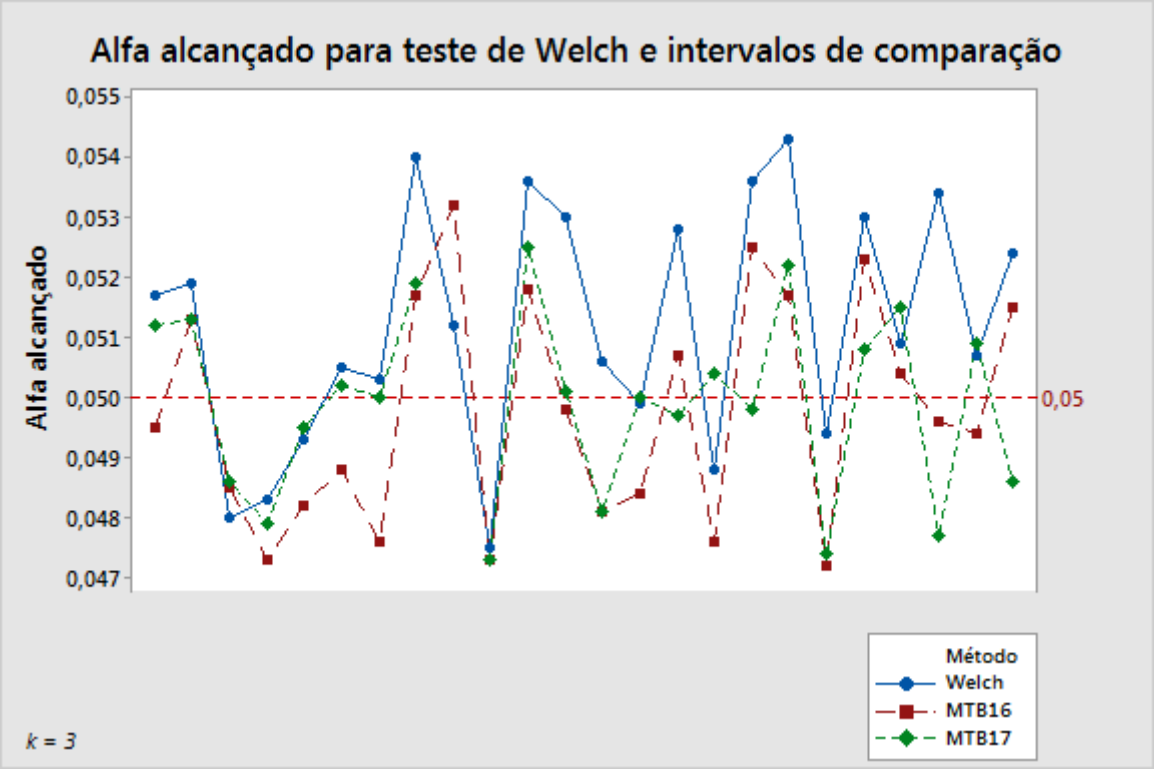


Figura 2 Teste de Welch comparado com dois métodos de calcular intervalos de comparação para 3 amostras

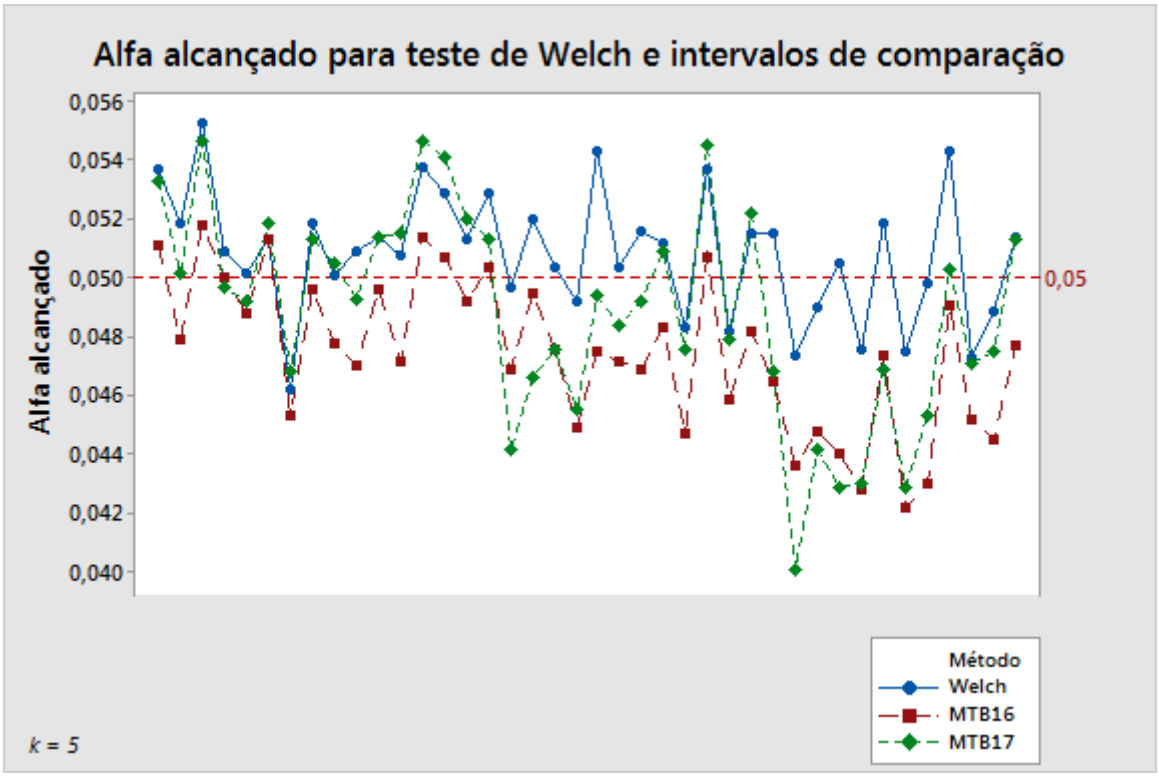


Figura 3 Teste de Welch comparado com dois métodos de calcular intervalos de comparação para 5 amostras

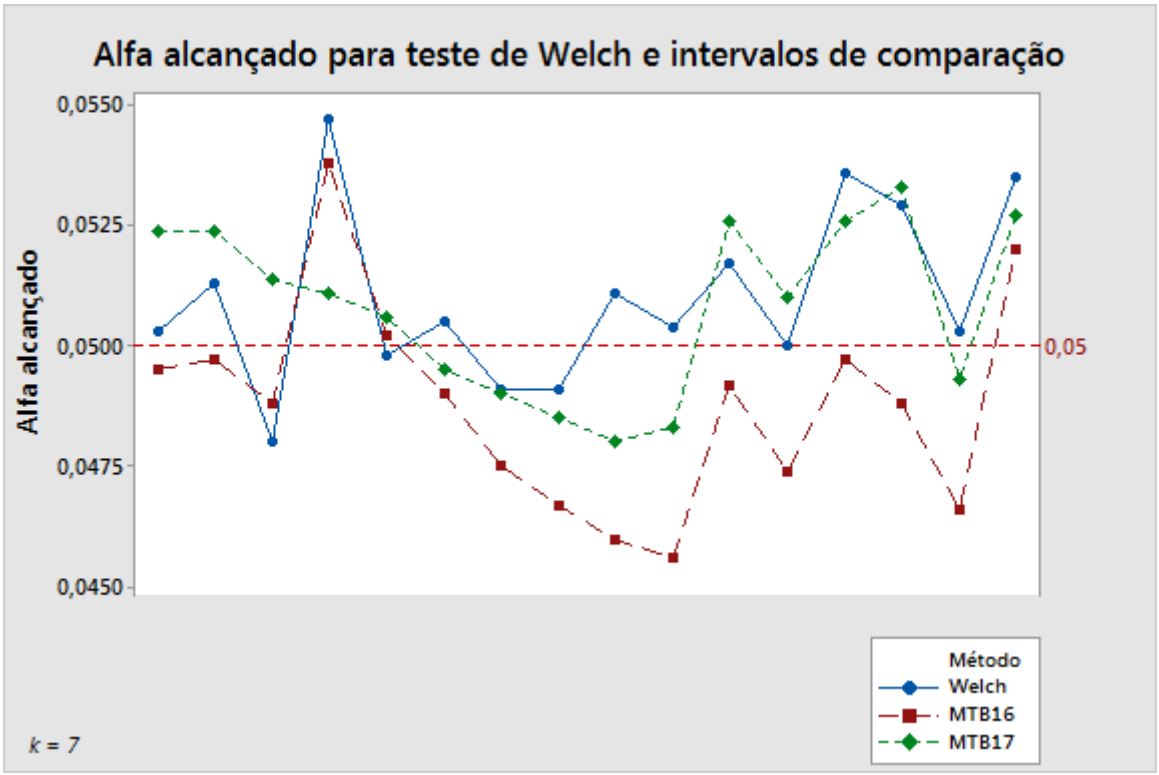


Figura 4 Teste de Welch comparado com dois métodos de calcular intervalos de comparação para 7 amostras

Esses resultados mostram valores alfa simulados em uma amplitude estreita em torno do valor alvo de 0,05. Além disso, os resultados usando-se o método com base em Games-Howell implementados na versão 17 do Minitab são indiscutivelmente mais alinhados de perto com os resultados do teste de Welch que foi o método usado na versão 16 do Minitab.

Há evidências de que a probabilidade de cobertura de intervalos pode ser sensível aos desvios padrão diferentes. Mas a sensibilidade não é tão extrema como aquela do teste F. O gráfico a seguir ilustra essa dependência no caso de k = 5.

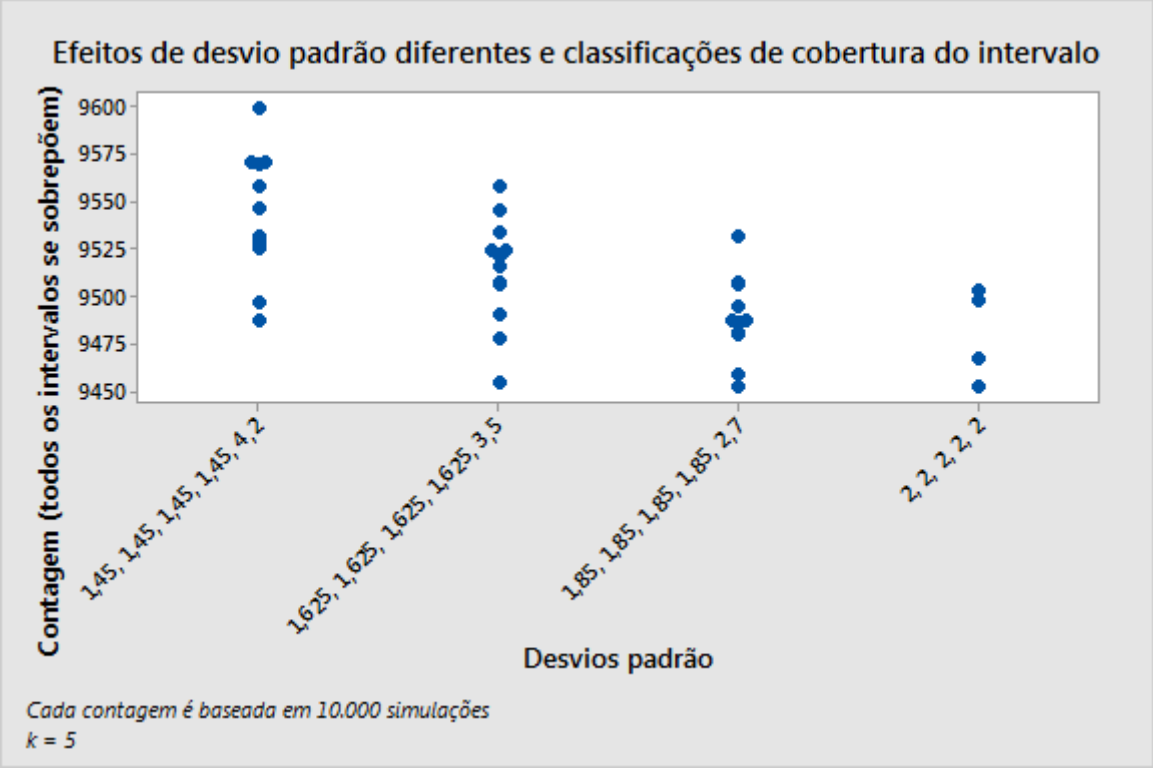


Figura 5 Resultados da simulação com desvios padrão diferentes

Usando o teste de hipótese e intervalos de comparação juntos

Em casos raros, é possível que o teste de hipóteses e a comparação não venham a concordar sobre rejeitar a hipótese nula. O teste pode rejeitar a hipótese nula, apesar de todos os intervalos de comparação ainda se sobreporem. Por outro lado, o teste pode deixar de rejeitar a hipótese nula, apesar de haver intervalos que não se sobrepõem. Essas divergências são raras porque ambos os métodos têm a mesma probabilidade de rejeitar a hipótese nula quando ela é real.

Quando isso acontece, consideramos primeiro os resultados do teste e usamos as comparações para investigação adicional, no caso de um teste significativo. Se o teste rejeitar a hipótese nula em um nível de significância α, qualquer intervalo de comparação que deixe de se sobrepor com pelo menos um outro será marcado em vermelho. Isso é usado como uma indicação visual de que a média de grupo correspondente difere pelo menos de uma outra. Ainda que todos os intervalos se sobreponham, o par com a menor

quantidade de sobreposição é colorido de vermelho se o teste for significativo para indicar a diferença "mais provável" (consulte a Figura 6 a seguir). Essa é uma escolha um tanto arbitrária, especialmente se existirem outros pares que têm muito pouca sobreposição. Mas nenhum outro par tem um limite na sua diferença que esteja mais próximo de zero.

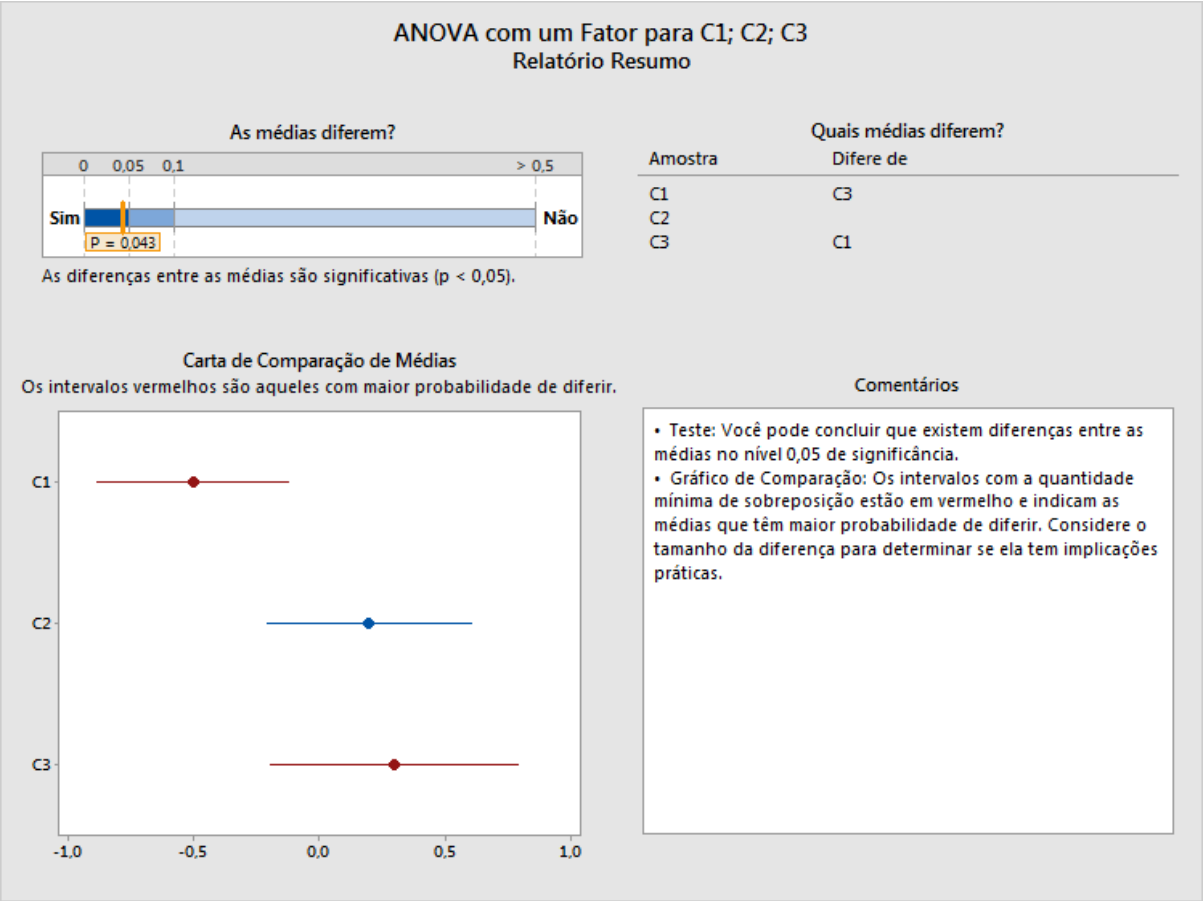


Figura 6 Teste significativo, intervalos marcados em vermelho mesmo quando eles se sobrepõem entre amostras

Se o teste não rejeitar a hipótese nula, então nenhum dos intervalos será marcado em vermelho, ainda que existam intervalos que não se sobrepõem (consulte a Figura 7 a seguir). Apesar de os intervalos sugerirem que existem diferenças entre as médias, lembre-se de que não rejeitar a hipótese nula não é o mesmo que concluir que a hipótese nula é real. Isso indica somente que as diferenças observadas não são grandes o bastante para descartar a chance como a causa. Vale observar também que o salto entre intervalos não sobrepostos será geralmente muito pequeno nesta situação, de modo que diferenças muito pequenas permanecem consistentes com os intervalos, e não indicam, necessariamente, que existem uma diferença com implicações práticas.

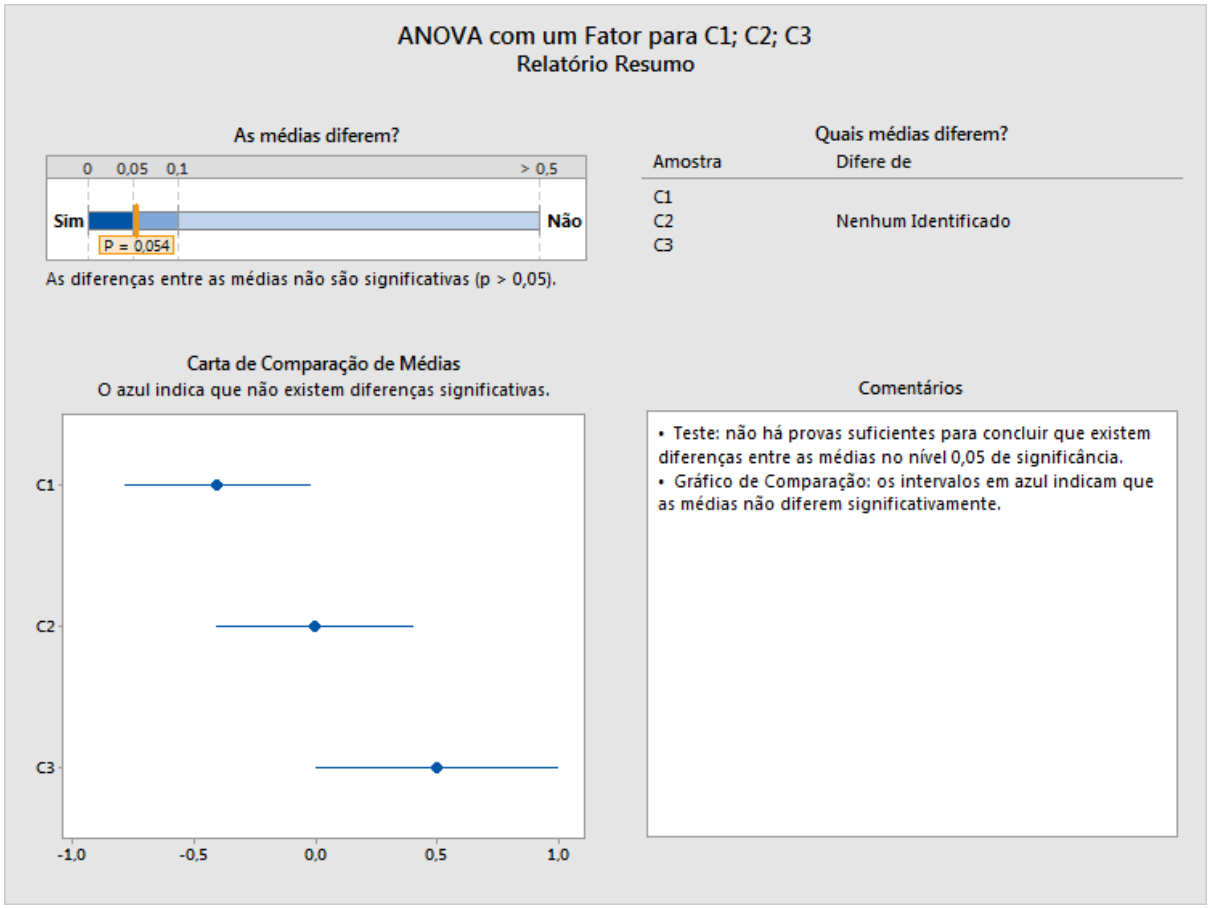


Figura 7 Teste falha, nenhum intervalo marcado em vermelho, mesmo quando não há sobreposições entre as amostras

Apêndice C: Tamanho amostral

No ANOVA para um fator, os parâmetros que estão sendo testados são as médias populacionais μ_1, μ_2, ... μ_k dos diferentes grupos ou populações. Os parâmetros satisfazem a hipótese nula se eles forem todos iguais. Se houver quaisquer diferenças entre as médias, elas satisfarão as hipóteses alternativas. A probabilidade de rejeitar a hipótese nula não deve ser maior do que α para médias que satisfazem a hipótese nula. As probabilidades reais dependem do desvio padrão das distribuições e do tamanho das amostras. O poder para detectar qualquer desvio das hipóteses nulas aumenta com desvios padrão menores ou amostras maiores.

Podemos calcular o poder do teste F sob a suposição de distribuições normais com desvios padrão iguais usando uma distribuição F não central. O parâmetro de não centralidade é:

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

em que μ é a média ponderada das médias:

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

e σ é o desvio padrão, que se supõe que seja constante. Todas as outras coisas sendo iguais, o poder aumenta com θ_F. Esse é o sentido preciso em que o poder aumenta conforme as médias desviam para ainda mais longe da hipótese nula.

Diferente do teste F, o teste Welch não tem uma fórmula exata simples do poder. Mas examinaremos duas fórmulas aproximadas, razoavelmente boas. A primeira usa uma distribuição F não central de uma maneira similar ao poder do teste F. O parâmetro de não centralidade que será usado ainda é da forma:

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

em que μ é a média ponderada:

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

mas os pesos irão depender dos desvios padrão e também dos tamanhos amostrais, ou seja, $w_i = \; n_i / \sigma_i^2$ ou $w_i = \; n_i / s_i^2$, dependendo se estivermos simulando os resultados para desvios padrão desconhecidos σ_i^2 ou estimando o poder, com base em desvios padrão da amostra s_i^2. O poder aproximado é calculado como:

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

em que os graus de liberdade do denominador são

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

Conforme mostrado a seguir, isso fprnece aproximações razoavelmente boas para o poder observado nas simulações. E apesar de usarmos uma aproximação diferente para calcular o poder no menu do Assistente, este fornece boa compreensão, e a base para selecionar a configuração de médias na qual calculamos o poder no menu do Assistente.

Configuração de médias

Ao manter a abordagem usada para poder e tamanho amostral no Minitab (Stat > ANOVA > Um fator), o Assistente não pede ao usuário um conjunto completo de médias nas quais poderá avaliar o poder. Em vez disso, ele pede ao usuário uma diferença entre médias que tem implicações práticas. Para uma dada diferença, existe um número infinito de configurações possíveis de médias nas quais as médias maiores e menores diferem naquela quantidade. Por exemplo, todos os seguintes têm uma diferença máxima de 10 dentre um conjunto de cinco médias:

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

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

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

e há infinitamente muitos mais.

Seguimos a abordagem usada para poder e tamanho amostral no Minitab (Stat > Poder e tamanho amostral > ANOVA para um fator), a saber, escolhendo um caso onde todos, exceto duas das médias estão na média (ponderada) das médias, e as duas médias restantes diferem da quantidade declarada. Contudo, devido à possibilidade de variâncias e tamanhos amostrais diferentes, o parâmetro de não centralidade (e, portanto, o poder) ainda dependem de quais duas médias supostamente diferem.

Considere a configuração das médias μ1, … , μk nas quais todas, exceto duas das médias são iguais à média ponderada geral μ, e duas médias, digamos μi > μj, diferem entre si e da média geral. Permita $\Delta = \mu_i - \mu_j$ denote a diferença entre as duas médias. Permita $\Delta_i = \mu_i - \mu$ e $\Delta_j = \mu - \mu_j$. Por isso, $\Delta = \Delta_i + \Delta_j$. Além disso, como μ representa a média ponderada de todas as médias k e supõe-se que (k – 2) das médias sejam iguais a μ temos:

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

Por isso:

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

e portanto,

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

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

Por esta determinada configuração de médias, podemos calcular o parâmetro de não centralidade relacionada ao teste 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 quantidade está aumentando em wi para wj fixo e vice-versa. Portanto, ela é maximizada no par
(i, J) com os dois maiores pesos e minimizada no par com os dois menores pesos. Todos os cálculos de poder consideram esses dois casos extremos, que maximizam e minimizam o poder sob a suposição de que exatamente duas médias diferem da média ponderada geral das médias.

Se você especificar uma diferença para o teste, os valores de poder mínimo e máximo são avaliados para esta diferença. A amplitude desses poderes é indicada nos relatórios relativos a uma barra colorida nas quais os poderes iguais ou abaixo de 60% estão em vermelho, os poderes iguais ou acima de 90% estão em verde e os poderes entre 60% e 90% estão em amarelo. Os resultados do Cartão de Relatórios dependem de onde a amplitude dos poderes cai em relação a essa escala codificada em cores. Se toda a amplitude estiver em vermelho, então o poder de qualquer par de grupos é menor do que ou igual a 60%, e o ícone vermelho aparece no cartão de relatório para indicar um problema de poder insuficiente. Se toda a amplitude estiver em verde, o poder de qualquer grupo é de, pelo menos 90% e o ícone verde no Cartão de Relatórios indica a condição do poder suficiente. Todas as outras condições são tratadas como situações intermediárias indicadas por um ícone amarelo no Cartão de Relatórios.

Em casos onde a condição verde não é alcançada, o Assistente calcula um tamanho amostra que poderia levar à condição verde dada a diferença especificada pelo usuário e os desvios padrão amostrais observados. O poder estimado depende dos tamanhos amostrais via os pesos $w_i = n_i/s_i^2$.. Se supõe-se que todas as amostras têm que ter o mesmo tamanho amostral, os dois menores pesos correspondem aos dois grupos com os maiores desvios padrão de amostra. O Assistente encontra um tamanho amostral que dá o poder de, no mínimo, 90% se a diferença especificada estiver entre os dois grupos com a maior variabilidade. Portanto, traçar um tamanho amostral, no mínimo com esse tamanho para todos os grupos iria resultar na amplitude completa dos valores de poder sendo, no mínimo 90%, o que satisfaz a condição verde.

Se o usuário não especificar uma diferença para o cálculo do poder, o Assistente encontra a maior diferença na qual o máximo da amplitude dos poderes calculados seria 60%. Esse valor é rotulado no limite entre as seções vermelha e amarela da barra, correspondendo a 60% do poder. Ele também encontra a menor diferença na qual o mínimo da amplitude de poderes calculados seria 90%. Esse valor é rotulado no limite entre as seções amarela e verde da barra, correspondendo a 90% do poder.

Cálculo do poder

O poder é calculado usando-se a aproximação devida a Kulinskaya et al. (2003):

Defina:

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

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

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

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

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

Os três primeiros cumulantes do numerador $\sum_{i=1}^{k} w_i (\bar{x}_i - \hat{\mu})^2$ da estatística de Welch podem ser estimados como:

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

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

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

Permita que $F_{k-1, f, 1-\alpha}$ denote o (1 – α) quantil da distribuição F(k – 1, f). Lembre-se de que $W^* \geq F_{k-1, f, 1-\alpha}$ é o critério para rejeição da hipótese nula em um teste de Welch tamanho α.

Permita

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

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

$c = \kappa_3 / (4\kappa_2)$ [Observe: a expressão para c é conhecida em Kulinskaya et al. (2003) sem os parênteses.]

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

Então o poder aproximado estimado do teste de Welch é:

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

em que χ_ν^2 é uma variável aleatória qui-quadrado com ν graus de liberdade.

Os seguintes resultados comparam o poder dos dois métodos de aproximação e o poder simulado para uma faixa de exemplos, com base em 10.000 simulações.

Tabela 3 Cálculos de poder para os dois métodos de aproximação comparados ao poder simulado

Exemplo	Alfa	Poder simulado	F não 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

Exemplo	Alfa	Poder simulado	F não central	Kulinskaya et al.
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, -0,862, 4,138	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
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	1,00
σ's: 2, 2, 2, 2, 2, 2, 5	0,05	0,9895	0,996653	1,00
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	1,00
σ's: 2, 2, 2, 2, 2, 2, 5	0,05	0,9995	0,999926	1,00
n's: 20, 20, 20, 20, 20, 20, 10	0,01	0,9943	0,99891	1,00
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, -0,5, 0,5	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

Exemplo	Alfa	Poder simulado	F não central	Kulinskaya et al.
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	0,33443
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	0,7817	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	1,00	0,999975	1,00
σ's: 2, 2, 2, 2, 2, 2, 2	0,05	0,9997	0,99987	1,00
n's: 12, 12, 12, 12, 12, 12, 12	0,01	0,9959	0,997927	0,99961
μ's: 0, 0, 0, 0, 0, -3,5, 3,5	0,10	1,00	1,00	1,00
σ's: 2, 2, 2, 2, 2, 2, 2	0,05	1,00	1,00	1,00
n's: 12, 12, 12, 12, 12, 12, 12	0,01	0,99998	0,99995	1,00
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	0,87562	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	0,9561	0,959796
n's: 12, 12, 12	0,01	0,8251	0,830726	0,823624

Exemplo	Alfa	Poder simulado	F não central	Kulinskaya et al.
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	1,00
σ's: 2, 2, 2	0,05	0,9997	0,999634	1,00
n's: 12, 12, 12	0,01	0,9949	0,994725	0,99909
μ's: 0, -3, 3	0,10	1,00	1,00	1,00
σ's: 2, 2, 2	0,05	1,00	1,00	1,00
n's: 12, 12, 12	0,01	0,9999	0,99985	1,00
μ's: 0, -3,5, 3,5	0,10	1,00	1,00	1,00
σ's: 2, 2, 2	0,05	1,00	1,00	1,00
n's: 12, 12, 12	0,01	0,9999	1,00	1,00
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	0,27424	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, -0,50, 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	0,848	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

Exemplo	Alfa	Poder simulado	F não central	Kulinskaya et al.
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	1,00
σ's: 2, 2, 4	0,05	0,989	0,99222	1,00
n's: 14, 12, 8	0,01	0,9222	0,940972	0,96383
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

Os resultados acima são resumidos no gráfico abaixo, que mostra as discrepâncias entre cada aproximação e valor do poder estimado por simulação.

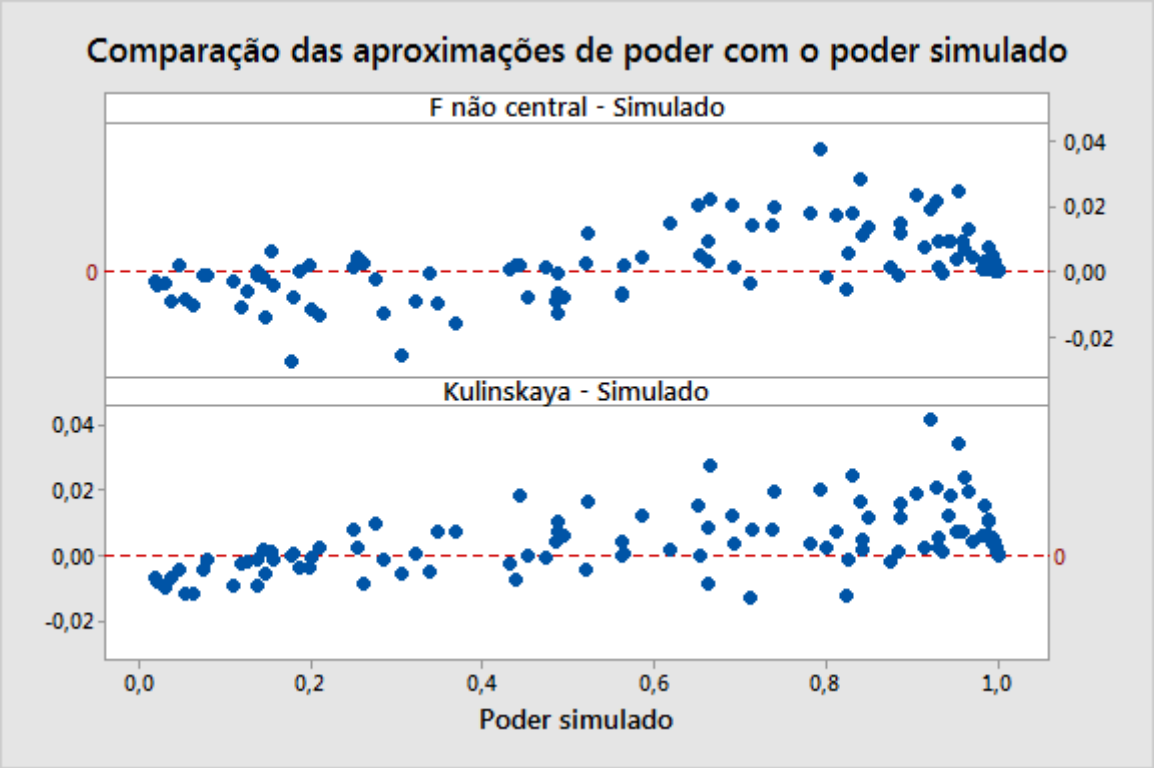


Figura 8 Comparisão de duas aproximações de poder e o poder estimado pela simulação

Apêndice D: Normalidade

Nesta seção, apresentamos as simulações que examinam o desempenho do teste de Welch e os intervalos de comparação com amostras de tamanho de pequeno a moderado de diversas distribuições normais.

As tabelas a seguir resumem os resultados de simulação para diferentes tipos de distribuições sob a hipótese nula de médias iguais. Para esses exemplos, todos os desvios padrão também são iguais e todas as amostras são de tamanho igual. O número de amostras é k = 3, 5 ou 7.

Cada célula mostra a estimativa do erro Tipo I com base em 10.000 simulações. O nível de significância alvo (alvo α) é 0,05.

Tabela 4 Os resultados da simulação do teste de Welch com média igual para distribuições diferentes

	Tamanho amostral n = 10			**Tamanho amostral n = 15**		
Distribuição	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
Qui-quadrado(3)	0,0473	0,0579	0,0753	0,0499	0,0588	0,0703
Qui-quadrado(5)	0,0458	0,0594	0,0643	0,0504	0,0606	0,0679
Qui-quadrado(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

As taxas de erros do Tipo I estão todas dentro de 3 pontos percentuais do alvo α mesmo com amostras de tamanho 10. Desvios maiores tendem a ocorrer com mais grupos e com distribuições que estão distantes do normal. Em tamanhos amostras de 10, os únicos casos em que a probabilidade de aceitação estava desativada em mais de 2 pontos percentuais são para k = 7. Eles ocorrem para a distribuição uniforme, que tem caudas mais curtas do que o normal, e para as distribuições exponenciais altamente assimétricas, qui-quadrado(3) e beta(8, 1). Aumentar os tamanhos amostrais para 15 marcadamente aprimora os resultados para a distribuição uniforme, mas não para as distribuições altamente assimétricas.

Realizamos uma simulação similar para intervalos de comparação. O simulado α neste caso é o número de simulações em 10.000 no qual alguns intervalos não se sobrepõem. O alvo $\alpha = 0{,}05$.

Tabela 5 Os resultados da simulação de intervalos de comparação com médias iguais para distribuições diferentes

	Tamanho amostral n = 10			**Tamanho amostral n = 15**		
Distribuição	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
Qui-quadrado(3)	0,0375	0,0305	0,0296	0,0384	0,0359	0,0339
Qui-quadrado(5)	0,0405	0,0390	0,0353	0,0417	0,0433	0,0416
Qui-quadrado(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

Desvios maiores tendem a ocorrer com mais amostras e com distribuições que estão distantes do normal. Em tamanhos amostrais de 10, as taxas de erro estão algumas vezes distantes em mais de 2 pontos percentuais para k = 7 (e em um caso, para k = 5). Esses casos ocorrem para a distribuição t de cauda extremamente pesada com 3 graus de liberdade, a distribuição Laplace e as distribuições exponencial altamente assimétrica e Qui-quadrado (3). Aumentar os tamanhos amostras para 15 aprimoram os resultados, deixando somente as distribuições t(3) e exponencial com valores α simulados que estão distantes do alvo em mais de 2 pontos percentuais. Observe que diferente dos resultados para o teste de Welch, os maiores desvios para intervalos de comparação estão no lado conservador.

O ANOVA para um fator no Assistente permite amostras k = 12, portanto, em seguida consideramos os resultados para mais de 7 amostras. A tabela a seguir mostra as taxas de erro do Tipo I usando o teste de Welch para dados não-normais em grupos k = 9. Novamente, o alvo $\alpha = 0{,}05$.

Tabela 6 Os resultados da simulação do teste de Welch para distribuições diferentes com 9 amostras

Distribuição	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
Qui-quadrado(3)	0,0774
Qui-quadrado(5)	0,0686
Qui-quadrado(10)	0,0581
Beta(8, 1)	0,0863

Como poderia ser esperado, as distribuições altamente assimétricas mostram os maiores desvios do alvo α. Mesmo assim, nenhuma das taxas de erro se desviam do alvo em mais de 4 pontos percentuais, apesar de o desvio para a distribuição exponencial estar próximo. O Cartão de Relatórios trata amostras de tamanho 15 o suficiente para não sinalizar um problema para dados não-normais porque todos os resultados estão, no mínimo, razoavelmente próximos do alvo α.

Amostras de tamanho n = 15 não apresentam desempenho tão bom quando temos amostras k = 12. A seguir consideramos os resultados simulados para o teste de Welch para uma amplitude de tamanhos amostrais usando distribuições extremamente não-normais, que irão nos ajudar no desenvolvimento de um critério razoável para o tamanho amostral.

Tabela 7 Os resultados da simulação do teste de Welch para distribuições diferentes com 12 amostras

n	T(3)	Uniforme	Qui-quadrado(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 essas distribuições n = 15 é aceitável se estivermos dispostos a aceitar um desvio de ligeiramente mais de 2 pontos percentuais do alvo α. Para manter o desvio abaixo de 2 pontos percentuais o tamanho amostral seria 20. Agora, consideramos os resultados das distribuições qui-quadrado (3) e exponenciais mais assimétricas.

Tabela 8 Os resultados da simulação do teste de Welch para distribuições qui-quadrado e exponenciais com 12 amostras

n	Qui-quadrado(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

Essas distribuições altamente assimétricas apresentam mais que um desafio. Se estivermos dispostos a aceitar um desvio bem acima de 3 pontos percentuais do alvo $\alpha = 0{,}05$, n = 15 poderia ser considerado suficiente mesmo para a distribuição qui-quadrado (3), mas a distribuição exponencial iria exigir algo mais perto de n = 30. Apesar de o critério de um tamanho amostral específico ser um tanto arbitrário, e que n = 20 funciona muito bem para uma ampla faixa de distribuições e marginalmente bem para distribuições extremamente assimétricas, usamos n = 20 como o tamanho amostral mínimo recomendado para 10 a 12 amostras. Claramente, se houver uma necessidade de manter o desvio pequeno mesmo para distribuições extremamente assimétricas, amostras maiores são recomendadas.