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.

Teste para desvios padrão (duas ou mais amostras)

Visão geral

O Assistente do Minitab inclui duas análises para comparar amostras independentes para determinar se sua variabilidade difere significativamente. O teste Desvio padrão para 2 amostras compara os desvio padrão de 2 amostras, e o teste Desvios padrão compara os desvios padrão de mais de 2 amostras. Neste artigo, nos referimos a experimentos de amostra k com k = 2 como experimentos de 2 amostras e experimentos de amostra k com k > 2 como experimentos de várias amostras. Geralmente, esses dois tipos de experimentos são estudados separadamente (consulte o Apêndice A).

Como o desvio padrão é a raiz quadrada da variância, um teste de hipótese que compara desvios padrão é equivalente a um teste de hipótese que compara variâncias. Diversos métodos estatísticos foram desenvolvidos para comparar as variâncias de duas ou mais populações. Dentre esses testes, o teste de Levene/Brown-Forsythe é um dos mais robustos e mais utilizados. Entretanto, o desempenho do poder do teste de Levene/Brown-Forsythe é menos satisfatório do que suas propriedades de erro do Tipo I em experimentos de 2 amostras. Pan (1999) mostra que para algumas populações, incluindo a população normal, o poder do teste em experimentos de 2 amostras, tem um limite superior que pode estar bem abaixo de 1 independentemente da magnitude da diferença entre os desvios padrão. Em outras palavras, para esses tipos de dados, é mais provável que o teste conclua que não há diferença entre os desvios padrão, independentemente de quão grande seja a diferença. Por esses motivos, o Assistente usa um novo teste, o teste de Bonett, para o teste Desvio padrão para 2 amostras. Para o teste de desvios padrão com experimentos de várias amostras, o Assistente use um procedimento de comparações múltiplas (MC).

O teste de Bonett (2006), uma versão modificada do teste de Layard (1978) da igualdade de duas variâncias, aprimora o desempenho do teste com pequenas amostras. Banga and Fox (2013A) derivam os intervalos de confiança associados ao teste de Bonett e mostram que

eles são tão exatos quanto os intervalos de confiança associados com o teste de Levene/Brown-Forsythe e que são mais precisos para a maioria das distribuições. Além disso, Banga e Fox (2013A) determinaram que o teste de Bonett é tão robusto quanto o teste de Levene/Brown-Forsythe e é mais poderoso para a maioria das distribuições.

O procedimento de comparações múltiplas (MC) inclui um teste geral da homogeneidade ou igualdade dos desvios padrão (ou variâncias) para comparações múltiplas, que são baseadas nos intervalos de comparação de cada par de desvios padrão. Os intervalos de comparação são derivados para que o teste de MC seja significativo se, e somente se, pelo menos um par dos intervalos de comparação não se sobrepuser. Banga and Fox (2013B) mostram que o teste de MC tem propriedades de erro Tipo I e Tipo II que são similares ao teste de Levene/Brown-Forsythe para a maioria das distribuições. Uma vantagem importante do teste de MC é a exibição gráfica dos intervalos de comparação, que fornecem uma ferramenta visual eficaz para identificação das amostras com desvios padrão diferentes. Quando há somente duas amostras no experimento, o teste de MC é equivalente ao teste de Bonett.

Neste artigo, avaliamos a validade do teste de Bonett e o teste de MC para diferentes distribuições de dados e tamanhos amostrais. Além disso, investigamos a análise do poder e do tamanho amostral para o teste de Bonett, que é baseado em um método de aproximação de amostras grandes. Com base nesses fatores, desenvolvemos as seguintes verificações que o Assistante automaticamente realiza em seus dados e exibe no Cartão de relatório:

- Dados incomuns
- Normalidade
- Validade do teste
- Tamanho amostral (somente teste de Desvio padrão para 2 amostras)

Testes para métodos de desvios padrão

Validade do teste de Bonett e do teste de MC

Em seu estudo comparativo dos testes de variâncias iguais, Conover, et al. (1981) descobriram que o teste de Levene/Brown-Forsythe estava entre os testes de melhor desempenho, com base em suas taxas de erro do Tipo I e do Tipo II. Desde aquela época, outros métodos foram propostos para teste de variâncias iguais em experimentos para 2 amostras e para várias amostras (Pan, 1999; Shoemaker, 2003; Bonett, 2006). Por exemplo, Pan mostra que a despeito de sua robustez e simplicidade de interpretação, o teste de Levene/Brown-Forsythe não tem poder suficiente para detectar diferenças importantes entre 2 desvios padrão quando as amostras se originam das mesmas populações, incluindo a população normal. Devido a essa limitação crítica, o Assistente usa o teste de Bonett para o teste Desvio padrão para 2 amostras (consulte o Apêndice A ou Banga and Fox, 2013A). Para o teste de desvios padrão com mais de 2 amostras, o Assistente usa um procedimento de MC com intervalos de comparação que fornece uma exibição gráfica para identificar amostras com diferentes desvios padrão quando o teste de MC é significativo (consulte o Apêndice A e Banga and Fox, 2013B).

Objetivo

Primeiro, queríamos avaliar o desempenho do teste de Bonett ao comparar desvios padrão de duas populações. Em segundo, queríamos avaliar o desempenho do teste de MC ao comparar os desvios padrão entre mais de duas populações. Especificamente, queríamos avaliar a validade desses testes, quando eles são realizados em amostras de diversos tamanhos de diferentes tipos de distribuições.

Método

Os métodos estatísticos usados para o teste de Bonett e o teste de MC são definidos no Apêndice A. Para avaliar a vailidade dos testes, precisamos examinar se suas taxas de erro do Tipo I permaneceram perto do nível alvo de significância (alfa) sob diferentes condições. Para fazer isto, realizamos um conjunto de simulações para avaliar a validade do teste de Bonett ao comparar os desvios padrão de 2 amostras independentes e outros conjuntos de simulações para avaliar a validade do teste de MC ao comparar os desvios padrão de múltiplas amostras independentes (k), quando k > 2.

Geramos 10.000 pares ou amostras aleatórias (k) múltiplas de diversos tamanhos de diversas distribuições, usando experimentos balanceados e não balanceados. Depois, realizamos um teste de Bonett bilateral para comparar os desvios padrão das 2 amostras ou realizamos um teste de MC para comparar os desvios padrão das amostras k em cada experimento, usando um nível de significância alvo de $\alpha = 0{,}05$. Contamos o número de vezes em 10.000 réplicas em que o teste rejeitou a hipótese nula (quando na realidade os desvios padrão reais foram iguais) e comparamos esta proporção, conhecida como nível de significância simulada, para

o nível de significância alvo. Se o teste for corretamente realizado, o nível de significância simulado, que representa a taxa de erro real do Tipo I, deve estar muito próxima ao nível de significância alvo. Para obter mais detalhes sobre os métodos específicos usados para as simulações para 2 amostras e para k amostras, consulte o Apêndice B.

Resultados

Para comparações para 2 amostras, as taxas de erro Tipo I simuladas do teste de Bonett estavam perto do nível alvo de significância, quando as amostras eram moderadas ou grandes em tamanho, independentemente da distribuição e independentemente se o experimento estava balanceado ou não balanceado. Entretanto, quando amostras pequenas foram extraídas de populações extremamente assimétricas, o teste de Bonett foi geralmente conservador, e tinha taxas de erro do Tipo I que eral ligeiramente inferiores ao nível alvo de significância (isto é, a taxa de erros do Tipo I alvo).

Para comparações de várias amostras, as taxas de erro do Tipo I do teste de MC estavam perto do nível alvo de significância quando as amostras eram moderadas ou grandes em tamanho, independentemente da distribuição e se o experimento era balanceado ou não balanceado. Para amostras pequenas e extremamente assimétricas, contudo, o teste foi geralmente menos conservador, e havia taxas de erro do Tipo I que eram maiores do que o nível alvo de significância quando o número de amostras no experimento é grande.

Os resultados dos nossos estudos foram consistentes com aqueles de Banga and Fox (2013A) and (2013B). Concluímos que o teste de Bonett e o teste de MC apresentam bom resultado quando o tamanho da menor amostra é de, no mínimo, 20. Portanto, usamos este requisito de tamanho amostral mínimo na Validade da verificação de teste no Cartão de Relatório do Assistente (consulte a seção Verificação de dados).

Intervalos de comparação

Quando um teste para comparar dois ou mais desvios padrão é estatisticamente significativo, indicando que, no mínimo, um dos desvios padrão é diferente dos outros, o próximo passo na análise é determinar quais amostras são estatisticamente diferentes. Uma maneira intuitiva de fazer esta comparação é representar graficamente os intervalos de confiança associados a cada amostra 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.

Objetivo

Queríamos desenvolver um método para calcular intervalos de comparação individuais que podem ser usados como ambos um teste geral da homogeneidade das variâncias e um método para identificar amostras com variâncias diferentes, quando o teste geral for significativo. Um requisito crítico para o procedimento de MC é que o teste geral é significativo se, e somente se, pelo menos um par dos intervalos de comparação não se sobrepuserem, o que indica que os desvios padrão de, no mínimo, duas amostras são diferentes.

Método

O procedimento de MC que usamos para comparar desvios padrão múltiplos é derivado de comparações múltiplas pareadas. Cada par de amostras é comparado usando-se o teste de Bonett (2006) de igualdade de desvios padrão de duas populações. As comparações pareadas usam uma correção de multiplicidade com base em uma aproximação de grandes amostras mostrado em Nakayama (2009). A aproximação de grandes amostras é preferível sobre a correção de Bonferroni normalmente usada porque a correção de Bonferroni torna-se cada vez mais conservadora conforme o número de amostras aumenta. Por fim, os intervalos de comparação resultam das comparações pareadas com base no melhor procedimento aproximado de Hochberg et al. (1982). Para obter detalhes, consulte o Apêndice A.

Resultados

O procedimento MC satisfaz o requisito de que o teste geral de igualdade de desvios padrão é significativo se, e somente se, pelo menos dois intervalos de comparação não se sobrepuserem. Se o teste geral não for significativo, todos os intervalos de comparação devem se sobrepor.

O Assistente exibe os intervalos de comparação no Gráfico de comparações de desvios padrão no Relatório de resumo. Próximo a este gráfico, o Assistente exibe o valor p do teste de MC, que é o teste geral para a homogeneidade dos desvios padrão. Quando o teste de desvios padrão é estatisticamente significativo, qualquer intervalo de comparação que não se sobrepõe com, no mínimo, um outro intervalo, é marcado em vermelho. Se o teste de desvios padrão não for estatisticamente significativo, nenhum dos intervalos é marcado em vermelho.

Desempenho do poder teórico (somente para experimentos para 2 amostras)

As funções de poder teórico dos testes de Bonett e de MC são necessárias para planejamento de tamanhos amostrais. Para experimentos para 2 amostras, uma função de poder teórico aproximado do teste pode ser derivada usando-se métodos teóricos para grandes amostras. Como esta função resulta de métodos de aproximação de grandes amostras, precisamos avaliar suas propriedades, quando o teste é conduzido usando-se pequenas amostras geradas de distribuições normais e não-normais. Ao comparar os desvios padrão de mais de dois grupos, contudo, a função de poder teórico do teste de MC não é obtida facilmente.

Objetivo

Queríamos determinar se poderíamos usar a função de poder teórico com base na aproximação de grandes amostras para avaliar os requisitos de poder e de tamanho amostral para o teste Desvio padrão para 2 amostras no Assistente. Para fazer isso, precisamos avaliar se a função de poder teórico reflete com exatidão o poder real alcançado pelo teste de Bonett quando ele é realizado nos dados de diversos tipos de distribuições, incluindo distribuições normais e não-normais.

Método

A função de poder teórico aproximado do teste de Bonett para experimentos para 2 amostras é derivada no Apêndice C.

Realizamos simulações para estimar os níveis de poder real (que nos referimos como níveis de poder simulado) usando o teste de Bonett. Primeiro, geramos pares ou amostras aleatórias de diversos tamanhos de diversas distribuições, incluindo distribuições normais e não-normais. Para cada distribuição, realizamos o teste de Bonett em cada um dos 10.000 pares de réplicas de amostras. Para cada par de tamanhos amostrais, calculamos o poder simulado do teste para detectar uma dada diferença como a fração das 10.000 pares de amostras para as quais o teste é significativo. Para comparação, também calculamos o nível de poder correspondente usando a função de poder teórico aproximado do teste. Se a aproximação funcionar bem, os níveis de poder teórico e simulado devem ser próximos. Para obter mais detalhes, consulte o Apêndice D.

Resultados

Nossas simulações mostraram que, para a maioria das distribuições, as funções de poder teórico e simulado do teste de Bonett são praticamente iguais para tamanhos amostrais pequenos e estão mais perto quando o tamanho amostral mínimo alcança 20. Para distribuições simétricas e praticamente simétricas, com caudas leves a moderadas, os níveis de poder teórico são ligeiramente mais altos do que os níveis de poder (real) simulado. Contudo, para distribuições assimétricas e distribuições de cauda pesada eles são menores do que os níveis de poder (real) simulado. Para obter mais detalhes, consulte o Apêndice D.

No geral, nossos resultados mostram que a função de pode teórico fornece uma boa base para planejamento de tamanhos amostrais.

Verificações dos dados

Dados incomuns

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. Os estudos de simulação mostram que quando os dados contêm outliers, o teste de Bonett e o teste de MC são conservadores (consulte o Apêndice B). Os níveis atuais de significância dos teste são marcadamente menores do que o nível alvo, particularmente quando a análise é realizada com pequenas amostras.

Objetivo

Queríamos desenvolver um método para verificar valores de dados que são muito grandes ou muito pequenos, em relação à amostra geral, e que podem afetar os resultados 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 25º e 75º 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.

Normalidade

Diferente da maioria dos testes de igualdade de variâncias, que são derivados sob a suposição de normalidade, o teste de Bonett e o teste de MC para igualdade de desvios padrão não fazem uma suposição sobre a distribuição específica dos dados.

Objetivo

Apesar de o teste de Bonett e do teste de MC serem baseados em métodos de aproximação de grandes amostras, queríamos confirmar que eles têm bom desempenho para dados normais e não-normais em pequenas amostras. Também queríamos informar ao usuário sobre como a normalidade dos dados se refere aos resultados dos testes de desvio padrão.

Método

Para avaliar a validade dos testes sob diferentes condições, realizamos simulações para examinar a taxa de erros do Tipo I do teste de Bonett e do teste de MC, com dados normais e não-normais de diversos tamanhos amostrais. Para obter mais detalhes, consulte a seção Testes para métodos de desvios padrão e o Apêndice B.

Resultados

Nossas simulações mostraram que a distribuição dos dados não tem um importante efeito nas propriedades de erros do Tipo I do teste de Bonett ou do teste de MC para amostras suficientemente grandes (tamanho amostral mínimo ≥ 20). Os testes produzem taxas de erros do Tipo I que são consistentemente próximas da taxa de erros de destino para ambos os dados normais e não-normais.

Com base nesses resultados, relativos à taxa de erros do Tipo I, o Assistente exibe as informações sobre a normalidade no Cartão de Relatório.

Para experimentos para 2 amostras, o Assistente exibe o seguinte indicador:

Status	Condição
	Esta análise usa o teste de Bonett. Com amostras suficientemente grandes, o teste apresenta bom desempenho para ambos os dados normais e não-normais.

Para experimentos para múltiplas amostras, o Assistente exibe o seguinte indicador:

Status	Condição
	Esta análise usa um Teste de comparações múltiplas. Com amostras suficientemente grandes, o teste apresenta bom desempenho para ambos os dados normais e não-normais.

Validade do teste

Nos Testes para a seção de métodos de desvios padrão, mostramos que para ambas as comparações de 2 amostras e múltiplas (k) amostras, o teste de Bonett e o teste de MC produzem taxas de erros do Tipo I para dados normais e não-normais, em experimentos

balanceados e não balanceados, quando as amostras forem de tamanho moderado ou grande. Contudo, quando as amostras forem pequenas, os testes de Bonett e de MC não apresentam, geralmente, bom desempenho.

Objetivo

Queríamos aplicar uma regra para avaliar a validade dos resultados do teste de desvio padrão para 2 amostras e para múltiplas (k) amostras, com base nos dados do usuário.

Método

Para avaliar a validade dos testes sob diferentes condições, realizamos simulações para examinar a taxa de erros do Tipo I do teste de Bonett e do teste de MC, com diversas distribuições de dados, números de amostras e tamanhos amostrais, conforme descritos anteriormente na seção Testes para métodos de desvios padrão. Para obter mais detalhes, consulte o Apêndice B.

Resultados

O teste de Bonett e o teste de MC apresentam bom resultado quando o tamanho da menor amostra é de, no mínimo, 20. Portanto, o Assistente exibe os seguintes indicadores de status no Cartão de Relatório para avaliar a validade dos testes de desvio padrão.

Status	Condição
	Os tamanhos amostrais são de, no mínimo, 20, portanto, o valor p deve ser exato.
	Alguns dos tamanhos amostrais são de, no mínimo, 20, portanto, o valor p pode não ser exato. Considere aumentar os tamanhos amostrais para, no mínimo, 20.

Tamanho amostral (somente para teste de Desvios padrão para 2 amostras)

Tipicamente, um teste de hipótese é realizado para coletar evidências para rejeitar a hipótese nula de “nenhuma diferença”. Se a amostra é muito pequena, o poder do teste pode não ser adequado para detectar uma diferença que realmente existe, que resulta em um erro do Tipo II. Portanto, é crucial assegurar que os tamanhos amostrais sejam suficientemente grandes para detectar diferenças praticamente importantes com alta probabilidade.

Objective

Se os dados não fornecerem evidência suficiente para rejeitar a hipótese nula, queríamos determinar se os tamanhos amostrais são grandes o suficiente para o teste para detectar diferenças práticas de interesse com alta probabilidade. Apesar de o objetivo do planejamento de tamanho amostral ser assegurar que os tamanhos amostrais sejam grandes o suficiente para detectar diferenças importantes com alta probabilidade, elas não devem ser tão grandes que diferenças inexpressivas tornem-se estatisticamente significativas com alta probabilidade.

Método

O poder e a análise do tamanho amostral do teste deFrdviod padrão para 2 amostras estão baseados em uma aproximação da função de poder do teste de Bonett, que normalmente fornece boas estimativas da função do poder real do teste (consulte os resultados da simulação resumidos em Desempenho da funçã de poder teórico na seção Método).

Resultados

Quando os dados não fornecem evidências suficientes contra a hipótese nula, o Assistente usa a função de de poder aproximado do teste de Bonett para calcular as diferenças práticas que podem ser detectadas com uma probabilidade de 80% e de 90% para o tamanho amostral dado. Além disso, se o usuário fornecer uma determinada diferença prática de interesse, o Assistente usa a função de poder do teste de aproximação normal para calcular os tamanhos amostrais que produzem uma chance de 80% e de 90% de detecção da diferença.

Para ajudar a interpretar os resultados, o Catão de Relatório do Assistente para o Teste de desvios padrão para 2 amostras exibe os seguintes indicadores de status ao verificar o poder e o tamanho amostrais:

Status	Condição
	O teste encontra uma diferença entre os desvios padrão, portanto, o poder não é um problema. OU O poder é suficiente. O teste não encontrou nenhuma diferença entre os desvios padrão, mas a amostra é grande o suficiente para fornecer pelo menos 90% de chance de detectar a determinada diferença.
	O poder pode ser suficiente. O teste não encontrou uma diferença entre os desvios padrão, mas a amostra é grande o suficiente para fornecer 80 a 90% de chance de detectar a determinada diferença. O tamanho amostral necessário para atingir 90% de poder é informado.
	Talvez o poder não seja suficiente. O teste não encontrou uma diferença entre os desvios padrão, mas a amostra é grande o suficiente para fornecer 60 a 80% de chance de detectar a determinada diferença. Os tamanhos amostrais necessários para atingir 80% e 90% de poder são informados.
	O poder não é suficiente. O teste não encontrou nenhuma diferença entre os desvios padrão, mas a amostra não é grande o suficiente para fornecer pelo menos 60% de chance de detectar a determinada diferença. Os tamanhos amostrais necessários para atingir 80% e 90% de poder são informados.
	O teste não encontrou uma diferença entre os desvios padrão. Você não especificou uma diferença prática para detecção; portanto, o relatório indica as diferenças que você pôde detectar com 80% e de 90% de chance, com base no seu alfa e tamanho amostral.

Referências

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

Banga, S.J. and Fox, G.D. (2013A). Em Robust Confidence Interval for a Ratio of Standard Deviations de Bonett. *White paper, Minitab Inc*.

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

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

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

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

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

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

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

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

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

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

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

Apêndice A: Método para o teste de Bonett e o teste de Múltipla comparação

As suposições essenciais para fazer inferências sobre os desvios padrão ou variâncias usando o método de Bonett (experimentos para 2 amostras) ou o procedimento de múltiplas comparações (MC) (experimentos de múltiplas amostras) pode ser descrito da seguinte forma. Permitir que $X_{11}, \dots, X_{1n_1}, \dots, X_{k1}, \dots, X_{kn_k}$ seja as amostras aleatórias independentes k ($k \geq 2$), com cada amostra traçada a partir de uma distribuição com média μ_i e variância σ_i^2 desconhecidas, respectivamente, para $i = 1, \dots, k$. Vamos supor que as distribuições pai das amostras têm uma curtose finita comum $\gamma = E(Y - \mu)^4/\sigma^4 < \infty$. Apesar desta suposição ser decisiva para as derivações teóricas, ela não é crítica para a maioria das aplicações práticas, onde as amostras são suficientemente grandes (Banga and Fox, 2013A).

Método A1: Teste de Bonett de igualdade de duas variâncias

O teste de Bonett só se aplica a experimentos para 2 amostras onde duas variâncias ou desvios padrão são comparados. O teste é uma versão modificada do teste de Layard (1978) de igualdade de variâncias em experimentos de duas amostras. Um teste de Bonett bilateral da igualde de duas variâncias com nível de significância de α rejeita a hipótese nula da igualdade se, e somente se,

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

em que:

S_i é o desvio padrão amostral da amostra i

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

$z_{\alpha/2}$ refere-se ao percentil superior $\alpha/2$ da distribuição normal padrão

$\hat{\gamma}_P$ é o estimador da curtose combinada fornecida como:

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

Na expressão do estimador da curtose combinada, m_i é a média aparada da amostra i, com a proporção aparada, $1/[2(n_i - 4)^{1/2}]$.

No acima descrito, a constante c é incluída como um pequeno ajuste de amostra para reduzir o efeito de probabilidades de erro de cauda diferente em experimentos não balanceados. Esta constante é fornecida como $c = c_1/c_2$, em que

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

Se o experimento é balanceado, isto é, se $n_1 = n_2$, então o valor p do teste é obtido como

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

em que Z é uma variável aleatória distribuída como distribuição normal padrão e z o valor observado das seguintes estatísticas com base nos dados em mãos. A estatística é

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

em que

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

Por outro lado, se o experimento for não balanceado, o valor p do teste é obtido como

$$P = 2\mathrm{min}(\alpha_L, \alpha_U)$$

em que $\alpha_L = \mathrm{Pr}(Z > z_L)$ e $\alpha_U = \mathrm{Pr}(Z > z_U)$. A variável z_L é a menor raiz da função

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

e z_U é a menor raiz da função $L(z, S_2, S_1, n_2, n_1)$.

Método A2: Teste de múltipla comparação e intervalos de comparação

Suponha que haja k $(k \geq 2)$ grupos independentes ou amostras. Nosso objetivo era construir um sistema de intervalos k para os desvios padrão da população de tal forma que o teste de igualdade dos desvios padrão seja significativo se, e somente se, no mínimo dois dos intervalos k não se sobreponham. Esses intervalos são referidos como intervalos de comparação. Este método de comparação é similar aos procedimentos de múltiplas comparações das médias em modelos de ANOVA para um fator, que foram inicialmente desenvolvidos por Tukey-Kramer e depois generalizados por Hochberg, et al. (1982).

Comparando dois desvios padrão

Para experimentos de 2 amostras, os intervalos de confiança da razão dos desvios padrão associados ao teste de Bonett podem ser calculados diretamente para avaliar o tamanho da diferença entre os desvios padrão (Banga and Fox, 2013A). Na realidade, usamos esta abordagem para Estat > Estatísticas básicas > 2 variâncias na versão 17 do Minitab. No Assistente, contudo, queríamos fornecer intervalos de comparação que são mais fáceis de interpretar do que o intervalo de confiança da razão dos desvios padrão. Para fazer isso, usamos o procedimento de Bonett descrito no Método A1 para determinar os intervalos de comparação para duas amostras.

Quando há duas amostras, o teste de Bonett da igualdade das variâncias é significativo se, e somente se, o seguinte intervalo de aceitação associado com o teste de Bonett de igualdade das variâncias não contiver 0.

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

em que as estimativas da curtose combinada $\hat{\gamma}_P$ e $g_i, i = 1, 2$ são previamente fornecidas.

A partir deste intervalo, deduzimos os seguintes dois intervalos de comparação de tal forma que o teste de igualdade de variâncias ou do desvio padrão seja significativo se, e somente se, eles não se sobrepuserem. Esses dois intervalos são

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

em que

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

Usar esses intervalos como um procedimento de teste de igualdade do desvio padrão é equivalente ao teste de Bonett de igualdade dos desvios padrão. Especificamente, os intervalos não se sobrepõem se, e somente se, o teste de Bonett de igualdade dos desvios padrão for significativo. Observe, contudo, que esses intervalos não são intervalos de confiança de desvios padrão, mas são somente apropriados para múltiplas comparações de desvios padrão. Hochberg et al. referem-se a intervalos similares para comparação de médias como intervalos de incertezas pelo mesmo motivo. Nós nos referimos a esses intervalos como intervalos de comparação.

Devido ao procedimento de intervalos de comparação ser equivalente ao teste de Bonett de igualdade dos desvios padrão, o valor p associado aos intervalos de comparação é idêntico ao valor p do teste de Bonett de igualdade de dois desvios padrão, descrito anteriormente.

Comparando múltiplos desvios padrão

Quando há mais de dois grupos ou amostras, os k intervalos de comparação são deduzidos de $k(k-1)/2$ testes simultâneos pareados de igualdade dos desvios padrão com nível de significância por família α. Mais especificamente, permita que $X_{i1}, \ldots, X_{in_i}$ e $X_{j1}, \ldots, X_{jn_j}$ sejam os dados amostrais para qualquer par (i, j) de amostras. Similar ao caso de 2 amostras, o teste de igualdade dos desvios padrão do determinado par (i, j) de amostras é significativo em algum α' nível se, e somente se, o intervalo

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

não contiver 0. No $\hat{\gamma}_{ij}$ acima está o estimador de curtose combinada com base no par (i, j) de amostras e é fornecido como

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

Além disso, conforme definido anteriormente, m_i é a média aparada da amostra i, com a proporção aparada, $1/[2(n_i - 4)^{1/2}]$ e

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

Como há $k(k-1)/2$ testes pareados simultâneos, o nível α' deve ser escolhido de forma que a taxa de erros por família real fique perto do nível alvo de significância α. Um ajuste possível é baseado na aproximação de Bonferroni. Contudo, as correções de Bonferroni são bem conhecidas por serem cada vez mais conservadoras conforme o número de amostras no experimento aumenta. Uma abordagem melhor é baseada em uma aproximação normal fornecida por Nakayama (2008). Usar essa abordagem simplesmente substitui $z_{\alpha'/2}$ por $q_{\alpha,k}/\sqrt{2}$, em que $q_{\alpha,k}$ é o ponto α superior do intervalo de k variáveis aleatórias normais padrão independentes e identicamente distribuídas; que é

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

onde $Z_1, \ldots, Z_k$ são variáveis aleatórias normais padrão independentes e identicamente distribuídas.

Além disso, usar um método similar ao Hochberg et al. (1982), o procedimento que se aproxima da melhor forma do procedimento pareado descrito acima, rejeita a hipótese nula da igualdade dos desvios padrão se, e somente se, para algum par (i, j) de amostras

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

em que V_i é escolhido para minimizar a quantidade

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

com

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

A solução deste problema, conforme ilustrado em Hochberg et al. (1982) é escolher

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

Ocorre que o teste baseado no procedimento aproximado é significativo se, e somente se, pelo menos um par dos seguintes k intervalos não se sobrepuserem.

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

Para calcular o valor p geral associado com o teste MC, permitimos que P_{ij} seja o valor p associado a qualquer par (i, j) de amostras. Segue-se então que o valor p geral associado ao teste de múltiplas comparações é

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

Para calcular P_{ij} realizamos o algoritmo do experimento para 2 amostras fornecido no Método A1 usando

$$se = V_i + V_j$$

em que V_i é fornecido acima.

Mais especificamente, se $n_i \neq n_j$

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

em que $\alpha_L = \Pr(Q > z_L\sqrt{2})$, $\alpha_U = \Pr(Q > z_U\sqrt{2})$, a variável z_L é a menor raiz da função $L\big(z, S_i, S_j, n_i, n_j\big)$, a variável z_U é a menor raiz da função $L\big(z, S_j, S_i, n_j, n_i\big)$ e Q é uma variável aleatória que tem a distribuição da amplitude como anteriormente definido.

Se $n_i = n_j$ então $P_{ij} = \Pr(Q > |z_o|\sqrt{2})$ em que

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

Apêndice B: Validade do teste de Bonett e o teste múltiplas comparações

Simulação B1: Validade do teste de Bonett (modelos para 2 amostras, experimentos balanceados e não balanceados)

Geramos pares de amostras aleatórias que são de pequenas a moderadas em tamanho de distribuições com propriedades diferentes. A distribuições incluíram:

- Distribuição normal padrão (N(0, 1))
- Distribuições simétricas e de cauda leve, incluindo a distribuição uniforme (U(0, 1)) e a distribuição Beta com ambos os parâmetros definidos como 3 (B(3, 3))
- Distribuições simétricas e de cauda pesada, incluindo distribuições t com 5 e 10 graus de liberdade (t(5),t(10)) e a distribuição Laplace com locação 0 e escala 1 (Lpl))
- Distribuições assimétricas e com cauda pesada, incluindo a distribuição exponencial com escala 1 (Exp) e distribuições qui-quadrado com 5 e 10 graus de liberdade (qui(5), qui(10))
- Distribuição assimétrica à esquerda e com cauda pesada; especificamente, a distribuição Beta com os parâmetros definidos como 8 e 1, respectivamente (B(8,1))

Além disso, para avaliar o efeito direto dos outliers, geramos pares das amostras de distribuições normais contaminadas definidas como

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

em que p é o parâmetro de combinação e $1 - p$ é a proporção de contaminação (que iguala a proporção de outliers). Selecionamos duas populações normais contaminadas para o estudo: $CN(0{,}9, 3)$, em que 10% da população são outliers; e $CN(0{,}8, 3)$, em que 20% da população são outliers. Essas duas distribuição são simétricas e têm longas caudas devido aos outliers.

Realizamos um teste de Bonett bilateral com um nível de significância alvo de $\alpha = 0{,}05$ em cada par de amostras de cada distribuição. Como os níveis de significância simulados estavam, em cada caso, baseados em 10.000 pares de réplicas de amostras, e como usamos o nível de significância alvo de 5%, o erro de simulação foi de $\sqrt{0{,}95(0{,}05)/10.000} = 0{,}2\%$.

Os resultados da simulação estão resumidos na Tabela 1 a seguir.

Tabela 1 Níveis de significância simulados para um teste de Bonett bilateral em experimentos para 2 amostras balanceados e não balanceados. O nível alvo de significância é 0,05.

Distribuição	n_1, n_2	Nível simulado	Distribuição	n_1, n_2	Nível simulado
N(0, 1)	10, 10	0,038	Exp	10, 10	0,052
	20, 10	0,043		20, 10	0,051
	20, 20	0,045		20, 20	0,049
	30, 10	0,044		30, 10	0,044
	30, 20	0,046		30, 20	0,042
	25, 25	0,048		25, 25	0,043
	30, 30	0,048		30, 30	0,042
	40, 40	0,051		40, 40	0,042
	50, 50	0,047		50, 50	0,039
t(5)	10, 10	0,044	Qui(5)	10, 10	0,040
	20, 10	0,042		20, 10	0,043
	20, 20	0,046		20, 20	0,040
	30, 10	0,041		30, 10	0,039
	30, 20	0,046		30, 20	0,043
	25, 25	0,048		25, 25	0,042
	30, 30	0,043		30, 30	0,043
	40, 40	0,046		40, 40	0,040
	50, 50	0,050		50, 50	0,039

Distribuição	n_1, n_2	Nível simulado	Distribuição	n_1, n_2	Nível simulado
t(10)	10, 10	0,041	Qui(10)	10, 10	0,044
	20, 10	0,040		20, 10	0,042
	20, 20	0,045		20, 20	0,041
	30, 10	0,046		30, 10	0,043
	30, 20	0,045		30, 20	0,045
	25, 25	0,046		25, 25	0,046
	30, 30	0,048		30, 30	0,038
	40, 40	0,045		40, 40	0,042
	50, 50	0,051		50, 50	0,049
Lpl	10, 10	0,054	B(8, 1)	10, 10	0,053
	20, 10	0,056		20, 10	0,045
	20, 20	0,055		20, 20	0,048
	30, 10	0,057		30, 10	0,042
	30, 20	0,058		30, 20	0,047
	25, 25	0,057		25, 25	0,041
	30, 30	0,053		30, 30	0,040
	40, 40	0,047		40, 40	0,042
	50, 50	0,048		50, 50	0,038
B(3, 3)	10, 10	0,032	CN(0,9, 3)	10, 10	0,024
	20, 10	0,037		20, 10	0,022
	20, 20	0,042		20, 20	0,018
	30, 10	0,039		30, 10	0,019
	30, 20	0,038		30, 20	0,020
	25, 25	0,039		25, 25	0,019
	30, 30	0,041		30, 30	0,015
	40, 40	0,044		40, 40	0,020
	50, 50	0,046		50, 50	0,017

Distribuição	n_1, n_2	Nível simulado	Distribuição	n_1, n_2	Nível simulado
U(0, 1)	10, 10	0,030	CN(0,8, 3)	10, 10	0,022
	20, 10	0,032		20, 10	0,019
	20, 20	0,031		20, 20	0,020
	30, 10	0,034		30, 10	0,017
	30, 20	0,034		30, 20	0,020
	25, 25	0,034		25, 25	0,021
	30, 30	0,037		30, 30	0,017
	40, 40	0,043		40, 40	0,023
	50, 50	0,043		50, 50	0,020

Conforme mostrado na Tabela 1, quando os tamanhos amostrais são menores, os níveis de significância simulados do teste de Bonett são menores do que o nível alvo de significância (0,05) para distribuições simétricas ou quase simétricas com caudas de leves a moderadas. Por outro lado, os níveis simulados tendem a ser um pouco maiores do que o nível alvo quando pequenas amostras se originam de distribuições altamente assimétricas.

Quando as amostras são moderadamente grandes ou grandes em tamanho, os níveis de significância simulados estão mais próximos do nível alvo de todas as distribuições. Na realidade, o teste apresenta um desempenho razoavelmente bom, mesmo para distribuições altamente assimétricas, como a distribuição exponencial e a distribuição Beta(8, 1).

Além disso, outliers parecem ter mais impacto em pequenas amostras do que em grandes amostras. Os níveis de significância simulados para as populações normais contaminadas estabilizaram em aproximadamente 0,020 quando o tamanho mínimo das duas amostras alcançaram 20.

Quando o tamanho mínimo das duas amostras é 20, os níveis de significância simulados consistentemente se encaixam no intervalo [0,038, 0,058], exceto pela distribuição uniforme plana e distribuições normais contaminadas. Apesar do nível de significância simulado de 0,040 ser ligeiramente conservador para um nível alvo de 0,05, esse taxa de erros do Tipo I pode ser aceitável para a maioria dos propósitos práticos. Portanto, concluímos que o teste de Bonett é válido quando o tamanho mínimo das duas amostras é de, no mínimo, 20.

Simulação B2: Validade do teste de MC (modelos de múltiplas amostras)

Parte I: Experimentos balanceados

Realizamos uma simulação para examinar o desempenho do teste de MC em modelos de múltiplas amostras com experimentos balanceados. Geramos k amostras de tamanho igual da mesma distribuição, usando o conjunto de distribuições previamente listadas na

simulação B1. Selecionamos o número de amostras em um experimento para serem $k = 3$, $k = 4$ e $k = 6$ e fixamos o tamanho das k amostras em cada experimento em 10, 15, 20, 25, 50 e 100.

Realizamos um teste de MC bilateral com um nível de significância alvo de $\alpha = 0{,}05$ nas mesmas amostras de cada caso do experimento. Como os níveis de significância simulados foram, em cada caso, baseados em 10.000 pares de réplicas de amostras, e como usamos o nível de significância alvo alvo de 5%, o erro de simulação foi $\sqrt{0{,}95(0{,}05)/10.000} = 0{,}2\%$.

Os resultados da simulação estão resumidos nas Tabelas 2a e 2b a seguir.

Tabela 2a Níveis de significância simulados para um teste bilateral de múltiplas comparações em experimentos balanceados, de múltiplas amostras. O nível alvo de significância do teste é 0,05.

Distribuição	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nível simulado	n_i	Nível simulado	n_i	Nível simulado
N(0, 1)	10	0,038	10	0,038	10	0,036
	15	0,040	15	0,041	15	0,039
	20	0,039	20	0,040	20	0,041
	25	0,045	25	0,047	25	0,047
	50	0,046	50	0,046	50	0,052
	100	0,049	100	0,049	100	0,052
t(5)	10	0,042	10	0,044	10	0,042
	15	0,041	15	0,044	15	0,046
	20	0,043	20	0,045	20	0,045
	25	0,046	25	0,048	25	0,046
	50	0,040	50	0,039	50	0,038
	100	0,038	100	0,040	100	0,040
T(10)	10	0,033	10	0,037	10	0,038
	15	0,040	15	0,042	15	0,041
	20	0,042	20	0,043	20	0,043
	25	0,041	25	0,042	25	0,045
	50	0,047	50	0,044	50	0,047
	100	0,048	100	0,046	100	0,047

Distribuição	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nível simulado	n_i	Nível simulado	n_i	Nível simulado
Lpl	10	0,056	10	0,063	10	0,071
	15	0,056	15	0,061	15	0,063
	20	0,054	20	0,058	20	0,059
	25	0,051	25	0,056	25	0,58
	50	0,045	50	0,051	50	0,049
	100	0,044	100	0,047	100	0,050
B(3, 3)	10	0,031	10	0,031	10	0,031
	15	0,037	15	0,036	15	0,034
	20	0,035	20	0,036	20	0,037
	25	0,039	25	0,038	25	0,040
	50	0,044	50	0,044	50	0,044
	100	0,044	100	0,046	100	0,043
U(0, 1)	10	0,029	10	0,025	10	0,023
	15	0,026	15	0,027	15	0,026
	20	0,028	20	0,030	20	0,028
	25	0,034	25	0,033	25	0,032
	50	0,041	50	0,036	50	0,036
	100	0,048	100	0,047	100	0,045
Exp	10	0,063	10	0,073	10	0,076
	15	0,056	15	0,058	15	0,064
	20	0,051	20	0,053	20	0,057
	25	0,043	25	0,045	25	0,050
	50	0,033	50	0,037	50	0,038
	100	0,033	100	0,035	100	0,035

Tabela 2b Níveis de significância simulada para um teste bilateral de múltiplas comparações em experimentos balanceados, de múltiplas amostras. O nível alvo de significância do teste é 0,05.

Distribuição	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nível simulado	n_i	Nível simulado	n_i	Nível simulado
Qui(5)	10	0,040	10	0,046	10	0,048
	15	0,043	15	0,046	15	0,049
	20	0,040	20	0,040	20	0,042
	25	0,040	25	0,045	25	0,042
	50	0,037	50	0,038	50	0,040
	100	0,036	100	0,037	100	0,038
Qui(10)	10	0,042	10	0,045	10	0,045
	15	0,038	15	0,044	15	0,047
	20	0,036	20	0,039	20	0,040
	25	0,043	25	0,044	25	0,045
	50	0,041	50	0,040	50	0,042
	100	0,038	100	0,040	100	0,042
B(8, 1)	10	0,058	10	0,060	10	0,066
	15	0,057	15	0,061	15	0,064
	20	0,049	20	0,051	20	0,055
	25	0,044	25	0,046	25	0,050
	50	0,037	50	0,037	50	0,039
	100	0,037	100	0,038	100	0,039
CN(0,9, 3)	10	0,020	10	0,018	10	0,016
	15	0,022	15	0,020	15	0,017
	20	0,014	20	0,012	20	0,008
	25	0,011	25	0,011	25	0,008
	50	0,009	50	0,007	50	0,006
	100	0,010	100	0,008	100	0,008

Distribuição	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Nível simulado	n_i	Nível simulado	n_i	Nível simulado
CN(0,8, 3)	10	0,017	10	0,015	10	0,011
	15	0,013	15	0,011	15	0,008
	20	0,012	20	0,012	20	0,009
	25	0,013	25	0,010	25	0,009
	50	0,011	50	0,011	50	0,009
	100	0,014	100	0,012	100	0,010

Conforme mostrado nas Tabelas 2a e 2b, quando o tamanho amostral é pequeno, o teste de MC é, geralmente, conservador para distribuições simétricas e quase simétricas em experimentos balanceados. Por outro lado, o teste é liberal para pequenas amostras obtidas de distribuições altamente assimétricas como as distribuições exponenciais e a Beta(8, 1). Conforme o tamanho amostral aumenta, contudo, os níveis de significância simulados se aproximam do nível de significância alvo (0,05). Além disso, o número de amostras não parece ter um efeito forte no desempenho do teste para amostras que são moderadas em tamanhos. Quando os dados são contaminados com outliers, entretanto, há um impacto notável no desempenho do teste. O teste é excessiva e consistentemente conservador quando outliers estão presentes nos dados.

Parte II: Experimentos não balanceados

Realizamos uma simulação para examinar o desempenho do teste de MC em experimentos não balanceados. Geramos 3 amostras da mesma distribuição, usando o conjunto de distribuições anteriormente descritas na Simulação B1. No primeiro conjunto de experimentos, o tamanho das primeiras duas amostras foi $n_1 = n_2 = 10$ e o tamanho do terceiro conjunto de amostras foi $n_3 = 15, 20, 25, 50, 100$. No segundo conjunto de experimentos, o tamanho das primeiras duas amostras foi $n_1 = n_2 = 15$ e o tamanho do terceiro conjunto de amostras foi $n_3 = 20, 25, 30, 50, 100$. No terceiro conjunto de experimentos, definimos o tamanho amostral mínimo em 20, com o tamanho das primeiras duas amostras em $n_1 = n_2 = 20$ e o tamanho da terceira amostra em $n_3 = 25, 30, 40, 50, 100$.

Realizamos um teste bilateral de MC com um nível de significância alvo de $\alpha = 0{,}05$ nas mesmas três amostras de cada distribuição. Como os níveis de significância simulados estavam, em cada caso, baseados em 10.000 pares de réplicas de amostras, e como usamos o nível de significância alvo de 5%, o erro de simulação foi de $\sqrt{0{,}95(0{,}05)/10.000} = 0{,}2\%$..

Os resultados da simulação estão resumidos nas Tabelas 3a e 3b a seguir.

Tabela 3a Níveis de significância simulada para um teste de múltiplas comparações em experimentos não balanceados, de múltiplas amostras. O nível alvo de significância do teste é 0,05.

Distribuição	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nível simulado	n_3	Nível simulado	n_3	Nível simulado
N(0, 1)	15	0,032	20	0,040	25	0,045
	20	0,037	25	0,039	30	0,041
	25	0,038	30	0,037	40	0,043
	50	0,041	50	0,044	50	0,041
	100	0,042	100	0,042	100	0,044
t(5)	15	0,040	20	0,042	25	0,043
	20	0,036	25	0,040	30	0,037
	25	0,044	30	0,036	40	0,038
	50	0,033	50	0,036	50	0,035
	100	0,032	100	0,031	100	0,032
t(10)	15	0,039	20	0,042	25	0,042
	20	0,038	25	0,041	30	0,040
	25	0,040	30	0,041	40	0,041
	50	0,037	50	0,043	50	0,042
	100	0,036	100	0,039	100	0,040
Lpl	15	0,059	20	0,060	25	0,054
	20	0,057	25	0,054	30	0,051
	25	0,056	30	0,051	40	0,050
	50	0,049	50	0,051	50	0,050
	100	0,048	100	0,047	100	0,046
B(3, 3)	15	0,034	20	0,033	25	0,037
	20	0,031	25	0,035	30	0,039
	25	0,031	30	0,034	40	0,039
	50	0,036	50	0,039	50	0,038
	100	0,035	100	0,039	100	0,039

Distribuição	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nível simulado	n_3	Nível simulado	n_3	Nível simulado
U(0, 1)	15	0,027	20	0,030	25	0,032
	20	0,030	25	0,030	30	0,031
	25	0,028	30	0,032	40	0,036
	50	0,039	50	0,034	50	0,037
	100	0,042	100	0,038	100	0,042
Exp	15	0,061	20	0,053	25	0,042
	20	0,060	25	0,052	30	0,047
	25	0,054	30	0,049	40	0,043
	50	0,050	50	0,046	50	0,041
	100	0,044	100	0,040	100	0,040

Tabela 3b Níveis de significância simulados para o teste de MC em experimentos de múltiplas amostras, não balanceados. O nível alvo de significância do teste é 0,05.

Distribuição	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nível simulado	n_3	Nível simulado	n_3	Nível simulado
Qui(5)	15	0,047	20	0,045	25	0,041
	20	0,043	25	0,042	30	0,039
	25	0,043	30	0,039	40	0,040
	50	0,039	50	0,037	50	0,040
	100	0,034	100	0,035	100	0,034
Qui(10)	15	0,043	20	0,042	25	0,042
	20	0,039	25	0,038	30	0,041
	25	0,040	30	0,041	40	0,038
	50	0,038	50	0,041	50	0,042
	100	0,035	100	0,034	100	0,035
B(8, 1)	15	0,056	20	0,052	25	0,048
	20	0,054	25	0,046	30	0,044
	25	0,050	30	0,047	40	0,046

Distribuição	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Nível simulado	n_3	Nível simulado	n_3	Nível simulado
	50	0,046	50	0,043	50	0,043
	100	0,043	100	0,042	100	0,044
CN(0,9, 3)	15	0,017	20	0,020	25	0,017
	20	0,020	25	0,019	30	0,012
	25	0,017	30	0,016	40	0,013
	50	0,019	50	0,016	50	0,012
	100	0,014	100	0,016	100	0,010
CN(0,8, 3)	15	0,012	20	0,013	25	0,013
	20	0,016	25	0,012	30	0,012
	25	0,014	30	0,010	40	0,010
	50	0,015	50	0,010	50	0,013
	100	0,012	100	0,011	100	0,010

Os níveis de significância simulados mostrados nas Tabelas 3a e 3b são consistentes com aqueles informados anteriormente para múltiplas amostras com experimentos balanceados. Portanto, o desempenho do teste de MC não parece ser afetado por experimentos não balanceados. Além disso, quando o tamanho amostral mínimo for de pelo menos 20, os níveis simulados de significância estarão próximos do nível alvo, exceto pelos dados contaminados.

Em conclusão, quando a menor amostra é de, no mínimo, 20, o teste de MC apresenta bom desempenho para múltiplas (k) amostras em ambos os experimentos balanceados e não balanceados. Para amostras menores, contudo, o teste é conservador para dados simétricos e quase simétricos e liberal para dados altamente assimétricos.

Apêndice C: Função de poder teórico

A função de poder teórico exato do teste de MC não está disponível. Contudo, para experimentos para 2 amostras, pode ser obtida uma função de poder aproximado com base em métodos de teoria de grandes amostras. Para experimentos de múltiplas amostras, mais esforços de pesquisa são necessários para derivar uma aproximação similar.

Para experimentos para 2 amostras, contudo, a função de poder teórico do teste de Bonett pode ser obtida usando-se métodos de teoria de grandes amostras. Mais especificamente, a estatística do teste, T, fornecida a seguir é assintoticamente distribuída como uma distribuição qui-quadrado com 1 grau de liberdade:

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

Nesta expressão de T, $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$, $g_i = (n_i - 3)/n_i$ e γ é a curtose comum desconhecida das duas populações.

Segue-se então que a função de poder teórico de um teste bilateral de Bonett de igualdade de variâncias com um nível de significância aproximado de α pode ser determinado como

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

em que

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

Para testes unilaterais, a função de poder aproximado ao testar contra $\sigma_1 > \sigma_2$ é

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

e ao testar contra $\sigma_1 < \sigma_2$, a função de poder aproximada é

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

Observe que, durante o planejamento da fase do tamanho amostral da análise de dados, a curtose comum das populações, γ, é desconhecida. Portanto, o investigador tipicamente deve confiar nas opiniões dos especialistas ou nos resultados dos experimentos anteriores para obter um valor de planejamento para γ. Se aquela informação não estiver disponível, ela é frequentemente uma boa prática para realizar um pequeno estudo piloto para desenvolver os planos para o estudo principal. Usando as amostras do estudo piloto, um valor de planejamento de γ é obtido conforme a curtose combinada fornecida por

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

No menu do Assistente, a estimativa do planejamento de γ é obtida retrospectivamente com base nos dados do usuário em mãos.

Apêndice D: Comparação de poder teórico e simulado

Simulação D1: Poder simulado (real) do teste de Bonett

Realizamos uma simulação para comparar os níveis de poder simulado do teste de Bonett aos níveis de poder com base na função de poder aproximado, derivado no Apêndice C.

Geramos 10,000 pares de amostras para cada uma das distribuições descritas anteriormente (consulte a Simulação B1). Em geral, os tamanhos de amostra selecionados eram grandes o bastante para o nível de significância simulado do teste ser razoavelmente próximo do nível de significância alvo, com base em nossos resultados anteriores na Simulação B1.

Para avaliar os níveis de poder simulado em uma razão de desvios padrão $\rho = \sigma_1/\sigma_2 = 1/2$, multiplicamos a segunda amostra em cada par de amostras pela constante 2. Como resultado, para uma determinada distribuição e para determinados tamanhos amostrais n_1 e n_2, o nível de poder simulado foi calculado como a fração dos 10.000 pares de réplicas de amostras para as quais o teste bilateral de Bonett foi significativo. O nível alvo de significância alvo do teste foi fixado em $\alpha = 0{,}05$. Para comparação, calculamos os níveis de poder teóricos correspondentes com base na função de poder aproximado derivada no Apêndice C.

Os resultados são apresentados na Tabela 4 abaixo.

Tabela 4 Comparação dos níveis de poder simulado aos níveis de poder aproximado de um teste bilateral de Bonett. O nível de significância alvo é 0,05.

Distribuição	n_1, n_2	Poder Aprox.	Poder Simulado	Distribuição	n_1, n_2	Poder Aprox.	Poder Simulado
N(0, 1)	20, 10	0,627	0,527	Exp	20, 10	0,222	0,227
	20, 20	0,83	0,765		20, 20	0,322	0,368
	20, 30	0,896	0,846		20, 30	0,377	0,434
	20, 40	0,925	0,886		20, 40	0,412	0,475
	30, 15	0,825	0,771		30, 15	0,32	0,307
	30, 30	0,954	0,925		30, 30	0,458	0,50
	30, 45	0,98	0,97		30, 45	0,531	0,579
	30, 60	0,989	0,984		30, 60	0,575	0,622
t(5)	20, 10	0,222	0,379	Qui(5)	20, 10	0,355	0,347

Distribuição	n_1, n_2	Poder Aprox.	Poder Simulado	Distribuição	n_1, n_2	Poder Aprox.	Poder Simulado
	20, 20	0,322	0,569		20, 20	0,517	0,53
	20, 30	0,377	0,637		20, 30	0,597	0,616
	20, 40	0,412	0,69		20, 40	0,644	0,661
	30, 15	0,32	0,545		30, 15	0,513	0,51
	30, 30	0,458	0,733		30, 30	0,701	0,711
	30, 45	0,531	0,795		30, 45	0,781	0,793
	30, 60	0,575	0,828		30, 60	0,823	0,833
t(10)	20, 10	0,476	0,45	Qui(10)	20, 10	0,454	0,414
	20, 20	0,673	0,673		20, 20	0,646	0,631
	20, 30	0,756	0,749		20, 30	0,73	0,717
	20, 40	0,80	0,803		20, 40	0,776	0,771
	30, 15	0,668	0,659		30, 15	0,641	0,618
	30, 30	0,85	0,852		30, 30	0,828	0,819
	30, 45	0,91	0,911		30, 45	0,892	0,882
	30, 60	0,936	0,937		30, 60	0,921	0,912
Lpl	20, 10	0,321	0,33	B(8, 1)	20, 10	0,363	0,278
	20, 20	0,469	0,519		20, 20	0,528	0,463
	20, 30	0,545	0,585		20, 30	0,609	0,549
	20, 40	0,59	0,632		20, 40	0,655	0,60
	30, 15	0,466	0,475		30, 15	0,524	0,419
	30, 30	0,647	0,673		30, 30	0,713	0,634
	30, 45	0,729	0,758		30, 45	0,792	0,737
	30, 60	0,773	0,80		30, 60	0,833	0,777
B(3, 3)	20, 10	0,777	0,628	CN(0,9, 3)	20, 10	0,238	0,284
	20, 20	0,939	0,869		20, 20	0,346	0,452
	20, 30	0,973	0,936		20, 30	0,405	0,517
	20, 40	0,984	0,964		20, 40	0,442	0,561
	30, 15	0,935	0,871		30, 15	0,343	0,374

Distribuição	n_1, n_2	Poder Aprox.	Poder Simulado	Distribuição	n_1, n_2	Poder Aprox.	Poder Simulado
	30, 30	0,993	0,98		30, 30	0,491	0,598
	30, 45	0,998	0,995		30, 45	0,567	0,70
	30, 60	0,999	0,999		30, 60	0,612	0,719
U(0, 1)	20, 10	0,916	0,74	CN(0,8, 3)	20, 10	0,26	0,223
	20, 20	0,992	0,95		20, 20	0,379	0,396
	20, 30	0,998	0,985		20, 30	0,444	0,467
	20, 40	0,999	0,995		20, 40	0,484	0,52
	30, 15	0,991	0,941		30, 15	0,376	0,354
	30, 30	1,0	0,996		30, 30	0,535	0,549
	30, 45	1,0	1,0		30, 45	0,614	0,65
	30, 60	1,0	1,0		30, 60	0,661	0,706

Os resultados mostram que, em geral, os níveis de poder aproximado e os níveis de poder simulado estão próximos um do outro. Eles se tornam próximos conforme os tamanhos amostrais aumentam. Os níveis de poder aproximado são normalmente ligeiramente maiores do que os níveis de poder simulado para distribuições simétricas e quase simétricas com caudas de moderadas a leves. Eles são, contudo, ligeiramente menores do que os níveis de poder simulado para distribuições simétricas com caudas pesadas ou para distribuições altamente assimétricas. A diferença entre as duas funções de poder normalmente não é importante, exceto no caso em que as amostras são geradas da distribuição t com 5 gradus de liberdade.

No geral, quando o tamanho amostral mínimo alcança 20, os níveis de poder aproximado e os níveis de poder simulado são notavelmente próximos. Portanto, o planejamento de tamanhos amostrais pode ser baseado nas funções de poder aproximadas.