

Método de Bonett

SOBRE O ARTIGO BONETT'S ROBUST CONFIDENCE INTERVAL FOR A RATIO OF STANDARD DEVIATIONS

Senin J. Banga e Gregory D. Fox
18 de junho de 2013

RESUMO

Propomos um procedimento alternativo para corrigir um passo em falso sutil no intervalo de confiança (IC) de Bonett (2006) para a razão de dois desvios padrão. O estimador da curtose combinada para a estatística de teste de Layard (1973) na qual o intervalo de Bonett está baseado, é consistente somente quando as variâncias populacionais são iguais. Derivamos um estimador alternativo que é consistente quando as variâncias populacionais são iguais e quando elas são desiguais e usa o novo estimador para calcular o IC correto. Estudos de simulação revelam que o novo IC é, em geral, mais exato e mais preciso que o IC baseado no teste de Levene/Browne-Forsythe W_{50} e o teste de Pan (1999) L_{50} . Consistente com Pan, observamos que ICs baseados no teste de W_{50} exibem uma perda de precisão com pequenas amostras, frequentemente resultando em intervalos que têm largura infinita. ICs que estão baseados no teste de L_{50} apresentam bom desempenho com distribuições simétricas e quase simétricas, mas apresentam um desempenho ruim quando as populações são assimétricas.

Termos do índice: homogeneidade de variâncias, teste de Levene, teste de Brown-Forsythe, teste de Layard, intervalo de confiança (IC) para a razão de variâncias

1. Introdução

É amplamente conhecido que o teste F clássico, e os intervalos de confiança associados (ICs) são extremamente sensíveis a desvios da normalidade — tão sensíveis que, de fato, o teste F clássico não é adequado para a maioria das aplicações práticas. Por este motivo, muitos propuseram alternativas mais robustas. Dentre essas, o teste conhecido como "Test W50" é frequentemente preferido, porque ele tem propriedades muito boas de erro tipo I, contudo, é simples de calcular e de interpretar. (Para análises comparativas, consulte Conover et al. (1981), Balakrishnan e Ma (1990), Lim e Loh (1996).) O teste de W_{50} é baseado em um procedimento que foi originalmente proposto por Levene (1960) e posteriormente aprimorado por Brown e Forsythe (1974). O teste de W_{50} foi amplamente adotada e está

disponível na maioria dos pacotes de software estatístico bem conhecidos, como o Minitab Statistical Software, SAS, R e JMP.

As propriedades de erro tipo II do teste W_{50} são um tanto menos notáveis do que suas propriedades de erro tipo I. Pan (1999) mostra que, para algumas distribuições, incluindo a distribuição normal, o poder do teste de W_{50} em problemas de amostras tem um limite superior que é possivelmente bem abaixo de 1. E esse limite superior não é afetado pela magnitude da diferença entre as variâncias da população. Esta deficiência naturalmente se estende aos ICs que são baseados no teste de W_{50} . Pan mostra que há uma probabilidade não desprezível de que um IC para a razão das variâncias da população que está baseado no teste de W_{50} será infinito $(0, +\infty)$ e, desta forma, não informativo. A observação de Pan é consistente com os resultados nossas próprias simulações, reportamos mais adiante nesse artigo.

Pan propõe um procedimento alternativo, chamado L_{50} , para corrigir as limitações do procedimento W_{50} . Com base nos resultados da simulação, Pan conclui que o teste de L_{50} é mais poderoso do que o teste de W_{50} , contudo, ele é igualmente robusto e compartilha suas propriedades assintóticas desejadas. As amostras das simulações de Pan, contudo, foram tiradas de distribuições simétricas ou um pouco assimétricas, com caudas de pesadas a leves. O impacto potencial da assimetria no desempenho do teste L_{50} em pequenas amostras não foi especificamente discutido.

Pan também argumenta que o L_{50} procedimento é tão poderoso quanto outros procedimentos notavelmente robustos tais como o teste de postos Fligner-Killeen modificado e o teste adaptativo Hall-Padmanabhan. Praticamente, contudo, o teste de postos Fligner-Killeen modificado e o teste adaptativo Hall-Padmanabhan são um pouco menos úteis do que os testes L_{50} e W_{50} porque eles são computacionalmente trabalhosos e intensivos.

Recentemente, Bonett (2006) propôs um procedimento de IC alternativo que está baseado na versão para duas amostras do teste de homogeneidade de variâncias de Layard (1973). Bonett inclui diversos ajustes para aprimorar o desempenho do procedimento de Layard para pequenas amostras. Por exemplo, Bonett propõe um estimador da curtose combinada que é assintoticamente equivalente ao de Layard, mas que exibe menos vício para pequenas amostras.

Infelizmente, nem o estimador da curtose combinada original de Layard, nem a substituição proposta por Bonett são consistentes quando as variâncias da população não são iguais. Desta forma, os intervalos que Bonett (2006) propõe não são ICs adequados, mas são melhores descritos como intervalos de aceitação para o teste de igualdade das variâncias. Assim, subtrair as probabilidades de cobertura simuladas relatadas em Bonett (2006) da unidade, resulta em taxas de erro tipo I do teste de igualdade de variâncias. Comparando essas taxas de erro tipo I com aquelas do teste original de Layard confirma que os ajustes de Bonett aprimoram com sucesso o desempenho do teste de Layard para pequenas amostras. O IC para a razão das variâncias proposto por Bonett, contudo, deve ser revisitado.

Observe também que Bonett compara os intervalos propostos com ICs baseados no teste F aproximado de Shoemaker (2003). Contudo, o IC da razão de variâncias associado ao teste de Shoemaker — conforme brevemente descrito na página 106 do artigo do Shoemaker — também é baseado no estimador da curtose combinada de Layard. Portanto, os ICs

calculados na seção 7 do artigo de Shoemaker também são melhores descritos como intervalos de aceitação para o teste de igualdade das variâncias. Apesar desses erros, pode-se concluir a partir dos resultados da simulação de Bonett que seu ajuste melhorou o desempenho de pequenas amostras do teste de Layard da igualdade das variâncias, e que o teste resultante para a igualdade de variâncias tem melhor desempenho que o teste de Shoemaker.

No presente documento, corrigimos o passo em falso em Bonett (2006), estendendo a forma para duas amostras do teste de Layard para testar a hipótese nula sobre a razão das variâncias ou desvios padrão. Para realizar isso, propomos um estimador de curtose combinada que é consistente para qualquer razão hipotética. Depois, invertemos a estatística de teste para obter o IC para a razão. Por fim, conduzimos estudos de simulação para avaliar as propriedades de robustez do novo IC em experimentos de pequenas amostras. Além disso, comparamos o desempenho do novo IC em pequenas amostras com o desempenho dos ICs associados ao teste F clássico, o teste W_{50} e o teste L_{50} .

2. Teste de Layard e alguma extensão

Sejam o $Y_{i1}, \dots, Y_{in_i}, \dots, Y_{k1}, \dots, Y_{kn_k}$ k amostras independentes, cada amostra sendo independente e identicamente distribuída com a média de $E(Y_{ij}) = \mu_i$ e variância de $\text{Var}(Y_{ij}) = \sigma_i^2 > 0$. Além disso, suponha que as amostras se originem de populações com uma curtose comum de $\gamma = E(Y - \mu)^4 / \sigma^4 < \infty$. Observamos que Layard usa o excesso de curtose de $\gamma_e = \gamma - 3$.

Sejam \bar{Y}_i e S_i a média e o desvio padrão da amostra i , respectivamente. Além disso, seja $\tau^2 = 2 + (1 - 1/\bar{n})\gamma_e = 2 + (1 - 1/\bar{n})(\gamma - 3)$, onde $\bar{n} = \sum n_i/k$. Conforme indicado em Layard (1973), para grandes amostras, $\tau^2 \cong \text{Var}((n_i - 1)^{1/2} \ln S_i^2)$.

Para testar a hipótese nula da igualdade de variâncias, Layard realiza uma transformação ortogonal no vetor cujos componentes $Z_i = (n_i - 1)^{1/2} \ln S_i^2 / \tau$ são assintoticamente distribuídos como a distribuição normal padrão sob a hipótese nula. Depois, ele usa a propriedade de preservação da distância de transformações ortogonais para mostrar que a estatística de teste de S' (dada a seguir) é assintoticamente distribuída como uma distribuição qui-quadrado com $k - 1$ graus de liberdade sob a hipótese nula da igualdade de variâncias:

$$S' = \sum_{i=1}^k (n_i - 1) \left(\ln S_i^2 - \frac{\sum_{i=1}^k (n_i - 1) \ln S_i^2}{\sum_{i=1}^k (n_i - 1)} \right)^2 / \tau^2$$

Em geral, o $Z_i = (n_i - 1)^{1/2} (\ln S_i^2 - \ln \sigma_i^2) / \tau$ é assintoticamente distribuído como a distribuição normal padrão. Portanto, pode-se aplicar as técnicas de Layard para derivar a estatística de teste mais generalizada T'_k .

$$T'_k = \sum_{i=1}^k (n_i - 1) \frac{(\ln S_i^2 - \ln \sigma_i^2)^2}{\tau^2} - \left(\sum_{i=1}^k (n_i - 1) \frac{\ln S_i^2 - \ln \sigma_i^2}{\tau \sqrt{\sum_{i=1}^k (n_i - 1)}} \right)^2$$

O T'_k é assintoticamente distribuído como uma distribuição qui-quadrado com $k - 1$ graus de liberdade sob ambas as hipóteses nula e a alternativa.

Pode-se expressar o T'_k de uma forma que é mais similar àquela do S' . Expressando o termo quadrático como uma soma em dobro e realizando um pouco de álgebra, obtivemos o seguinte resultado:

$$T'_k = \sum_{i=1}^k (n_i - 1) \left(\ln S_i^2 - \ln \sigma_i^2 - \frac{\sum_{i=1}^k (n_i - 1) (\ln S_i^2 - \ln \sigma_i^2)}{\sum_{i=1}^k (n_i - 1)} \right)^2 / \tau^2$$

Se todas as variâncias forem iguais, então $T'_k = S'$. Portanto, o S' e T'_k são as mesmas estatísticas de teste ao testar a hipótese nula da igualdade de variâncias. Entretanto, o T'_k também pode ser usado mais geralmente para testar qualquer hipótese que seja expressa como funções das variâncias. Por exemplo, podemos usar o T'_k para testar qualquer hipótese nula na forma do $H_0: \sigma_i = \sigma_{0i}$ para qualquer $\sigma_{0i} > 0, i = 1, \dots, k$ dado.

Como o $\tau^2 = 2 + (1 - 1/\bar{n})(\gamma - 3)$ é desconhecido, um teste baseado em S' ou T'_k requer um estimador para a curtose comum das populações, γ . Por exemplo, para testar a hipótese nula da homogeneidade das variâncias, Layard propôs o seguinte estimador combinado da curtose comum:

$$\hat{\gamma} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^4}{\left[\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \right]^2} \sum_{i=1}^k n_i$$

Layard destaca contudo, que o $\hat{\gamma}$ não é necessariamente um estimador consistente da curtose comum quando as variâncias não são iguais.

No caso especial de experimentos de duas amostras, pode-se avaliar a magnitude da diferença entre os desvios padrão testando-se a hipótese nula do $H_0: \sigma_1/\sigma_2 = \rho_0$ para uma dada razão hipotética $\rho_0 > 0$. Contudo, pode-se avaliar esta diferença mais diretamente calculando-se o IC da razão dos desvios padrão.

Se $\rho_0 = 1$, então a hipótese nula é equivalente à hipótese da homogeneidade da variância. Portanto, pode-se basear o teste no $T'_2 = S'$, após substituir-se a versão de duas amostras do estimador de curtose de Layard para o γ na expressão do $\tau^2 = 2 + (1 - 1/\bar{n})(\gamma - 3)$ para obter $\hat{\tau}^2$.

Contudo, se $\rho_0 \neq 1$, o teste deverá ser baseado no T'_2 em vez de no S' . Além disso, se $\rho_0 \neq 1$, então o estimador de curtose combinada de Layard não é necessariamente consistente e, portanto, não pode ser usado para estimar a curtose comum das populações. Portanto, um estimador alternativo da curtose combinada — um que seja consistente para qualquer razão hipotética $\rho_0 > 0$ — é necessário.

Em seguida derivamos esse estimador. Como se trata de uma função do ρ_0 , denotamos o estimador como $\hat{\gamma}_p(\rho_0)$. Também definimos a estatística de teste $T_2 = \tau^2 T'_2 / \hat{\tau}^2$, onde $\hat{\tau}^2 = 2 + (1 - 1/\bar{n})(\hat{\gamma}_p(\rho_0) - 3)$. De acordo com o teorema de Slutsky, T_2 é assintoticamente distribuído como uma distribuição qui-quadrado com 1 grau de liberdade. Por fim, invertemos T_2 para obter ICs para $\rho = \sigma_1/\sigma_2$.

3. IC para a razão dos desvios padrão

A seção anterior detalha a necessidade de um estimador alternativo da curtose ao testar as hipóteses nulas que são declaradas em termos da razão das variâncias ou dos desvios padrão. O resultado a seguir fornece aquele estimador.

RESULTADO 1

Para qualquer $\rho = \sigma_1/\sigma_2 > 0$ dado, um estimador consistente da curtose combinada da curtose da população comum no modelo de duas amostras pode ser dado como

$$\hat{\gamma}_P(\rho) = (n_1 + n_2) \frac{\sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^4 + \rho^4 \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^4}{[(n_1 - 1)S_1^2 + \rho^2(n_2 - 1)S_2^2]^2}$$

A prova deste resultado pode ser encontrada no Apêndice A.

Conforme esperado, o $\hat{\gamma}_P(1)$ é idêntico ao estimador de curtose combinada de Layard, $\hat{\gamma}$, uma vez que o $\sigma_1/\sigma_2 = 1$ implica que os desvios padrão (ou variâncias) são iguais.

A estatística T'_2 , que é a versão de duas amostras da estatística geral T'_k , é dada como

$$T'_2 = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1}\right) \tau^2}$$

onde $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$ e $\tau^2 = 2 + (1 - 1/\bar{n})\hat{\gamma}_e = 2 + (1 - 1/\bar{n})(\gamma - 3)$.

Conforme indicado no Layard (1973), em grandes amostras, $\tau^2 \cong \text{Var}((n_i - 1)^{1/2} \ln S_i^2)$. Bonett (2006) usa uma aproximação alternativa, que é também adotada em Shoemaker (2003), $\text{Var}((n_i - 1)^{1/2} \ln S_i^2) \cong \gamma - (n_i - 3)/n_i$. Em grandes amostras, essas aproximações são equivalentes. Contudo, Shoemaker relata que a última versão é vantajosa ao usar seu teste da igualdade de variâncias com pequenas amostras. Usando este ajuste, a estatística T'_2 pode ser modificada como

$$T'_2 = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

onde $g_i = (n_i - 3)/n_i$.

Segue-se então, que a estatística de teste $T_2 = \tau^2 T'_2 / \hat{\tau}^2$ para teste das hipóteses nulas $H_0: \rho = \rho_0$ é dada como

$$T_2 = \frac{(\ln \hat{\rho}^2 - \ln \rho_0^2)^2}{\frac{\hat{\gamma}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho_0) - g_2}{n_2 - 1}}$$

Nesta expressão do T_2 , a raiz quadrada do denominador pode ser exibida como uma estimativa para grandes amostras do erro padrão para a curtose combinada.

Além disso, na expressão de $\hat{\gamma}_P(1) \equiv \hat{\gamma}$, Bonett (2006) usa as médias de amostras aparadas com as médias amostrais aparadas $1/[2(n_i - 4)^{1/2}]$. Neste sentido, fizemos o mesmo ajuste ao estimador de curtose combinada:

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

onde m_i é a média aparada da amostra i , com a proporção da aparada $1/[2(n_i - 4)^{1/2}]$. Esta versão do estimador da curtose combinada e a versão anterior são assintoticamente equivalentes uma vez que a média aparada m_i é um estimador consistente da média da população μ_i . Esta versão alternativa, contudo, pode aprimorar o desempenho de pequenas amostras do teste baseado em T_2 .

A estatística do teste T_2 pode agora ser invertida para derivar um IC aproximado para a razão das variâncias ou desvios padrão. Contudo, primeiro, descreveremos brevemente o passo em falso na derivação dos ICs de Bonett (2006) para a razão dos desvios padrão.

3.1 Intervalos de Bonett

Em vez de inverter T_2 para obter o IC, Bonett (2006) inverte a seguinte estatística

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

Consequentemente, o intervalo resultante é simplesmente a região de aceitação para o teste de igualdade de variâncias. Isso é porque o estimador da curtose combinada, $\hat{\rho}_P(1)$, é consistente somente quando as variâncias são iguais, ou equivalentemente quando a razão hipotética é 1. O intervalo resultante é relatado em Bonett (2006) como

$$\exp[\ln(c S_1^2/S_2^2) \pm z_{\alpha/2} se]$$

onde

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

A constante c está incluída como um ajuste de pequenas amostras para mitigar o efeito de probabilidades de erro de caudas desiguais em experimentos não balanceados. Esta constante é dada por

$$c = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

A constante desaparece quando os projetos são equilibrados e seu efeito torna-se desprezível com o aumento do tamanho das amostras.

A Tabela 1 ilustra as consequências da interpretação errônea dos intervalos acima como ICs. Esses resultados são baseados em um pequeno estudo de simulação no qual calculamos as probabilidades de cobertura simuladas com base nos intervalos de Bonett (2006). Para os casos de variâncias iguais (coluna da esquerda), geramos duas amostras independentes da distribuição normal padrão. Para os casos de variâncias diferentes (coluna da direita), escalamos as observações da segunda amostra por um fator de constante de 4. As

probabilidades de cobertura estimadas estão baseadas em 100.000 réplicas. A cobertura nominal almejada é 0,95.

Tabela 1 Efeito de variâncias populacionais diferentes nos ICs de Bonett (2006) ($\alpha = 0,05$)

n_1, n_2	Probabilidades de cobertura simuladas	
	Variâncias iguais	Variâncias diferentes
10, 10	0,963	0,972
50, 50	0,952	0,991
100, 100	0,952	0,994

Se os intervalos foram baseados em um estimador de curtose combinada consistente, então espera-se que as probabilidades de cobertura nos dois casos sejam idênticas. Contudo, observe que os intervalos são consistentemente mais conservadores quando as variações são diferentes. Além disso, as probabilidades de cobertura se aproximam de 1 conforme o tamanho das amostras aumenta. Observe que resultados similares são obtidos com ICs aproximados do Shoemaker (2003).

3.2 Cálculos do IC

Considere o problema de testar a hipótese nula $H_0: \rho = \rho_0$ contra a hipótese alternativa $H_A: \rho \neq \rho_0$, onde $\rho = \sigma_1/\sigma_2$ e $\rho_0 > 0$, com base na estatística de teste T_2 dada anteriormente. Sob a hipótese nula, a estatística de teste

$$T_2 = \frac{(\ln \hat{\rho}^2 - \ln \rho_0^2)^2}{\frac{\hat{Y}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho_0) - g_2}{n_2 - 1}}$$

é assintoticamente distribuída como uma distribuição qui-quadrado com 1 grau de liberdade. Assim, o teste rejeita a hipótese nula ao nível α de significância se e somente se

$$(\ln \hat{\rho}^2 - \ln \rho_0^2)^2 > z_{\alpha/2}^2 \left(\frac{\hat{Y}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho_0) - g_2}{n_2 - 1} \right)$$

onde z_α denota o $\alpha \times 100$ o. ponto percentílico superior da distribuição normal padrão. Observe que o $\alpha \times 100$ o. ponto percentílico superior da distribuição qui-quadrado com 1 grau de liberdade, $\chi_{1,\alpha}^2$, satisfaz a seguinte condição: $\chi_{1,\alpha}^2 = z_{\alpha/2}^2$.

Os resultados da simulação de Bonett (2006) mostram que o ajuste de pequenas amostras, para reduzir o efeito de probabilidades de erro de caudas diferentes em experimentos não balanceados, funcionou bem. Assim, fizemos um ajuste similar para o teste com base em T_2 . Quando este ajuste é feito, o teste rejeita a hipótese nula se e somente se

$$(\ln \rho_0^2 - \ln(c\hat{\rho}^2))^2 > z_{\alpha/2}^2 \left(\frac{\hat{Y}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho_0) - g_2}{n_2 - 1} \right)$$

onde c é a constante de ajuste de Bonett dada como

$$c = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Equivalentemente, um percentual de confiança aproximado $(1 - \alpha)100$ definida para $\rho = \sigma_1/\sigma_2$ com base em T_2 é dada por

$$\left\{ \rho \in (0, \infty): (\ln \rho^2 - \ln(c\hat{\rho}^2))^2 - z_{\alpha/2}^2 \left(\frac{\hat{Y}_P(\rho) - g_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho) - g_2}{n_2 - 1} \right) \leq 0 \right\}$$

Observe que o c não tem efeito em experimentos balanceados e tem somente um efeito desprezível em experimentos não-balanceados de grandes amostras.

O próximo resultado fornece uma expressão alternativa da confiança definida de uma forma que é conveniente para descrever sua natureza. Nesta expressão, o estimador de curtose combinada é reescrito em termos das curtoses amostrais individuais dadas como

$$\hat{\gamma}_i = n_i \frac{\sum_{j=1}^{n_i} (Y_{ij} - m_i)^4}{[(n_i - 1)S_i^2]^2}, i = 1, 2$$

RESULTADO 2

Um percentual de confiança aproximado de $(1 - \alpha)100$ definido para $\rho = \sigma_1/\sigma_2$ com base em T_2 pode ser expresso como

$$\hat{\rho}\sqrt{c} \{r \in (0, \infty): H(r^2) \leq 0\}$$

ou equivalentemente, a confiança definida para $\rho^2 = \sigma_1^2/\sigma_2^2$ pode ser expressa como

$$c\hat{\rho}^2 \{r \in (0, \infty): H(r) \leq 0\}$$

onde

$$H(x) = (\ln x)^2 - z_{\alpha/2}^2 se^2(cx), x > 0$$

$$se^2(x) = A \frac{\hat{\gamma}_1 K^2/n_1 + \hat{\gamma}_2 x^2/n_2}{(K + x)^2} - B$$

$$A = \frac{(n_1 + n_2)(n_1 + n_2 - 2)}{(n_1 - 1)(n_2 - 1)}, B = \frac{g_1}{n_1 - 1} + \frac{g_2}{n_2 - 1}, K = \frac{n_1 - 1}{n_2 - 1}$$

Para a prova deste resultado, consulte o Apêndice B.

É facilmente verificado que a função $H(x)$ é contínua na reta real positiva, com $H(0) = H(+\infty) = +\infty$ e $H(1) < 0$. Portanto, pelo teorema dos valores intermediários, a função $H(x)$ admite pelo menos uma raiz no intervalo $(0, 1)$ e pelo menos uma raiz no intervalo $(0, +\infty)$.

O próximo resultado descreve o conjunto de confiança definido como um intervalo ou união de intervalos disjuntos.

RESULTADO 3

Se a função $H(x)$ tem exatamente duas raízes, x_L e x_U , então $0 < x_L < 1 < x_U$ e a confiança definidos para $\rho^2 = \sigma_1^2/\sigma_2^2$ é o intervalo dado por

$$[c\hat{\rho}^2 x_L, c\hat{\rho}^2 x_U]$$

Ocorre, então, que o IC do $\rho = \sigma_1/\sigma_2$ é o intervalo,

$$[\hat{\rho}\sqrt{cx_L}, \hat{\rho}\sqrt{cx_U}]$$

Por outro lado, se a função $H(x)$ tiver mais de duas raízes, a confiança definida para $\rho^2 = \sigma_1^2/\sigma_2^2$ é a união de intervalos não sobrepostos. Os pontos extremos de cada intervalo são as raízes consecutivas em que a função se abre para cima.

Para a prova deste resultado, consulte o Apêndice C.

OBSERVAÇÃO

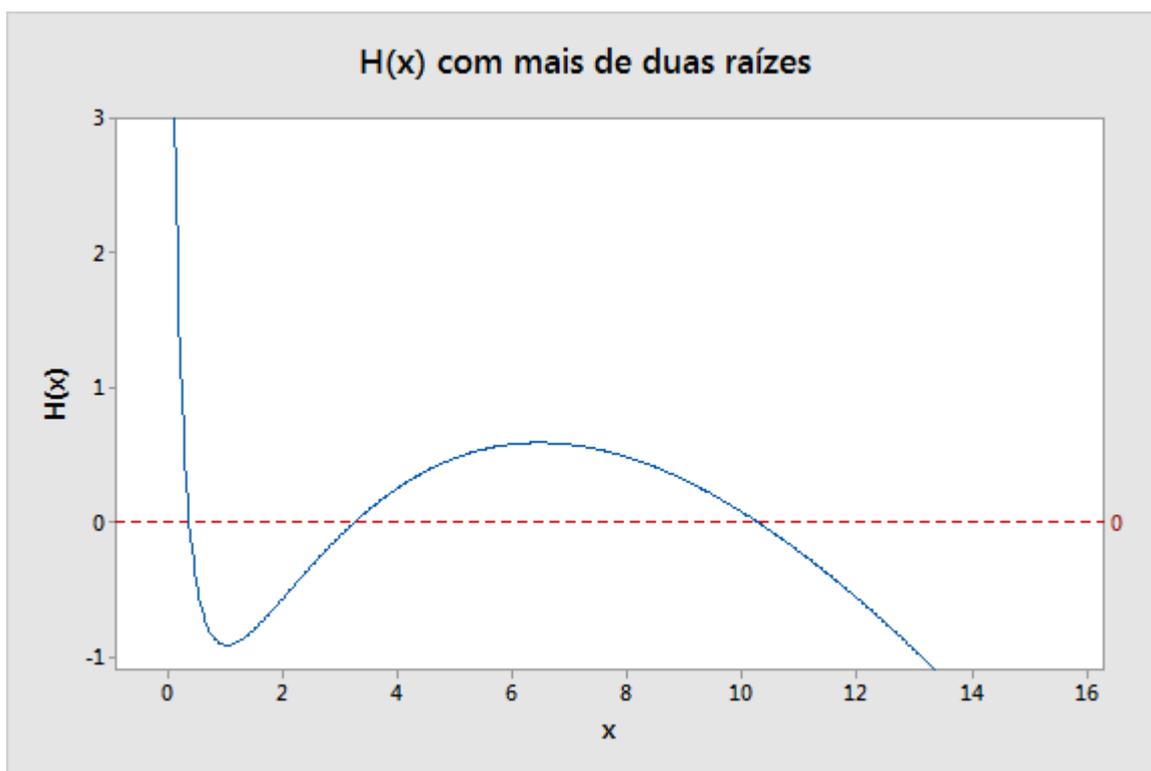
Apesar de ser matematicamente possível para a função $H(x)$ admitir mais de duas raízes, observamos que isso ocorre somente com experimentos extremamente incomuns e praticamente insignificantes nos quais uma ou ambas as amostras são muito pequenas ou gravemente não balanceadas. Conjecturamos que $H(x)$ tem duas ou quatro raízes.

O exemplo a seguir está baseado em dados que foram fabricados para forçar a função $H(x)$ a ter mais de duas raízes. Os dados são sumarizados da seguinte maneira: $n_1 = 169$, $n_2 = 7$, $S_1 = 301,855$, $S_2 = 4606,170$, $\hat{\gamma}_1 = 1,877$, $\hat{\gamma}_2 = 6,761$, $c = 0,728$, $A = 30,381$, $B = 0,101$ e $K = 28,000$.

Para $\alpha = 0,05$, a função $H(x)$ é dada como

$$H(x) = (\ln x)^2 - 1,960^2 \left(30,381 \frac{1,877 \times 28^2 / 169 + 6,761 \times (0,728x)^2 / 7}{(28,000 + 0,728x)^2} - 0,101 \right)$$

A função $H(x)$ neste caso tem quatro raízes. O gráfico da função é exibido a seguir. Observe que a quarta raiz não é visível no gráfico porque ela é muito grande. Entretanto, sabemos que a quarta raiz existe porque $H(+\infty) = +\infty$.



As quatro raízes são numericamente calculadas como $x_1 = 0,389$, $x_2 = 3,282$, $x_3 = 10,194$ e $x_4 = 39685,0$. A razão estimada dos desvios padrão é $\hat{\rho} = S_1/S_2 = 0,066$. A confiança definida para $\rho^2 = \sigma_1^2/\sigma_2^2$ pode ser expressa como

$$[c \hat{\rho}^2 x_1, c \hat{\rho}^2 x_2] \cup [c \hat{\rho}^2 x_3, c \hat{\rho}^2 x_4] = [0,001, 0,010] \cup [0,032, 124,072]$$

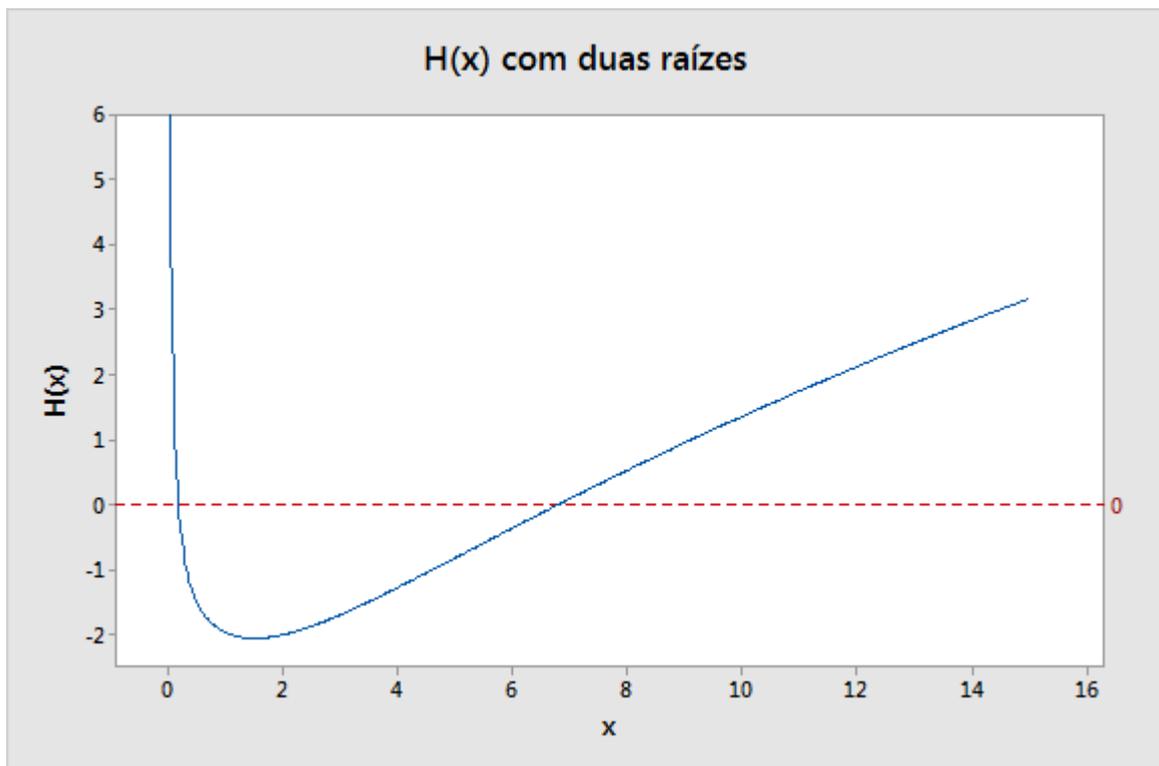
A confiança definida para a razão dos desvios padrão, ρ , é obtida tomando-se a raiz quadrada dos pontos extremos dos intervalos.

Quando as amostras não são muito pequenas ($n_i \geq 10$) e a disparidade entre seus tamanhos não é grande, a função $H(x)$ tipicamente admite duas raízes. Uma raiz está abaixo da unidade, e a outra raiz está acima da unidade conforme descrito no Resultado 2. Aqui está um exemplo baseado em dados gerados aleatoriamente. Os dados podem ser resumidos da seguinte maneira: $n_1 = 10$, $n_2 = 12$, $S_1 = 1,150$, $S_2 = 1,043$, $\hat{\gamma}_1 = 2,704$, $\hat{\gamma}_2 = 3,671$, $c = 1,041$, $A = 4,444$, $B = 0,146$ e $K = 0,818$.

Para $\alpha = 0,05$, a função $H(x)$ é dada neste caso como

$$H(x) = (\ln x)^2 - 1,960^2 \left(4,444 \frac{2,704 \times 0,818^2/10 + 3,671 \times (1,041x)^2/12}{(0,818 + 1,041x)^2} - 0,146 \right)$$

A função $H(x)$ tem duas raízes conforme mostrado a seguir:



As duas raízes são numericamente calculadas como $x_1 = 0,200$ e $x_2 = 6,824$. A razão estimada dos desvios padrão é $\hat{\rho} = S_1/S_2 = 1,102$. O conjunto de confiança de $\rho^2 = \sigma_1^2/\sigma_2^2$ é o intervalo dado como

$$[c \hat{\rho}^2 x_1, c \hat{\rho}^2 x_2] = [0,253, 8,634]$$

O IC para a razão dos desvios padrão, ρ , é obtido tomando-se a raiz quadrada dos pontos extremos do intervalo acima.

Descrevemos dois algoritmos para encontrar os limites de confiança.

O primeiro algoritmo consiste em usar um procedimento de busca de raiz numérica para encontrar as raízes da função $H(x)$. A raiz que corresponde a um limite de confiança inferior para a razão das variâncias está confinada no intervalo $(0, 1)$. Se denotarmos esta raiz por x_L e, depois, pelo Resultado 3, o limite de confiança inferior da razão das variâncias é calculado como $c\hat{\rho}^2 x_L$, e o limite de confiança inferior da razão dos desvios padrão é obtido como $\hat{\rho}\sqrt{c x_L}$. Similarmente, o limite de confiança superior para a razão das variâncias é $c\hat{\rho}^2 x_U$, e o

limite de confiança superior para a razão dos desvios padrão é $\hat{\rho}\sqrt{cx_U}$, no qual $x_U > 1$ é a outra raiz de $H(x)$. Uma abordagem simples para encontrar o limite de confiança superior é usar o fato de que o limite inferior do $1/\rho^2$ é o limite superior de ρ^2 . Primeiro, as funções da primeira amostra e da segunda amostra são intercambiadas na expressão da função $H(x)$ como se estivesse calculando o limite de confiança da razão $1/\rho^2 = \sigma_2^2/\sigma_1^2$. Em segundo lugar, o algoritmo para localizar o limite inferior é aplicado à nova função $H(x)$. Por fim, o limite resultante é invertido para obter o limite de confiança superior desejado.

Uma abordagem alternativa consiste em calcular recursivamente o limite de confiança inferior da razão das variâncias usando a relação de recorrência dada por

$$\rho_0^2 = 1$$

$$\rho_{i+1}^2 = \exp \left[\ln(c \hat{\rho}^2) - z_{\alpha/2} \sqrt{\frac{\hat{Y}_P(\rho_i) - g_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho_i) - g_2}{n_2 - 1}} \right], i = 0, 1, 2, \dots$$

O limite de confiança inferior para a razão das variâncias é ρ_{j+1}^2 , tal que $|\rho_{j+1}^2 - \rho_j^2| < \varepsilon$, no qual $j > 0$ e ε são escolhidos para serem pequenos (por exemplo $\varepsilon = 10^{-6}$). Para encontrar o limite de confiança superior, simplesmente substituímos $-z_{\alpha/2}$ por $+z_{\alpha/2}$ acima.

Evidentemente, os dois algoritmos para cálculo dos limites de confiança são equivalentes porque o procedimento recursivo é essencialmente um procedimento iterativo para solução da equação $H(\rho^2/(c\hat{\rho}^2)) = 0$ para ρ^2 . O algoritmo recursivo é mais fácil de implementar e, portanto, é uma alternativa útil quando um procedimento de busca de raiz não está disponível.

4. Estudos de simulação e resultados

Neste artigo, derivamos um procedimento para estender o teste de Layard para igualdade de duas variâncias, para testar a razão das variâncias. Chamamos este procedimento de Teste de Layard Estendido para a Razão ou ELTR. Nesta seção, investigamos as propriedades de pequenas amostras de ICs baseados no procedimento ELTR. Seguimos a abordagem geral adotada por Bonett (2006).

Comparamos os ICs que são baseados no procedimento ELTR a ICs que são baseados no teste L_{50} (Pan, 1999) e no teste W_{50} (o teste de Levene/Brown-Forsythe). Para o Estudo 1, também incluímos, para comparação, os ICs que são baseados no teste F clássico. É bem conhecido que, quando os dados são normalmente distribuídos, o teste F clássico é ótimo. Observe que os cálculos dos ICs com base nos testes W_{50} e L_{50} são fornecidos em Pan (1999). Os cálculos dos ICs baseados no teste F podem ser encontrados em diversos livros introdutórios de estatística. Eles também são fornecidos em Bonett (2006).

Conduzimos três estudos de simulação, cada um com 100.000 réplicas amostrais. Cada réplica consiste em duas amostras independentes que são de tamanho pequeno a moderado. Cada amostra foi tirada de uma distribuição pai com propriedades conhecidas incluindo simetria, assimetria, caudas pesadas e caudas leves. O erro padrão associado com

cada simulação é aproximadamente 0,0009, 0,0007 e 0,0003 para níveis de confiança nominal de 90%, 95% e 99% respectivamente.

Para avaliar o desempenho de cada procedimento, relatamos a probabilidade de cobertura alcançada e a largura média dos intervalos simulados da razão das variâncias. Alguns dos intervalos associados com o teste W_{50} têm largura infinita (uma possibilidade exposta por Pan (1999)). Nesses casos, relatamos a largura média dos intervalos finitos e o percentual dos intervalos com largura infinita. Todas as simulações foram conduzidas usando-se a Versão 8 do pacote de software Mathematica.

Estudo 1: Comparação das probabilidades de cobertura para Dados Normais

No primeiro estudo, geramos amostras aleatórias de vários tamanhos a partir da distribuição normal. Os resultados estão apresentados na Tabela 2.

Table 2 Comparação das Probabilidades de Cobertura e Largura Média dos Intervalos

$1 - \alpha$	n_1, n_2	Medir	Procedimento			
			F	ELTR	L_{50}	W_{50}
0,90	10, 10	Cobertura	0,898	0,918	0,913	0,921
		Largura	3,72	5,06	4,72	8,03 (0,01%)
	30, 10	Cobertura	0,90	0,909	0,897	0,911
		Largura	2,42	3,01	3,58	3,17
	25, 25	Cobertura	0,902	0,907	0,914	0,916
		Largura	1,61	1,73	1,85	1,938
	50, 50	Cobertura	0,90	0,901	0,906	0,907
		Largura	1,03	1,06	1,13	1,15
0,95	10, 10	Cobertura	0,949	0,963	0,958	0,964
		Largura	4,90	7,72	6,52	497,24 (0,20%)
	30, 10	Cobertura	0,95	0,957	0,945	0,959
		Largura	2,98	4,91	4,67	4,07
	25, 25	Cobertura	0,951	0,955	0,958	0,961
		Largura	1,99	2,24	2,31	2,49
	50, 50	Cobertura	0,951	0,952	0,953	0,954
		Largura	1,25	1,31	1,38	1,41

$1 - \alpha$	n_1, n_2	Medir	Procedimento			
			F	ELTR	L_{50}	W_{50}
0,99	10, 10	Cobertura	0,989	0,993	0,992	0,994
		Largura	8,29	17,76	12,52	$> 10^4$ (8,8%)
	30, 10	Cobertura	0,99	0,992	0,986	0,994
		Largura	4,26	15,76	8,26	6,77
	25, 25	Cobertura	0,99	0,992	0,992	0,993
		Largura	2,86	3,66	3,43	4,03
	50, 50	Cobertura	0,99	0,991	0,991	0,991
		Largura	1,71	1,89	1,92	2,02

As linhas sombreadas exibem as probabilidades de cobertura alcançadas (Cobertura) de cada procedimento em cada nível de confiança ($1 - \alpha$) e cada combinação de tamanhos de amostras (n_1, n_2). A média das larguras de intervalo (Largura) é exibida abaixo de cada probabilidade de cobertura. Se quaisquer intervalos de uma condição forem infinitos, relataremos a média dos intervalos finitos e o percentual de intervalos que foram infinitos.

Conforme esperado, os resultados mostram que os ICs associados ao procedimento F são os mais exatos e os mais precisos. As probabilidades de cobertura alcançadas com o procedimento F são mais próximas da cobertura alvo do que aquelas associadas aos outros procedimentos. E as larguras médias dos intervalos associadas ao procedimento F são menores do que aquelas associadas aos outros procedimentos. A tabela também revela, contudo, que ICs que são construídos usando-se os procedimentos ELTR e L_{50} são quase tão exatos e precisos quanto aqueles baseados no procedimento F.

Os intervalos baseados no teste W_{50} também são bastante exatos. Contudo, os intervalos do W_{50} podem ser muito amplos e podem até ter largura infinita, dependendo do tamanho das amostras. Observe que quando ambas as amostras têm somente 10 observações, pelo menos 0,01% dos intervalos produzidos pelo procedimento W_{50} são infinitamente amplos. E o percentual de intervalos infinitos aumenta quando a cobertura alvo aumenta. Sob a maioria das condições, as larguras médias do ELTR e os intervalos do L_{50} são menores do que as larguras médias dos intervalos do W_{50} .

Estudo 2: Comparação das probabilidades de cobertura para Dados Não-normais

O segundo estudo foi projetado para avaliar e comparar o desempenho do ELTR, L_{50} , e os procedimentos do W_{50} quando distribuições não são normais. Também incluímos uma distribuição normal contaminada para avaliar o impacto de outliers no desempenho dos procedimentos. Denotamos esta distribuição contaminada como CN(0,1, 3) para indicar que, enquanto 90% das observações são geradas da distribuição normal padrão, as 10% restantes

são geradas de uma população normal com uma média de 0 e um desvio padrão de 3. Os resultados são apresentados na Tabela 3.

Tabela 3 Comparação das probabilidades de cobertura e larguras médias dos intervalos em alguns modelos não normais o nível de confiança nominal é $1 - \alpha = 0,95$

Distribuição o [γ] n_1, n_2	ELTR	L_{50}	W_{50}	Distribuição o [γ] n_1, n_2	ELTR	L_{50}	W_{50}
Uniforme				$\chi^2(5)$			
[1,8]				[5,4]			
10, 10	0,971	0,971	0,966	10, 10	0,956	0,938	0,956
	5,27	4,87	42,08 (0,1%)		11,61	8,78	$> 10^4$ (2,6%)
10, 30	0,964	0,961	0,957	10, 30	0,959	0,923	0,956
	2,51	2,4	2,89		6,25	4,14	190,645 (0,3%)
25, 25	0,967	0,972	0,968	25, 25	0,956	0,944	0,954
	1,43	1,79	1,88		3,66	2,92	3,26
50, 50	0,959	0,962	0,959	50, 50	0,959	0,946	0,952
	0,83	1,06	1,08		2,07	1,7	1,77
Beta (3, 3)				Exponencia I			
[2,5]				[9]			
10, 10	0,968	0,966	0,966	10, 10	0,947	0,916	0,95
	6,26	5,59	254,62 (0,1%)		20,99	14,47	$> 10^4$ (9,1%)
10, 30	0,96	0,954	0,96	10, 30	0,954	0,896	0,953
	3,14	2,76	3,71		10,46	6,19	$> 10^4$ (4,1%)
25, 25	0,959	0,966	0,965	25, 25	0,956	0,931	0,951
	1,81	2,06	2,18		6,09	4,13	5,48 (0,008%)
50, 50	0,957	0,959	0,958	50, 50	0,962	0,942	0,952
	1,06	1,23	1,26		3,18	2,24	2,38

Distribuição o [γ] n_1, n_2	ELTR	L_{50}	W_{50}	Distribuição o [γ] n_1, n_2	ELTR	L_{50}	W_{50}
Laplace				$\chi^2(1)$			
[6]				[15]			
10, 10	0,946	0,935	0,961	10, 10	0,928	0,889	0,947
	13,47	10,45	$> 10^4$ (3,0%)		55,09	37,4	$> 10^5$ (25,1%)
10, 30	0,947	0,919	0,957	10, 30	0,943	0,882	0,956
	6,78	4,82	$> 10^4$ (0,4%)		18,71	11,14	$> 10^6$ (25,7%)
25, 25	0,945	0,94	0,952	25, 25	0,952	0,925	0,954
	4,00	3,372	3,86		10,97	6,84	$> 10^4$ (0,4%)
50, 50	0,952	0,949	0,955	50, 50	0,958	0,936	0,951
	2,19	1,91	1,99		5,08	3,31	3,75 (0,001%)
t(5)				Lognormal			
[9]				[113,9]			
10, 10	0,957	0,946	0,965	10, 10	0,923	0,876	0,955
	11,07	8,81	$> 10^3$ (2,0%)		59,22	46,15	$> 10^5$ (23,0%)
10, 30	0,957	0,93	0,959	10, 30	0,949	0,866	0,958
	6,06	4,24	$> 10^3$ (0,7%)		29,13	17,67	$> 10^6$ (31,6%)
25, 25	0,954	0,948	0,96	25, 25	0,947	0,917	0,965
	3,54	2,93	4,86 (0,01%)		16,21	8,73	$> 10^4$ (2,4%)
50, 50	0,954	0,947	0,954	50, 50	0,955	0,928	0,96
	2,10	1,71	1,77 (0,003%)		8,62	4,11	164,38 (0,2%)

Distribuição [γ] n_1, n_2	ELTR	L_{50}	W_{50}	Distribuição [γ] n_1, n_2	ELTR	L_{50}	W_{50}
Half Normal				CN(0,1, 3)			
[3,9]				[8,3]			
10, 10	0,956	0,942	0,954	10, 10	0,977	0,965	0,979
	10,41	7,89	$> 10^4$ (1,5%)		12,64	9,52	$> 10^4$ (4,9%)
10, 30	0,959	0,93	0,954	10, 30	0,981	0,952	0,979
	5,18	3,64	13,00 (0,02%)		7,82	4,71	944,68 (1,1%)
25, 25	0,959	0,952	0,959	25, 25	0,982	0,972	0,981
	3,01	2,62	2,88		4,63	3,22	3,71
50, 50	0,96	0,951	0,954	50, 50	0,983	0,972	0,978
	1,69	1,54	1,59		2,64	1,83	1,91

As linhas sombreadas exibem as probabilidades de cobertura alcançadas de cada procedimento, distribuição semelhante e cada combinação de tamanhos de amostras. A média das larguras dos intervalos é exibida abaixo de cada probabilidade de cobertura. Se quaisquer intervalos de uma condição forem infinitos, relataremos a média dos intervalos finitos e o percentual de intervalos que foram infinitos. A curtose (γ) de cada distribuição pai é exibida entre colchetes.

Para distribuições simétricas, de cauda leve, os resultados indicam que todos os três métodos produzem probabilidades de cobertura similarmente conservadoras. Contudo, o ELTR e os intervalos L_{50} são mais precisos para pequenas amostras do que são os intervalos W_{50} . Por exemplo, quando as amostras são geradas a partir de uma distribuição Beta com parâmetros de (3, 3), as probabilidades de cobertura alcançadas para o ELTR e os intervalos de L_{50} são pelo menos tão exatos quanto aqueles dos intervalos de W_{50} , mas os intervalos de W_{50} são consistentemente mais amplos.

O ELTR e os intervalos de W_{50} também são um pouco conservadores para distribuições simétricas, de caudas pesadas, enquanto os intervalos de L_{50} são liberais. Os intervalos de L_{50} são ainda mais liberais quando os experimentos são não balanceados. Por exemplo, quando amostras de tamanhos 10 e 30 são geradas a partir da distribuição Laplace, a probabilidade de cobertura alcançada para os intervalos de L_{50} é 0,919. E quando amostras do mesmo tamanho são geradas de uma distribuição-t com 5 graus de liberdade, a probabilidade de cobertura alcançada para os intervalos de L_{50} é 0,930.

Os intervalos de L_{50} também são bem liberais quando pequenas amostras são geradas de distribuições altamente assimétricas e com cauda pesada. Por exemplo, quando são geradas amostras a partir de uma distribuição lognormal, a cobertura alcançada pode ser tão baixa quanto 0,866. Para essas distribuições, o método W_{50} é o menos liberal dos três métodos.

Contudo, muitos dos intervalos W_{50} têm largura infinita. Por exemplo, quando são geradas amostras a partir da distribuição qui-quadrado com 1 grau de liberdade ($\chi^2(1)$), mais de 25% dos intervalos W_{50} podem ter largura infinita. Os intervalos de ELTR são um pouco menos exatos, mas consideravelmente mais estreitos e, assim, mais informativos do que os intervalos de W_{50} .

Por fim, observamos que todos os três procedimentos são adversamente afetados por outliers. O método L_{50} é o menos afetado, que pode ser esperado porque o método L_{50} foi derivado para reduzir o efeito dos outliers no teste W_{50} (Pan, 1999). Quando amostras são geradas a partir da distribuição normal contaminada, $CN(0,1, 3)$, o mínimo de probabilidades de cobertura alcançada para o ELTR e os procedimentos W_{50} é 0,977. Resultados adicionais da simulação (não mostrados) indicam que esses intervalos são aprimorados apenas lentamente com tamanhos de amostras crescentes.

Estudo 3: Sensibilidade à suposição de igualdade de curtoses

Nosso estudo final investiga a sensibilidade do procedimento do ELTR à suposição de igualdade de curtose sob a qual ele é derivado. Examinamos o desempenho do procedimento ELTR quando a curtose das populações pai não são iguais, que é quando $\gamma_1 \neq \gamma_2$. Também incluímos os procedimentos L_{50} e W_{50} , porque eles são derivados sob a suposição de que as populações pai não são iguais. Essa suposição de similaridade é minada quando a curtose das populações pai não são iguais. Os resultados são apresentados na Tabela 4.

Tabela 4 A sensibilidade do procedimento do ELTR ao nível de confiança nominal da suposição de igualdade de curtose é $1 - \alpha = 0,95$

Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	ELTR	L_{50}	W_{50}	Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	ELTR	L_{50}	W_{50}
Beta (3, 3), Normal [2,5, 3]				Normal, CN (0,9, 3) [3, 8,3]			
10, 10	0,964	0,961	0,964	10, 10	0,955	0,948	0,951
	0,27	0,23	204,50 (0,20%)		6,88	5,16	> 10 ⁴ (4,89%)
30, 10	0,946	0,939	0,946	30, 10	0,941	0,91	0,942
	0,16	0,17	0,15		5,26	3,77	3,20
10, 30	0,966	0,956	0,967	10, 30	0,961	0,95	0,958
	0,14	0,11	0,17		4,26	2,40	630,42 (1,10%)
50, 50	0,951	0,95	0,949	50, 50	0,936	0,91	0,907

Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	ELTR	L_{50}	W_{50}	Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	ELTR	L_{50}	W_{50}
	0,04	0,05	0,05		1,27	1,11	1,19
Normal, Laplace [3, 6]				Half Normal, $\chi^2(5)$ [3,9, 5,4]			
10, 10	0,941	0,935	0,947	10, 10	0,956	0,94	0,954
	6,67	5,17	$> 10^6$ (2,90%)		0,42	0,32	304,41 (2,60%)
30, 10	0,912	0,888	0,914	30, 10	0,954	0,918	0,949
	5,06	3,85	3,21		0,33	0,22	0,20
10, 30	0,963	0,943	0,955	10, 30	0,962	0,934	0,958
	3,33	2,25	$> 10^3$ (0,40%)		0,23	0,15	3,28 (0,30%)
50, 50	0,935	0,894	0,889	50, 50	0,955	0,941	0,945
	0,98	1,04	1,12		0,07	0,06	0,07
Normal, Half Normal [3, 3,9]				$\chi^2(5),$ Exponencial [5,4, 9]			
10, 10	0,956	0,948	0,957	10, 10	0,938	0,914	0,94
	28,16	20,65	$> 10^4$ (1,50%)		211,17	137,88	$> 10^6$ (9,10%)
30, 10	0,946	0,924	0,947	30, 10	0,928	0,875	0,929
	20,59	14,83	12,78		194,70	93,02	83,02
10, 30	0,961	0,946	0,962	10, 30	0,968	0,93	0,954
	14,06	9,37	49,11 (0,02%)		102,35	55,29	$> 10^5$ (3,90%)
50, 50	0,953	0,95	0,952	50, 50	0,95	0,92	0,923
	4,32	4,16	4,33		29,64	23,37	25,54

As linhas sombreadas exibem as probabilidades de cobertura alcançadas de cada procedimento, combinação de distribuições pai (Dist. 1, Dist. 2), e a combinação de tamanhos de amostras.. A média das larguras dos intervalos é exibida abaixo de cada probabilidade de cobertura. Se quaisquer intervalos de uma condição forem infinitos, relataremos a média dos intervalos finitos e o percentual de intervalos que foram infinitos. A curtose de cada distribuição pai (γ_1, γ_2) é exibida entre colchetes.

Em geral, o desempenho do procedimento ELTR não parece ter sido adversamente afetado por curtoses diferentes quando as amostras são suficientemente grandes. No entanto, quando os experimentos são desbalanceados e a menor amostra é obtida a partir da distribuição de caudas mais pesadas, as probabilidades de cobertura alcançadas são liberais. As probabilidades de cobertura alcançadas são melhores quando a amostra maior é gerada da distribuição de caudas mais pesadas.

Quando o tamanho das amostras é suficientemente grande, geralmente os intervalos L_{50} e W_{50} também parecem ser robustos à dissimilaridade das distribuições que resulta de curtoses diferentes. Contudo, observe que quando as amostras são geradas a partir de uma distribuição normal e de uma distribuição Laplace, ou de uma distribuição normal e de uma distribuição normal contaminada, as probabilidades de cobertura para os intervalos de L_{50} e de W_{50} não são estáveis, mesmo para amostras grandes tão grandes como 50.

Os intervalos de L_{50} são geralmente mais liberais do que o ELTR e os intervalos W_{50} . Em três casos, as probabilidades de cobertura alcançadas para os intervalos de L_{50} são menores do que 0,90. Em contraste, apenas uma das probabilidades de cobertura alcançadas para os intervalos de W_{50} é menor do que 0,90. A menor probabilidade de cobertura alcançada para os intervalos do ELTR é 0,912.

O estudo anterior (Tabela 3) mostra que todos os três procedimentos produzem intervalos que são visivelmente mais conservadores quando ambas as amostras são geradas a partir da distribuição normal contaminada, $CN(0,1, 3)$. O presente estudo mostra que todos os três procedimentos apresentam desempenho visivelmente melhor quando apenas uma amostra é gerada a partir de $CN(0,1, 3)$. Contudo, observe que o desempenho dos intervalos de L_{50} e de W_{50} parece degradar consideravelmente quando o tamanho das amostras aumenta para 50.

5. Exemplo

Nesta seção, aplicamos todos os quatro procedimentos — F, ELTR, L_{50} e W_{50} — para um conjunto de dados obtidos de Pan (1999). Ott (1993, página 352) descreve os dados da seguinte maneira:

Um químico em uma mina de ferro suspeita que a variância na quantidade (peso em onças) de óxido de ferro por libra de minério tende a aumentar conforme a quantidade média de óxido de ferro por libra aumenta. Para testar esta teoria, dez espécimes de 1 libra de minério de ferro são selecionados de cada um dos dois locais, um, local 1, contendo um conteúdo médio muito mais alto de óxido de ferro do que o outro, o local 2. As quantidades de óxido de ferro contidas no espécime de minério são mostradas a seguir:

Local 1	8,1	7,4	9,3	7,5	7,1	8,7	9,1	7,9	8,4	8,8
Local 2	3,9	4,4	4,7	3,6	4,1	3,9	4,6	3,5	4,0	4,2

Os ICs de 95% do $\sigma_2/\sigma_1 = 1/\rho$ calculado usando os quatro diferentes métodos são dados na tabela a seguir:

Procedimento	IC de 95%
F	(0,262, 1,055)

Procedimento	IC de 95%
ELTR	(0,277, 0,924)
L_{50} (Pan)	(0,295, 0,938)
W_{50} (Levene/Brown-Forsythe)	(0,237, 0,908)

6. Conclusão

Nossas simulações mostram que, em geral, os ICs baseados no procedimento de ELTR são tão exatos quanto os ICs derivados de teste L_{50} e W_{50} . Contudo, os intervalos do ELTR e os intervalos de L_{50} são mais precisos do que os intervalos de W_{50} para a maioria das distribuições. Os intervalos de W_{50} tendem a ser mais exatos do que os intervalos de ELTR e os intervalos de L_{50} quando amostras pequenas são geradas a partir de distribuições altamente assimétricas e com cauda pesada. Contudo, essa vantagem é normalmente compensada por uma perda notável de precisão. Os intervalos de W_{50} resultantes são tipicamente muito amplos e suscetíveis de ter uma largura infinita.

Conforme experimentado, os intervalos de L_{50} aprimoram a precisão dos intervalos de W_{50} . Para populações assimétricas contudo, os intervalos de L_{50} são tão curtos que eles produzem probabilidades de cobertura excessivamente liberais. Em contraste, os intervalos de ELTR são mais estáveis em geral. Os intervalos de ELTR não são, normalmente, muito longos ou muito curtos, portanto, as probabilidades de cobertura são normalmente nem muito conservadoras nem muito liberais. Portanto, o procedimento de ELTR parece ser o melhor procedimento para os fins mais práticos.

Os intervalos de ELTR são um pouco mais trabalhosos de calcular do que os intervalos baseados no teste de L_{50} ou no teste de W_{50} . Em geral, contudo, a precisão aumentada (comparada aos intervalos de W_{50}) e a precisão aumentada (comparada aos intervalos de L_{50}) mais do que compensam o esforço computacional extra. O procedimento de ELTR foi implementado como parte da análise Variância de duas amostras na Versão 17 do Minitab Statistical Software, na qual ela é referida como procedimento de Bonett.

Para pesquisa futura, pode-se considerar investigar as propriedades para pequenas amostras do teste de Layard em experimentos de múltiplas amostras quando o estimador de curtose combinada de Layard é substituído pela versão mais robusta de Bonett, dada como

$$\hat{\gamma}_A = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - m_i)^4}{\left[\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - m_i)^2 \right]^2} \sum_{i=1}^k n_i$$

onde m_i é a média aparada da amostra i , com a proporção da aparada $1/[2(n_i - 4)^{1/2}]$ e $i = 1, \dots, k$.

Além disso, pode ser benéfico usar a aproximação de Shoemaker da variância assintótica da variância amostral transformada por log.

Por fim, observamos que os intervalos propostos por Bonett (2006), apesar de não adequado como ICs, não deixam de ser notavelmente exatos e precisos para a maioria das distribuições, quando interpretados como regiões de aceitação para o teste da igualdade de

duas variâncias. Essas regiões de aceitação são bem adequadas para servir como base para um procedimento gráfico para comparação de múltiplas variâncias. Hochberg, Weiss e Hart (1982) propuseram um procedimento similar para teste da igualdade de médias. Tal procedimento foi implementado como parte da análise do Teste de Igualdade de Variâncias na Versão 17 do Minitab Statistical Software, na qual ele é referido como procedimento de Comparações Múltiplas.

7. Apêndices

Apêndice A: Prova do Resultado 1

Usando as notações e suposições da Seção 2, seja $X_j = \rho Y_{2j}$ para um dado $\rho = \sigma_1/\sigma_2$. Então

$$\text{Var}(X_j) = \rho^2 \text{Var}(Y_{2j}) = \rho^2 \sigma_2^2 = \sigma_1^2 = \text{Var}(Y_{1j})$$

e

$$E\left(\frac{X_j - \mu_{X_j}}{\sigma_{X_j}}\right)^4 = \rho^4 E\left(\frac{Y_{2j} - \mu_2}{\sigma_2}\right)^4 = E\left(\frac{Y_{2j} - \mu_2}{\sigma_2}\right)^4 = \gamma$$

Uma vez que $E\left(\frac{Y_{1j} - \mu_1}{\sigma_1}\right)^4 = \gamma$ por suposição, segue que as populações pai das duas amostras Y_{1j} e $X_j = \rho Y_{2j}$ têm a mesma variância σ_1^2 e a mesma curtose γ . Por Layard (1973), um estimador de curtose combinada consistente de γ com base nas duas amostras Y_{1j} e X_j é dado como

$$\begin{aligned} \hat{\gamma}' &= (n_1 + n_2) \frac{\sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^4 + \sum_{j=1}^{n_2} (X_j - \bar{X})^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_X^2]^2} \\ &= (n_1 + n_2) \frac{\sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^4 + \rho^4 \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^4}{[(n_1 - 1)S_1^2 + \rho^2(n_2 - 1)S_2^2]^2} = \hat{\gamma}_P(\rho) \end{aligned}$$

conforme necessário.

Apêndice B: Prova de resultado 2

Nós já estabelecemos que uma confiança percentual de $(1 - \alpha)100$ aproximada definida para $\rho = \sigma_1/\sigma_2$ com base em T_2 é dada por

$$\left\{ \rho \in (0, \infty): (\ln \rho^2 - \ln(c\hat{\rho}^2))^2 - z_{\alpha/2}^2 \left(\frac{\hat{\gamma}_P(\rho) - k_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho) - k_2}{n_2 - 1} \right) \leq 0 \right\}$$

O estimador de curtose combinada pode ser expresso em termos dos estimadores da curtose das amostras individuais, que são dadas por

$$\hat{\gamma}_i = n_i \frac{\sum_{j=1}^{n_i} (Y_{ij} - m_i)^4}{[(n_i - 1)S_i^2]^2}, i = 1, 2$$

Mais especificamente, se fizermos $t = \rho/\hat{\rho}$, então

$$\hat{\gamma}_P(\rho) = (n_1 + n_2) \frac{\sum_{j=1}^{n_1} (Y_{1j} - m_1)^4 + \rho^4 \sum_{j=1}^{n_2} (Y_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + \rho^2(n_2 - 1)S_2^2]^2} = (n_1 + n_2) \frac{\hat{\gamma}_1 K^2/n_1 + \hat{\gamma}_2 t^4/n_2}{(K + t^2)^2}$$

onde $K = (n_1 - 1)/(n_2 - 1)$.

Consequentemente, o termo de erro padrão quadrático pode ser expresso como

$$\frac{\hat{Y}_P(\rho) - k_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho) - k_2}{n_2 - 1} = A \frac{\hat{Y}_1 K^2/n_1 + \hat{Y}_2 t^4/n_2}{(K + t^2)^2} - B$$

onde

$$A = \frac{(n_1 + n_2)(n_1 + n_2 - 2)}{(n_1 - 1)(n_2 - 1)}, B = \frac{k_1}{n_1 - 1} + \frac{k_2}{n_2 - 1}$$

Assim, se permitimos $r^2 = \rho^2/(c\hat{\rho}^2)$, então, vê-se facilmente que

$$\begin{aligned} (\ln \rho^2 - \ln(c\hat{\rho}^2))^2 - z_{\alpha/2}^2 \left(\frac{\hat{Y}_P(\rho) - k_1}{n_1 - 1} + \frac{\hat{Y}_P(\rho) - k_2}{n_2 - 1} \right) \\ = (\ln r^2)^2 - z_{\alpha/2}^2 \left(A \frac{\hat{Y}_1 K^2/n_1 + \hat{Y}_2 c^2 r^4/n_2}{(K + c r^2)^2} - B \right) \end{aligned}$$

Disso resulta que uma confiança percentual aproximada de $(1 - \alpha)100$ definida para $\rho = \sigma_1/\sigma_2$ com base em T_2 pode ser dada como

$$\hat{\rho}\sqrt{c} \{r \in (0, \infty): H(r^2) \leq 0\}$$

ou equivalentemente, a confiança definida para $\rho^2 = \sigma_1^2/\sigma_2^2$ pode ser expressa como

$$c\hat{\rho}^2 \{r \in (0, \infty): H(r) \leq 0\}$$

onde

$$H(x) = (\ln x)^2 - z_{\alpha/2}^2 se^2(cx), x > 0$$

e

$$se^2(x) = A \frac{\hat{Y}_1 K^2/n_1 + \hat{Y}_2 x^2/n_2}{(K + x)^2} - B$$

Apêndice C: Prova de resultado 3

É facilmente verificado que $H(x)$ é contínuo na reta real positiva, com $H(0) = H(+\infty) = +\infty$ e $H(1) < 0$. Pelo teorema dos valores intermediários, a função $H(x)$ admite pelo menos uma raiz no intervalo de $(0, 1)$ e pelo menos uma raiz no intervalo de $(0, +\infty)$. Assim, se a função de $H(x)$ tiver exatamente duas raízes, então uma raiz está abaixo de 1 e a outra está acima de 1. Uma vez que esta função abre-se para cima, a desigualdade de $H(r) \leq 0$ é satisfeita se r situa-se entre as raízes. Essas raízes definem os pontos de extremidade do IC de $\rho^2/(c\hat{\rho}^2)$. Assim, se permitirmos que $x_L < 1 < x_U$ seja as duas raízes, então, pelo Resultado 2, o limite de confiança inferior da razão das variâncias, ρ^2 , é calculado como $c\hat{\rho}^2 x_L$, e o limite de confiança inferior da razão do desvio padrão é obtido como $\hat{\rho}\sqrt{c x_L}$. Similarmente, o limite de confiança superior da razão das variâncias é $c\hat{\rho}^2 x_U$ e o limite de confiança superior da razão dos desvios padrão é de $\hat{\rho}\sqrt{c x_U}$.

Por outro lado, se a função $H(x)$ tiver mais de duas raízes, a desigualdade de $H(r) \leq 0$ é satisfeita se r situa-se entre raízes consecutivas, nas quais a função se abre para cima. Assim, a confiança definida é uma união de intervalos não sobrepostos.

8. Referências

- Balakrishnan, N. e Ma, C. W.W. (1990). A Comparative Study of Various Tests for the Equality of Two Population Variances. *Journal of Statistical Computation and Simulation*, 35, 41–89.
- Bonett D. G. (2006). Robust Confidence Interval for a Ratio of Standard Deviations. *Applied Psychological Measurements*, 30, 432–439.
- Boos, D. D. e Brownie, C. (1989). Bootstrap Methods for Testing Homogeneity of Variances. *Technometrics*, 31, 69–82.
- Brown, M. B., e Forsythe A. B. (1974).(1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69, 364–367.
- Conover, W. J., Johnson, M. E. e Johnson, M. M. (1981).(1981). A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, 23, 351–361.
- Fligner, M. A. e Killeen, T. J. (1976). Distribution-Free Two-Sample Tests for Scale. *Journal of the American Statistical Association*, 71, 210–213.
- Hall, P. e Padmanabhan, A. R.R. (1997). Adaptive Inference for the Two-Sample Scale Problem. *Technometrics*, 39, 412–422.
- Hochberg, Y., Weiss, G. e Hart S., (1982). On Graphical Procedures for Multiple Comparisons. *Journal of the American Statistical Association*, 77, 767–772.
- Layard, M. W.W. J. (1973). Robust Large-Sample Tests for Homogeneity of Variances. *Journal of the American Statistical Association*, 68, 195–198.
- Levene, H. (1960). "Robust Tests for Equality of Variances," in I. Olkin, ed., *Contributions to Probability and Statistics*, Palo Alto, CA: Stanford University Press, 278–292.
- Lim, T.-S. and Loh, W.-Y. (1996). A Comparison of Tests of Equality of Variances. *Computational Statistics and Data Analysis*, 22, 287–301.
- Ott, L. (1993). *An Introduction to Statistical Methods and Data Analysis*, Belmont, CA: Duxbury Press.
- 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, 105–114.
- Wolfram, S. (1999). *The Mathematica Book*, 4th ed. Wolfram Media/Cambridge University Press.

© 2020 Minitab, LLC. All rights reserved. Minitab®, Minitab Workspace™, Companion by Minitab®, Salford Predictive Modeler®, SPM®, and the Minitab® logo are all registered trademarks of Minitab, LLC, in the United States and other countries. Additional trademarks of Minitab, LLC can be found at www.minitab.com. All other marks referenced remain the property of their respective owners.