I Encontro Português de Biometria & I Encontro Luso-Galaico de Biometria

Comunicação Oral

Combinação de Valores de Prova;
Valores de Prova Aleatórios e Valores de Prova Generalizados

Maria de Fátima Brilhante
Universidade dos Açores, Departamento de Matemática, e
CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa , fbrilhante@uac.pt

Dinis Pestana
Universidade de Lisboa, FCUL/DEIO
CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa,
Instituto de Investigação Científica Bento da Rocha Cabral, dinis.pestana@fc.ul.pt

Fernando Sequeira
Universidade de Lisboa, FCUL/DEIO e CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, fjsequeira@fc.ul.pt

Palavras–chave: Ampliação computacional de amostras, uniformidade, contaminaçã, valores de prova generalizados

Sumário:

A teoria clássica da combinação de valores de prova assume implicitamente que são valores observados de variáveis aleatórias uniformes. Discutimos os conceitos de valores de prova aleatórios e de valores de prova genereralizados, abordando depois a combinação de valores de prova na situação mais realista de os valores de prova poderem ser modelados usando densidades que são misturas convexas da densidade uniforme e da densidade Beta(1,2) ou Beta(2,1).

Palavras-Chave:

uniformidade, valores de prova generalizados, valores de prova aleatórios, combinação de valores de prova.

1 Introdução

Diz-se que rejeitar a hipótese nula é uma decisão forte, mantê-la é uma decisão fraca. De facto, a manutenção da hipótese nula é, frequentemente, simples consequência da escassez de dados. A globalização dos resultados de obtidos na cooperação (ou simples publicitação dos resultados) de diversas equipas de investigação pode permitir ultrapassar esse problema. Diversas formas de fazer uma síntese meta-analítica dos resultados podem, por outro lado, levar a uma conclusão global resultando da harmonização de resultados contraditórios.
A publicção em Medicina cada vez mais exige a disponibilização de dados, por forma a no futuro ser possível uma meta análise para determinar a magnitude de efeitos de diversos tratamentos. Por outro lado, ainda é comum nas publicações científicas apenas ser divulgado o valor de prova, havendo desde os anos 30 do século passado técnicas para os combinar, veja-se por exemplo [15, 24]. Quando dispomos de valores de prova independentes, sob validade da hipótese nula temos uma amostra de uma população uniforme padrão. Mas tal deixa de ser válido condicionalmente a ser a hipótese alternativa que é verdadeira, e torna-se bastante complexo quando parâmetros perturbadores dificultam o cálculo da distribuição amostral das estatísticas que pretendemos usar.
Encarar os valores de prova p como valores observados de uma variável aleatória P com suporte em $[0, 1]$ permite concetualizar melhor o conceito de significância. Por outro lado, quando há

parâmetros perturbadores, convém porventura definir uma *estatística de teste generalizada*, que nos dá acesso ao cálculo de um valor-p, referido como *valor-p generalizado*. Na secção 2 discutimos os conceitos de valor de prova aleatório e de valor de prova generalizado.
A combinção de valores de prova, usando-os directamente ou transformando-os (Pestana [24]) é simples, mas muitas vezes pouco eficaz, pois quando se tem como objectivo avaliar uma determinada hipótese nula usando os valores de prova obtidos em r testes independentes, é muito pouco adequado assumir que é válida a correspondente hipótese nula combinada. Consequentemente, alguns dos valores de prova observados **não** devem ser uniformes. Os resultados dessas sínteses meta-analíticas raramente integram uma análise do viés de publicação, cf. [25]
Na secção 3 analisamos a situação especial de considerar que em lugar de distribuição uniforme as variáveis aleatórias P têm uma densidade que é uma combinação convexa das densidades uniforme e Beta(2,1) ou Beta (1,2), mostrando que nesta classe de fvariáveis aleatórias X_m ($m \in [-2,2]$ é o coeficiente de mistura) se X_m e X_p forem independentes então

$$\min\left\{\frac{X_m}{X_p}, \frac{1-X_m}{1-X_p}\right\} = X_{\frac{mp}{6}}$$

e consequentemente sempre mais próxima da uniforme X_0 (coincidindo com a uniforme se alguma das variáveis originais for uniforme). Apresentamos também sucintamente uma extensão para o casso de dependência auto-regressiva.
Usamos aquele resultado para mostrar que aumentar computacionalmente amostras, como investigado em [12, 28, 3, 4] de valores de prova pode conduzir a um resultado contrário ao esperado, isto é piorar a potência do teste combinado.

2 Interpretação e uso de valores de prova

2.1 Valores de prova aleatórios

Seja T uma estatística de teste cuja distribuição amostral sob validade de H_0 é F_0, com inversa F_0^{-1} (ou inversa generalizada $F_0^{\leftarrow}$), mas que é F_θ sob validade da hipótese alternativa.
O valor de prova $p = \mathbb{P}(T > T(\text{obs.}) \mid H_0 \text{ verdadeira}) = 1 - F_0(\text{obs.})$ indica um grau de surpresa causado pela eventual discrepância entre o que é postulado na hipótese nula e o valor observado da estatística de teste, e é neste sentido restrito que nos é útil: aponta resultados que requerem observação mais cuidada, descartando por outro lado resultados que parecem não requerer a rejeição da hipótese nula — mas valores elevados do valor de prova nunca devem ser confundidos com um indicador probabilístico de que a hipótese nula é verdadeira. É também evidente que, como consequência direta do teorema da transformação uniformizante, sob validade de H_0 o valor de prova p é o valor observado de uma variável aleatória $P \frown Uniforme(0,1)$.
Rejeitar hipóteses nulas por haver evidência da sua possível falsidade é um dos pontos cardeais do método experimental, bem expresso no conceito popperiano de falsear hipóteses, ou na muito citada frase de Linus Pauling sobre "ter muitas ideias, e a coragem de deitar quase todas fora" caracterizar a postura, por excelência, do cientista.
Há porém que não esquecer que a repetibilidade de resultados experimentais é um outro fundamento da ciência — é aliás por confiarmos no critério de repetibilidade que o tradicional nível de significância 0.05 é usado como "ponto de viragem" na tomada de decisões. A questão de repetibilidade não é em geral discutida com a profundidade que merece, remetemos para Utts [31], que analisa a questão com rigor no contexto de meta análise, e veja-se também [13, 1]. E, naturalmente, faz sentido questionar se numa repetição da experiência é de facto expectável observar um valor p similar ao anteriormente obtido, e no caso de não ser investigar o por quê.
Ora já vimos que, sob a validade de H_0, p é o valor observado de $P \frown Uniforme(0,1)$, pelo que sob validade de H_0 qualquer valor entre 0 e 1 é igualmente expectável! Claro que o objetivo

de testar hipóteses é encontrar razões para rejeitar H_0, aumentando a credibilidade de H_A ser verdadeira, e sob H_A a distribuição amostral de P deixa de ser uniforme, e em geral é bastante assimétrica, com forte modalidade na zona próxima de 0, e é por isso que esperamos alguma consistência nos valores p obtidos em provas independentemente repetidas, e é isso que traz um valor acrescentado ao uso de testes de hipóteses no processo científico de extrair conhecimento da informação disponível. A assimetria da distribuição amostral de $P|\, H_A$ é muito apelativamente exibida em `http://demonstrations.wolfram.com/PValuesAreRandomVariables/`.
Mas nesta perspetiva importa muito discutir a distribuição amostral de P quando H_A é verdadeira [16, 20] — e isso ganha especial acuidade quando se procura uma síntese meta-analítica de valores de prova, combinando $p_k,\ k = 1, ..., n$, observados para decidir sob $H_0^* :\ H_0$ é verdadeira $\forall k = 1, ..., n$ *vs.* $H_A^* :\ \exists k \in \{1, ...n\}$ para o qual H_A é verdadeira.
De facto, como Kulinskaya *et al.* [19] observam, com excelente ilustração na p. 120, no contexto de testes sobre valores médios gaussianos, na repetição independente de uma experiência em que se obteve um valor $p = 0.05$, o valor esperado do valor de prova é $\mathbb{E}[P] = 0.122$, um resultado preocupante quando se pensa no viés de publicação em meta análise, cf. [25]. Que os valores esperados dos valores de prova aleatórios podem não ser interpretáveis, ou sequer dignos de menção, devido à possível forte assimetria de $P|\, H_A$ é também discutido em [13] numa perspetiva bayesiana, veja-se ainda a carta ao editor que Senn enviou sobre este artigo ao editor de *Statistics in Medicine* 10 anos após a publicação, e [27].
Diversos autores [16, 21, 27] discutiram valores de prova como variáveis aleatórias, e recomendamos também [8, 9, 11] no que se refere à informação veículada por valores de prova. A formalização do conceito de valor de prova aleatório pode ser feita pela exibição da função de distribuição, que por simplicidade apresentamos no contexto de variáveis aleatórias contínuas com função de distribuição invertível, deixando ao único leitor interessado a tarefa de formalização com inversa generalizada no caso mais geral.
Denote-se S, com função de distribuição F_S, o resultado de uma experiência ao qual se associa o valor de prova p, e defina-se o valor de prova aleatório $P_r = \mathbb{P}[S_0 > S]$ numa replicação independente com resultado S_0. A função de distribuição de P_r é assim

$$F_{P_r}(p) = \begin{cases} 0 & p < 0 \\ \mathbb{P}[F_{S_0}(S) \geq 1 - p] = 1 - F_S(F_{S_0}^{-1}(1-p) & 0 \leq p < 1 \\ 1 & p \geq 1 \end{cases} \tag{1}$$

Na secção 3 focamo-nos numa situação experimental "defeituosa" (explicando em quê), em que na investigação de uniformidade dos valores de prova faz sentido usar uma distribuição amostral em que uma proporção $1 - \frac{|m|}{2},\ m \in [-2, 2]$ de $P_k,\ k = 1, ..., \nu$ são uniformes, sendo os remanescentes provenientes de uma população de máximos ou de mínimos de uniformes independentes.

2.2 Valores de prova generalizados

Também a noção de valor de prova generalizado tem vindo a ganhar importância, estando aqui em causa ultrapassar as dificuldades geradas por haver parâmetros perturbadores; estranhamente nunca foi abordado o problema de valores de prova generalizados aleatórios — tarefa que também não empreendemos aqui — apesar de a questão, do ponto de vista conceptual, não necessitar de novo aparato teórico, sendo contudo a caracterização da distribuição amostral sob H_A muito mais complexa. Um exemplo é com certeza um bom veículo para introduzir as principais ideias da generalização de valores de prova quando há parâmetros perturbadores a ultrapassar.
Considere-se o problema de testar a igualdade de valores médios μ_1 e μ_2 de duas populações gaussianas, sem pressupor igualdade de variâncias, usando a informação de duas amostras independentes de tamanhos n_1 e n_2, respetivamente. Denotem-se $\overline{X}_1, \overline{X}_2, S_1^2, S_2^2$ os estimadores suais de valores médios e de variâncias.

Este problema (de Behrens-Fisher) goza de alguma celebridade por a solução de Fisher, no problema dual de estimação de um intervalo de confiança para $\mu_1 - \mu_2$, numa interpretação frequencista, a probabilidade de cobertura do estimador intervalar ser inferior à declarada. Fisher [10] remendou a sua solução inventando probabilidades fiduciais e intervalos de confiança fiduciais — ainda hoje polémicos — e acolheu nos seus *Annals of Eugenics* a solução bayesiana de Jeffreys [17], coincidente com a solução fiducial, ostensivamente ignorando a solução frequencista de Welch [35], usando uma estatística diferente. O problema de Behrens-Fisher permanece assim como uma questão em que frequencismo e bayseanismo não são apenas atitudes filosóficas, são de facto metodologias diversas que podem conduzir a soluções diferentes.
A dificuldade encontrada pelos pioneiros na solução do problema resultam de se pretender inferir sobre a diferença $\theta = \mu_1 - \mu_2$ de valores médios havendo um parâmetro perturbador $\boldsymbol{\eta} = (\sigma_1^2, \sigma_2^2)$. Em vez da complexa abordagem de Welch [35] , que recorre a uma sofisticada estimação de um número de graus de liberdade fracionário para uma estatística generalizada t_ν, usando o método dos momentos de modo a forçar que não haja estimativas inadmíssiveis da variância (cf. [23]) proceda-se da seguinte forma:
Sejam $\boldsymbol{X} = (\overline{X}_1 - \overline{X}_2, S_1^2, S_2^2)$, $\boldsymbol{x} = (\overline{x}_1 - \overline{x}_2, s_1^2, s_2^2)$ o valor observado daquele vetor aleatório, e defina-se a *estatística de teste generalizada*

$$T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta}) = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sqrt{\frac{s_1^2 \sigma_1^2}{n_1 S_1^2} + \frac{s_2^2 \sigma_2^2}{n_2 S_2^2}} \tag{2}$$

cujo valor observado é $t = \overline{x}_1 - \overline{x}_2$, e cujo valor esperado $[T]$ é uma função crescente de $\mu_1 - \mu_2$.
Uma vez que $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = Z \frown Gaussiana(0,1)$ e $Y_i = \frac{(n_i - 1) S_i^2}{\sigma_i^2} \frown \chi^2_{n_i - 1},\ i = 1, 2,$

$$T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta}) = Z \sqrt{\frac{(n_1 - 1) s_1^2}{n_1 Y_1} + \frac{(n_2 - 1) s_2^2}{n_2 Y_2}}, \tag{3}$$

em que as variáveis aleatórias Z, Y_1, Y_2 são independentes, podemos definir um *valor de prova generalizado* p_g no teste unilateral $H_0 :\ \mu_1 = \mu_2$ *vs.* $H_A :\ \mu_1 > \mu_2$ como

$$p_g = \mathbb{P}[T \geq t | \mu_1 = \mu_2] = \mathbb{P}\left[Z \sqrt{\frac{(n_1 - 1) s_1^2}{n_1 Y_1} + \frac{(n_2 - 1) s_2^2}{n_2 Y_2}} \geq \overline{x}_1 - \overline{x}_2 \right] \tag{4}$$

(No caso do teste bilateral $H_0 :\ \mu_1 = \mu_2$ *vs.* $H_A :\ \mu_1 \neq \mu_2$, o valor de prova generalizado seria $\mathbb{P}\left[Z^2 \left(\frac{(n_1 - 1) s_1^2}{n_1 Y_1} + \frac{(n_2 - 1) s_2^2}{n_2 Y_2} \right) \geq (\overline{x}_1 - \overline{x}_2 \right]$)^2.)
A vantagem desta abordagem é a estatística de teste generalizada permitir desenhar um teste exato.
O exemplo acima tipifica muitas situações em que há em jogo parâmetros perturbadores:
Pretendemos testar $H_0 :\ \theta \leq \theta_0$ *vs.* $H_A :\ \theta > \theta_0$, dispondo de uma variável aleatória X cuja distribuição depende do parâmetro de interesse θ mas também de parâmetros perturbadores $\boldsymbol{\eta}$. Vamos então especificar uma estatística de teste generalizada $T = T(X; x, \theta, \boldsymbol{\eta})$, i.e. uma função de X e do seu valor observado x, do parâmetro de interesse θ e do vetor de parâmetros perturbadores $\boldsymbol{\eta}$, que tenha as seguintes propriedades (com adaptações óbvias para o caso de teste unilateral à direita ou de teste bilateral):

- Para qualquer valor fixado x as distribuição amostral de T não depende de $\boldsymbol{\eta}$;
- Para $X = x$, $t = T(x; x, \theta, \boldsymbol{\eta}) = T_{\text{obs.}}(X : x, \theta, \boldsymbol{\eta})$ não depende de $\boldsymbol{\eta}$;
- Para x e $\boldsymbol{\eta}$ fixados, $\mathbb{P}[T(X; x, \theta, \boldsymbol{\eta}) > t]$ é uma função monótona não decrescente de θ.

O valor de prova generalizado p_g é o valor de prova correspondente a esta estatística de teste generalizada, isto é

$$p_g = \mathbb{P}[T(X; x, \theta_0, \boldsymbol{\eta}) > T(x; x, \theta_0, \boldsymbol{\eta})] \tag{5}$$

Os testes que usam estatísticas de teste generalizadas são exatos, o que é consequências de estas dependerem simultaneamente de vetores aleatórios observáveis e dos seus valores observados. Mas, ao invés do que acontece em abordagens bayesianas, os parâmetros não t em estatuto aleatório. O exemplo usado mostra como se pode testar componentes da variância num contexto heterocedástico, sem ser necessário recorrer a malabarismos como os descritos em [22].
Valores de prova generalizados devem-se a Tsui e Weerahandi [30], veja-se também [32, 33, 34], ou a panorâmica no capítulo 9 de [18]. Hanning *et al.* [14] apresentaram métodos gerais para a construção da quantidades pivotais adequados para estimação intervalar generalizada na controversa perspetiva fiducial.
O conceito de estatística generalizada foi sido inventado para resolver as dificuldades originadas por parâmetros perturbadores. Porém a noção de estatística generalizada é geral, não se confina à situação particular de ser necessário descartar ou acomodar parâmetros que perturbam a inferência estatística sobre os parâmetros de interesse, veja-se por exemplo [2].

3 Combinação de valores de prova

Um teste global H_0^* : H_0 é verdadeira $\forall k = 1, ..., n$ *vs.* H_A^* : $\exists k \in \{1, ...n\}$ para o qual H_A é verdadeira, cujo objetivo é decidir a manutenção ou rejeição global de H_0 quando H_A é a alternativa, sendo os dados disponíveis os valores de prova p_k obtidos em n ensaios independentes, é um problema complexo, havendo diversas abordagens, nenhuma das quais pode ser considerada ótima em todas as circunstâncias. De um modo geral a amostra de valores de prova disponíveis é escassa (e frequentemente censurada, por o viés de publicação obliterar valores tradicionalmente não significativos), cf. [25], pelo que tradicionais testes de ajustamento de um modelo uniforme (que, pelo atrás dito, deveria ser truncado) se torna pouco conveniente.
De entre os métodos mais populares, destacamos os que curiosamente foram os primeiros a ser propostos:

- método de Tippett [29]: rejeite-se a hipótese nula H_0^* ao nível α se

$$p_{1:n} < 1 - (1 - \alpha)^{1/n} \tag{6}$$

  pois sob H_0^* tem-se $P_{1:n} \frown Beta(1, n)$.

- método de Fisher [10]: rejeite-se H_0^* ao nível α se

$$-2 \sum_{k=1}^{n} \ln(p_k) > \chi^2_{2n;\ 1-\alpha}, \tag{7}$$

  uma vez que $P_k \frown Uniforme(0, 1) \Longrightarrow -\ln(P_k) \frown Exponencial(1)$.

Apesar de esta ser a indicação habitual, cf. [15], há boas razões para preferir um teste bilateral; de facto, se pensarmos em alternativas que correspondam a desequilibrar a uniforme, do tipo $X_m,\ m \in [-2, 2]$, com função densidade de probabilidade

$$f_m(x) = (1 + m(x - 0.5))\, \mathbb{I}_{(0,1)}(x), \tag{8}$$

valores negativos de m produzirão mínimos inferiores ao aceitável sob a hipótese nula, mas valores positivo de m produzirão o efeito contrário, pelo que parece mais sensato uma "proteção" quer contra valores muito pequenos ($< 1 - \left(1 - \frac{\alpha}{2}\right)^{1/n}$), quer contra valores muito grandes ($> 1 - \left(\frac{\alpha}{2}\right)^{1/n}$).

Por outro lado dever-se-ia, no caso de teste de Tippet, usar como pontos críticos $< 1-\left(1-\frac{\alpha}{2}\right)^{1/\nu}$ e $> 1-\left(\frac{\alpha}{2}\right)^{1/\nu}$, com ν estimado levando em linha de conta o viés de publicação, usando por exemplo como critério de estimação de $\nu - n$ o número de casos necessários para reverter a decisão no teste combinado, cf. [28].
A outra questão que não está devidamente tratada, que também se prende com o viés de publicação, é o modelo uniforme padrão não ser adequado, uma vez que só estão em geral reportados valores de prova p inferiores a 0.05. Faria então mais sentido considerar que sob validade de H_0, os p_k observados são modelados por $Beta(1,1;0,\omega)$, usando como estimador centrado de ω o valor $\frac{n+1}{n}\, P_{n:n}$.
Feitas estas observações, cujas consequências serão analisadas noutro trabalho, retomamos o fio do discurso no veio tradicional, usando os critérios (6) e (7). Em [5] a questão do efeito da ampliação de amostras com pseudo-valores de prova gerados usando algoritmos como

$$p_{n+k} = \min\left(\frac{u_k}{p_k}, \frac{1-u_k}{1-p_k},\right),\ k = 1,..n \tag{9}$$

e

$$p_{2n+k} = p_k + p_{n+k} - \lfloor p_k + p_{n+k} \rfloor,\quad k = 1,..n \tag{10}$$

onde $\lfloor a \rfloor$ é o maior inteiro que não excede a, e $u_k,\ k = 1, 2..., n$ são números pseudo-aleatórios uniformes é detalhadamente discutida. De facto, em [5], estabelece-se que se

$$X_{m_i} = \begin{cases} U & X_{2\,\frac{m_i}{|m_i|}} \\ & \\ 1-\frac{|m_i|}{2} & \frac{|m_i|}{2} \end{cases},\ i = 1, 2 \tag{11}$$

forem variáveis independentes da família definida em (8),

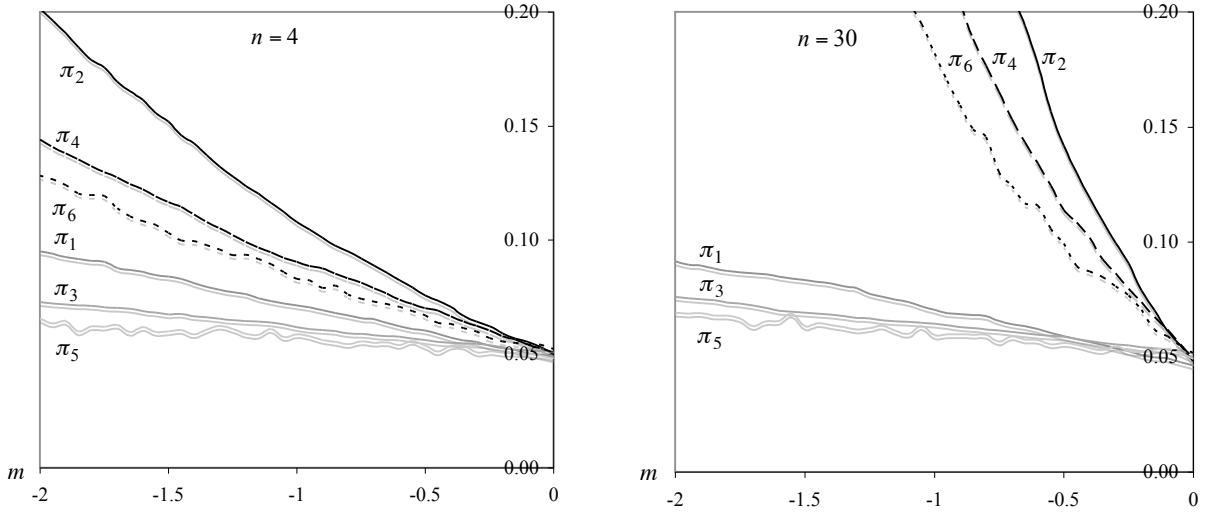
$$V_{m_1,m_2} = \min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}},\right) \ = \ X_{\frac{m_1\, m_2}{6}} = \begin{cases} U & X_{2\,\frac{m_1\, m_2}{|m_1\, m_2|}} \\ & \\ 1-\frac{|m_1\, m_2|}{12} & \frac{|m_1\, m_2|}{12} \end{cases}, \tag{12}$$

onde $X_{-2} \frown Beta(1,2)$ e $X_2 \frown Beta(2,1)$; e também

$$W_{m_1,m_2} = X_{m_1} + X_{m_2} - \lfloor X_{m_1} + X_{m_2} \rfloor = \begin{cases} U & Y \\ & \\ 1-\frac{|m_1\, m_2|}{12} & \frac{|m_1\, m_2|}{12} \end{cases} \tag{13}$$

com $Y \frown Beta(2,2)$, pelo que a ampliação da amostra de valores de prova com pseudo-valores gerados como descrito em (9) e (10), por gerar valores mais consentâneos com a uniforme, tem afinal o inesperado e contraproducente efeito de piorar o desempenho dos testes de uniformidade de Tippett e de Fisher, respectivamente.
O gráfico que apresentamos torna isso bem patente: usando amostras iniciais de $X_m,\ m \in [-2, 0]$, de tamanho $n = 4$ na figura da esquerda, $n = 30$ na figura da direita, seguidamente duplicadas usando (9) e triplicadas usando (10), calculou-se a proporção de *runs* em que a hipótese nula de uniformidade é rejeitada usando o teste de Tippett (π_1, π_2 e π_3, respectivamente para a amostra inicial, para a amostra de tamanho duplo, para a amostra de tamanho triplo), ou rejeitada usando o teste de Fisher (π_2, π_4 e π_6, respectivamente para a amostra inicial, para a amostra de tamanho duplo, para a amostra de tamanho triplo). Repare-se que X_m com $m \in [-2, 0)$ tem uma cauda esquerda mais pesada do que a cauda esquerda uniforme na mesma abcissa, pelo que X_m terá uma maior probabilidade de gerar valores mais próximos de 0, o que se espera que a realização dos testes detete. A observação dos gráficos indica não só que o teste de Fisher tem melhor desempenho do que o de Tippett (o que corresponde a uma opinião geralmente expressa sobre testes combinados),

como também que a duplicação e a triplicação da amostra usando com os algoritmos (9) e (10) diminui a potência — o que é uma consequência do exposto em (12) e (13). De facto, ao aumentar computacionalmente o tamanho da amostra com os referidos algoritmos, estamos a obter sempre amostras mais confundíveis com amostras uniformes.



A investigação deste problema, para além do seu interesse intrínseco, é na prática importante pois há razões para crer que os valores de prova reportados podem não ser uniformes, mas sim extremos de uniformes independentes. De facto, no caso de um experimentador obter resultados que não abonam a favor do que quer estabelecer, pode ter a tentação de proceder a um novo ensaio, e reportar o melhor dos resultados (em geral será o mínimo de uniformes independentes). Ora se o modelo apropriado para $P_k \stackrel{\mathrm{d}}{=} X_m$ com X_m como descrito em (11), trata-se de uma situação em que $\frac{|m|}{2}$ dos casos foi reportado o mínimo (se $m \in [-2, 0)$) ou o máximo (se $m \in (0, 2]$) de dois valores de prova obtidos em experiências independentes. Pires e Branco [26] usam este modelo para iluminar a controvérsia resultante da crítica de Fisher aos resultados "demasiado bons" que Mendel usou nos seus trabalhos pioneiros de Genética. Os nossos resultados mostram que na prática destrinçar essa "batota" da situação em que os valores de prova p_k são genuinamente observações de réplicas independentes da uniforme padrão é muito pouco seguro, e que aumentar a dimensão da amostra não é um remédio infalível, uma vez que pode diminuir a proporção de casos em que houve o reporte de um extremo de duas observações.

Por outro lado, na prática estatística admite-se que a amostra aleatória é constituída por unidades independentes. No caso de se estar a verificar a integridade de um trabalho, interessa por isso testar não só a uniformidade dos P_k como a sua independência. Uma forma interessante de colocar a questão é definir um processo auto-regressivo

$$X_k = \rho\, X_{k-1} + (1-\rho)\, U_k, \quad \rho \in (0,1),\ k = 1, 2, \ldots \tag{14}$$

onde $\rho \in [0, 1]$ e $X_0, , U_1, U_2, ...$ são réplicas independentes de $U \frown Uniforme(0, 1)$ e testar $H_0 :\ \rho = 0\ vs.\ H_A :\ \rho \in (0, 1]$.

Observe-se que

$$U_k = \frac{X_k - \rho\, X_{k-1}}{1-\rho},\ k = 1, ..., n, \tag{15}$$

de que facilmente se deduz que a função densidade de probabilidade conjunta de $X_1, ..., X_n$ é

$$f_{(X_1,...,X_n)}(x_1, ..., x_n) = \frac{1}{(1-\rho)^n}\, \mathbb{I}_{\boldsymbol{S}} \tag{16}$$

com $\boldsymbol{S} = \bigcap_{i=1}^{n} \left\{ 0 < \dfrac{x_i - \rho\, x_{i-1}}{1-\rho} < 1 \right\} = \left\{ (x_1, ..., x_n) : \min_{1\le k\le n} \min \left(\dfrac{x_k}{x_{k-1}}, \dfrac{1-x_k}{1-x_{k-1}} \right) \geq \rho \right\}$.

Uma vez que, como atrás referimos, $\min \left(\frac{U_k}{U_{k-1}}, \frac{1-U_k}{1-U_{k-1}} \right) \frown Uniforme(0,1)$, imediatamente se conclui da definição do processo auto-regressivo $\{X_k\}$, $k = 1, 2, ...$, que

$$\min \left(\frac{X_k}{X_{k-1}}, \frac{1-X_k}{1-X_{k-1}} \right) = \rho + (1-\rho) \min \left(\frac{U_k}{U_{k-1}}, \frac{1-U_k}{1-U_{k-1}} \right) \frown Uniforme(\rho, 1).$$

Consequentemente

$$\min_{1\le k\le n} \min \left(\frac{X_k}{X_{k-1}}, \frac{1-X_k}{1-X_{k-1}} \right) \frown Beta(1, n; \rho, 1). \tag{17}$$

A generalização para processos auto-regressivos construídos com variáveis da família X_m definida em (11) segue padrões idênticos, veja-se [6, 7].

Vamos agora exibir um exemplo que mostra que o recurso a valores de prova generalizados pode melhorar o desempenho de um teste. Suponha-se que pretendemos testar $H_0 : P_k \frown Uniforme(0,1)$ *vs.* $H_A : P_k \stackrel{\mathrm{d}}{=} X_m,\ m \in [-2, 0)$, onde X_m é a subclasse de variáveis aleatórias com densidade de declive negativo definida em (8).
Brilhante [2], não tendo obtido uma estatística suficiente adequada para servir como estatística de teste, usou o método exposto em [14] para construir uma estatística de teste generalizada:

$$T = T(\boldsymbol{X}; \boldsymbol{x}, m) = m - \frac{V + 2\sum_{k=1}^{n} \ln(x_k)}{n\,(1-\overline{x})} \tag{18}$$

onde $V = -2\sum_{k=1}^{n} \ln(F_m(X_k) \frown \chi^2_{2n}$ é uma variável fulcral invertível.
Assim, o valor de prova generalizado é

$$p_g = \mathbb{P}[T \leq 0 | m = 0] = 1 - [v \leq -2 \sum_{k=1}^{n} \ln(x_k) \tag{19}$$

sendo a potência do teste

$$\pi[\boldsymbol{x}; m) = \mathbb{P}[T(\boldsymbol{X}; \boldsymbol{x}, m) \leq 0 | m] = 1 - [P[V \leq mn\,(1-\overline{x}) - 2\sum_{k=1}^{n} \ln(x_k)] \tag{20}$$

Observe-se que há uma ligeira vantagem quando comparado com o método de Fisher, apontado como o que tem melhor desempenho numa grande generalidade de situações.

4 Conclusão

A investigação de valores de prova aleatórios levou-nos colateralmente a concluir que afinal a ampliação computacional das amostras pode não ser uma boa ideia, pois é possível que características estruturais do modelo — em particular, a entropia máxima da uniforme padrão, na classe das leis com suporte (0,1) — tenham um efeito perverso na potência dos testes que se pretende fazer.
Por outro lado, o uso de estatísticas de teste generalizadas, que foram introduzidas para lidar com parâmetros perturbadores mas que podem também ser usadas no caso de o conjunto de parâmetros perturbadores ser vazio, pode melhorar o desempenho de testes, nomeadamente de testes de uniformidade, de grande importância nas sínteses meta-analíticas usando valores de prova combinados.

Agradecimentos

Trabalho financiado por fundos nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projecto PEst-OE/MAT/UI0006 /2011.

Referências

[1] Boos, D. D., and Stefanski, L. A.. (2011). *P*-value precision and reproducibility, *The American Statistician* **65**, 213–221.

[2] Brilhante, M. F. (2013). Generalized p-values and random p-values when the alternative to uniformity is a mixture of a Beta(1,2) and Uniform. In Oliveira, P. et al. (eds), Recent Developments in Modeling and Applications in Statistics, Springer, Heidelberg, pp 159-167.

[3] Brilhante, M. F., Pestana, D. and Sequeira, F.: Combining p-values and random p-values, In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the 32nd International Conference on Information Technology Interfaces*, 515–520. (2010).

[4] Brilhante, M. F., Malva, M., Mendonça, S., Pestana, D., Sequeira, F., and Velosa, S. (2013). Uniformity. In Lita da Silva, J.; Caeiro, F.; Natário, I.; Braumann, C.A. (Eds.): *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and other Statistical Applications*, 73-81, Springer, Berlin.

[5] Brilhante, M. F., Pestana, D. and Sequeira, F.: Combining p-Values: an Overview, with a Discussion on Uniformity and on Mixtures of Uniform and Beta (1,2) or Beta (2,1), *Notas e Comunicações do CEAUL* 23/13.

[6] Brilhante, M. F., Pestana, D., and Sequeira, F. (2013). Auto-regressive extensions of uniform randomness characterisation, Abstracts Book of the 7th Workshop on Statistics, Mathematics and Computation, 38-39.

[7] Brilhante, M. F., Pestana, D., and Sequeira, F. (2013). Auto-regressive extensions of uniform randomness characterisation, *Notas e Comunicações do CEAUL* 23/13.

[8] Dempster, A., and Schatzoff, M. (1965). Expected significance level as a sensitivity index for test statistics. *J. Amer. Statist. Ass.* **60**, 420–436.

[9] Donahue, R. (1999). A note on information seldom reported via the p-value. *The American Statistician* **53**, 303–306.

[10] Fisher, R. A. (1935). The fiducial argument on statistical inference. Ann. Eugen. 6, 391–398.

[11] Gelman, A., and Stern, H. (2006). The difference between "significant" and"not significant" is not itself statistically significant, *The American Statistician* **60**, 328–331.

[12] Gomes, M. I, Pestana, D., Sequeira, F., Mendonça, S., and Velosa, S.: Uniformity of offsprings from uniform and non-uniform parents. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the 31st International Conference on Information Technology Interfaces*, 243–248 (2009).

[13] Goodman, S. N. (1992). A comment on replication, p-values and evidence, *Statistics in Medicine* **11**, 875–879 (ver também o comentário de S. Senn no mesmo jornal, **21** 2437–2444).

[14] Hanning, J., Iyer, H. K., and Patterson, P. D. (2006). Fiducial generalized confidence Intervalas, *J. Amer. Statist. Ass.* **101**, 254–269.

[15] Hartung, J., Knapp, G. and Sinha, B.K., (2008). *Statistical Meta-Analysis with Applications*, Wiley, New York.

[16] Hung, H., O'Neill, R., Bauer, R., and Kohne, K. (1997). The behaviour of the p value when the alternative is true. *Biometrics* **53**, 11–22.

[17] Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Ann. Eugen.* **10**, 48–51.

[18] Khuri, A. I., Mathew, T. and Sinha, B. K. (1998) Tests Using Generalized P-Values, in Statistical Texts for Mixed Linear Models, Wiley, Hoboken, NJ.

[19] Kulinskaya, E., Morgenthaler, S., and Staudte, R. G. (2008) *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*, Wiley, Chichester (2008).

[20] Mathew, T., Sinha, B. K., and Zhou, L. (1993). Some statistical procedures for combining independent tests. *J. Amer. Statist. Ass.* **88**, 912–919.

[21] Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). *P*-values are random variables, *The American Statistician* **62** , 242–245.

[22] Oehlert, G. W. (2000). *A First Course in Design and Analysis of Experiments*, Freeman, New York.

[23] Pestana, D., e Velosa, S. (2010). *Introdução à Probabilidade e à Estatística*, Fundação Calouste Gulbenkian, Lisboa.

[24] Pestana, D. (2011). Combining p-values. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, 1145-1147, Springer, New York

[25] Pestana, D., Rocha, M. L., Vasconcelos, R., and Velosa, S. (2013). Publication Bias and Meta-Analytic Syntheses. In Lita da Silva, J.; Caeiro, F.; Natário, I.; Braumann, C.A. (Eds.): *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and other Statistical Applications*, 347-354, Springer, Berlin.

[26] Pires, A.M., Branco, J.A.: A statistical model to explain the Mendel-Fisher controversy. Statistical Science **25** 545–565 (2010)

[27] Sackrowitz, H., and Samuel-Cahn, E. (1999). *P* values as random variables-expected *P* values. *The American Statistician* **53**, 326–331.

[28] Sequeira, F.: *Meta-Análise: Harmonização de Testes Usando os Valores de Prova*. PhD Thesis, DEIO, Faculdade de Ciências da Universidade de Lisboa (2009).

[29] Tippett, L. H. C. : *The Methods of Statistics*, Williams & Norgate, London (1931).

[30] Tsui, K., and Weerahandi, S. (1989). Generalized *p*-values in significance testing of hypothesis in the presence of nuisance parameters. *The American Statistician* **84**, 602–607.

[31] Utts, J. (1991). Replication and Meta-Analysis in Parapsychology, *Statistical Science* **6**, 363–403.

[32] Weerahandi, S. (1993). Generalized confidence intervals. *The American Statistician* **88**, 889–905.

[33] Weerahandi, S. (1995). *Exact Statistical Methods for Data Analysis*, Springer, New York.

[34] Weerahandi, S. (2004). *Generalized Inference in Repeated Measures*, Springer, New York.

[35] Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–361.