BETICES

Maria de Fátima Brilhante
Universidade dos Açores, DM, e CEAUL

Dinis Duarte Pestana
Universidade de Lisboa, DEIOe CEAUL

Maria Luísa Rocha
Universidade dos Açores, DEG e CEEAplA

Sumário: Sob validade de H_0 os valores de prova são observações de uma amostra aleatória uniforme padrão, pelo que testar a validade global de H_0 é equivalente a testar a uniformidade dos valores-p. A investigação de misturas de $Beta(1,1)$ — ou seja, uniforme padrão — com $Beta(1,2)$ ou com $Beta(2,1)$ mostra que a tentação de aumentar computacionalmente a amostra de valores-p num teste em que a alternativa é uma mistura apropriada de betas não resulta, devido ao carácter extremal da entropia da uniforme padrão entre as variáveis aleatórias com suporte em (0,1). Misturas mais gerais de betas levam à definição de uma nova família de variáveis aleatórias, conducentes a modelos populacionais que generalizam o de Verhulst, mais adequados para situações de crescimento rápido, como os de populações de tumores neoplásicos.

Palavras-chave: uniformes, betas, misturas de betas, crescimento populacional.

Abstract: As under the null hypothesis H_0 the available p-values are observations from a standard uniform random sample, testing the overall validity of H_0 is equivalent to testing uniformity of the p-values. The analysis of mixtures of $Beta(1,1)$ — i.e., standard uniform — with $Beta(1,2)$ or $Beta(2,1)$ shows that computational augmentation of small samples of p-values, when testing uniformity versus sensible mixtures of betas, doesn't work as expected, due to the extremal entropy of the standard uniform among all random variables with support in (0,1). More general mixtures of Beta distributions lead to the definition of a new family of random variables, tied to populational growth models generalizing the Verhults logistic model; the solution of the corresponding differential equation is the extremal Gumbel law, appropriate to model quick growth as in populations of neoplastic tumor cells.

Key-words: Uniform, beta random variables, mixtures of beta, population growth.

1 O teorema da transformação uniformizante, meta-análise e valores de prova combinados

A família de variáveis aleatórias $X_{p,q} \frown Beta(p,q)$, $p,q>0$, com funções densidade de probabilidade $f_{X_{p,q}}(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}\,\mathbb{I}_{(0,1)}(x)$, onde $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\,\mathrm{d}x$ é a função Beta de Euler, é muito versátil, prestando-se por isso a modelar uma

grande diversidade de fenómenos. Destacam-se nesta família a uniforme padrão $X_{1,1}$, que pela sua singularidade denotaremos $U \frown Uniforme(0,1)$, e $X_{2,2}$, esta última associada ao modelo populacional logístico.

Conta-se que Linus Pauling, quando um entrevistador lhe perguntou inopinadamente o que é necessário fazer para ganhar um prémio Nobel, respondeu quase de imediato: "Ter muitas ideias, e a coragem de deitar fora quase todas." Nesta resposta notável reconhece-se importância fundamental da teoria dos testes de hipóteses na construção do conhecimento a partir de informação empírica. No entanto a teoria dos testes de hipóteses é usada de forma simplista por muitos investigadores. Apesar de uma análise estatística ser actualmente um requisito incontornável na publicação científica, na prática o que se procura é "um $p-value$ significativo", em detrimento de um planeamento experimental adequado, com consequências adversas ao progresso da Ciência. De facto, uma consequência insuficientemente discutida do teorema da transformação uniformizante é que se a hipótese nula H_0 for verdadeira, o valor de prova p correspondente ao valor observado da estatística de teste T é uma observação de uma uniforme padrão. Assim

• ao usar como critério de publicação a obtenção de um *p-value* "significativo" cria-se um "viés de publicação" que tem uma influência nefasta em estudos de síntese, por estudos com resultados não-significativos estarem subrepresentados;

• como, afinal, sob validade de H_0 o valor de prova p pode ser um valor muito próximo de 0 (ou de 1), por ser uma observação da uniforme padrão, é imprescindível ter em linha de conta que um dos critérios incontornáveis na construção das ciências experimentais é a *repetibilidade* — recomendamos o notável trabalho de Utts (1991), que discute em profundidade o requisito de repetibilidade na construção de conhecimento válido.

Glass (1976) trouxe um progresso notável ao paradigma da ciência com o conceito de meta-análise, a teoria de como harmonizar resultados porventura discordantes, e de sintetizar informação insuficiente por forma a extrair informação conclusiva de vários estudos em si mesmo inconclusivos ou polémicos, cf. Pestana (2010) para uma panorâmica e novos resultados sobre combinação de valores de prova. Na secção 2, trazemos novas possibilidades à meta-análise investigando criticamente os efeitos de aumentar computacionalmente a amostra de valores-p; os resultados são parcialmente decepcionantes, mas por outro lado revelam a que ponto o carácter extremal da entropia da uniforme é uma propriedade relevante.

Por outro lado, na Secção 3 introduzimos uma "$Betinha(p,q)$", e modelos de crescimento populacional associados, discutindo crescimento de Gompertz como um descontrole do crescimento logístico de Verhults, apropriados no caso de crescimento desregrado como o que ocorre em desenvolvimentos tumorias maligno.

2 A Família X_m de misturas de Uniforme com Beta(1,2) ou Beta(2,1)

O gráfico da função densidade de probabilidade uniforme padrão no intervalo (0,1) é um segmento de recta horizontal, com ordenada na origem 1. É evidente que qualquer segmento de recta passando por $(\frac{1}{2}, 1)$, com declive $m \in [-2, 2]$, continua a ser uma função densidade de probabilidade de uma variável aleatória X_m com suporte em $\boldsymbol{S} = (0, 1)$. É também evidente que $m < p \Rightarrow X_m \prec X_p$; consequentemente, devido ao viés de publicação, quando o objectivo é avaliar a validade de uma hipótese nula H_0 com base na observação de valores de prova $\{p_k\}_{k=1}^n$, parece adequado testar a uniformidade dos $\{p_k\}_{k=1}^n$ — isto é, $m = 0$ — *vs.* uma alternativa à esquerda, $m \in [-2, 0)$, ou uma alternativa à direita, $m \in (0, 2]$. Por outras palavras, estamos a considerar misturas de uniformes com $Beta(1, 2)$ ou com $Beta(2, 1)$, consoante $m \in [-2, 0)$ ou $m \in (0, 2]$.

Como se $U \frown Uniforme(0, 1)$ e X com suporte em $\boldsymbol{S} = (0, 1)$ forem variáveis aleatórias independentes, $V = U + X - \boldsymbol{I}(U + X) \frown Uniforme(0, 1)$ (onde $\boldsymbol{I}(U+X)$ denota a parte inteira de $U + X$) e $W = \min\left\{\frac{U}{X}, \frac{1-U}{1-X}\right\} \frown Uniforme(0, 1)$, com V e W independentes de X, Gomes *et al.* (2009) investigaram o aumento computacional da amostra dos $\{p_k\}_{k=1}^n$ à custa de pseudo-aleatórios uniformes,

Quando se considera o efeito deste aumento artificioso da amostra sobre a potência dos testes, os resultados são deveras decepcionantes. A estranheza deste resultado não só alerta para o facto em geral ignorado de que aumentar a amostra pode piorar a análise estatística, como exige a pesquisa de uma explicação.

Em Brilhante *et al.* (2010a) estudam-se produtos e potências de produtos de variáveis aleatórias, mostrando em particular que se $U_1 \stackrel{\mathrm{d}}{=} U_2 \stackrel{\mathrm{d}}{=} U_3 \frown Uniforme(0, 1)$, independentes e independentes de X_{-2}, quer $(U_1\, U_2)^{U_3}$ quer $(U_1\, U_2\, U_3)^{X_{-2}}$ são uniformes padrões. Mais geralmente, a função densidade de probabilidade de $(X_m\, X_p)^{X_r}$, onde X_m, X_p, X_r são independentes, tem valores muito próximos de 1 no suporte $\boldsymbol{S} = (0, 1)$.

Por outro lado, se X_m, X_p forem independentes, $W_{m,p} = \min\left\{\frac{X_m}{X_p}, \frac{1-X_m}{1-X_p}\right\} \stackrel{\mathrm{d}}{=} X_{\frac{mp}{6}}$, bastando assim que X_m ou X_p seja uniforme (i.e., $m = 0$ ou $p = 0$) para se obter uma uniforme (cf. Brilhante *et al.* (2010b)

Assim, a uniforme parece atrair funções simples das misturas X_m de uniformes e $Beta(1, 2)$ ou $Beta2, 1)$; de facto, qualquer das operações consideradas aumenta a entropia, e a uniforme é a lei de entropia máxima no suporte $\boldsymbol{S} = (0, 1)$ (Kagan *et al.*, (1973)).

3 Betinhas e crescimento populacional

Uma vez que $\forall p,q>0,\ \int_0^1 x^{p-1}(-\ln x)^{q-1}\mathrm{d}x = \frac{\Gamma(q)}{p^q}$ e $x^{p-1}(-\ln x)^{q-1}\geq 0,\ \forall x\in(0,1)$, segue-se que $\forall p,q>0\ f_{X^*_{p,q}}(x) = \frac{p^q}{\Gamma(q)}\,x^{p-1}(-\ln x)^{q-1}\,\mathbb{I}_{(0,1)}(x)$ é uma função densidade de probabilidade, de uma variável aleatória $X^*_{p,q}$ a que chamaremos neste trabalho "*Betinha*(p,q)". O valor médio é $\mathbb{E}[X^*_{p,q}] = \left(\frac{p}{1+p}\right)^q$, função decrescente de q para p fixo; observe-se ainda que se $p,q>1$, $\mathbb{E}[X^*_{p,q}] < \mathbb{E}[X_{p,q}]$. Note também que $\forall x\in(0,1),\ \mathbb{P}[X^*_{p,q}\leq x]\geq\mathbb{P}[X_{p,q}\leq x]$, pelo que diremos que para $p,q>1$ se tem $X^*_{p,q}\prec X_{p,q}$.

O caso $q=2$ reveste-se de particular interesse. De facto, de $-\ln x = \sum_{k=1}^{\infty}\frac{(1-x)^k}{k}$, conclui-se que

• A função densidade de probabilidade da *Beta*$(2,2)$ (proporcional à "parábola logística") é, a menos de um factor normalizador, uma aproximação de grau 2 da função densidade de probabilidade da *Betinha*$(2,2)$. Mais geralmente, a função densidade de probabilidade da *Beta*$(p,2)$ é, a menos de um factor normalizador, uma aproximação de grau p+1 da função densidade de probabilidade da *Betinha*$(p,2)$.

• A *Betinha*$(2,2)$ é uma mistura convexa de *Beta*$(2,k),\ k=2,3,\ldots$, com pesos $w_k=\frac{4}{(k-1)k(k+1)},\ k=2,3,\ldots$. (E, mais geralmente, a *Betinha*$(p,2)$ é uma mistura convexa de *Beta*$(p,k),\ k=2,3,\ldots$, com pesos $w_k=\frac{p^2\,\Gamma(p)\,\Gamma(k)}{(k-1)\,\Gamma(p+k)},\ k=2,3,\ldots$)

É bem conhecida a dinâmica de $x_{n+1}=r\,x_n\,(1-x_n)$, e as suas implicações para a dinâmica de populações; a dinâmica de $x_{n+1}=r\,x_n\,(1-x_n)^{k-1}$ foi extensivamente estudada em Aleixo *et al.* (2009, 2010).

A equação diferencial de Verhulst $\frac{\mathrm{d}}{\mathrm{d}t}N(t) = rN(t)[1-N(t)]$, que corresponde à equação às diferenças $x_{n+1} = r\,x_n\,(1-x_n)$ quando se considera um modelo hierárquico, assumindo que $x\in(0,1)$ é o valor de uma função de distribuição N, tem como solução a função logística, que descreve adequadamente a dinâmica de muitas populações naturais.

Quando em lugar de partir de $r\,x\,(1-x)$ — que admite que o crescimento causado pela taxa de reprodução malthusiana é moderado por uma retroacção de segundo grau reflectindo a limitação de recursos naturais — se toma como modelo de partida $r\,x\,(-\ln x)$, a correspondente equação diferencial

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = rN(t)[-\ln(N(t))]$$

tem a solução $N(t)=\exp[-\exp(-rt)]$, conhecida em estudos populacionais como função de Gömpertz, e em estatística como distribuição de Gumbel. A distribuição de Gumbel é uma das leis estáveis para máximos de variáveis i.i.d., a cujo domínio de atracção pertencem por exemplo quer a lei gaussiana quer a lei exponencial, com

cauda direita moderadamente pesada, e tem-se revelado uma boa adaptação para o crescimento da população de células em tumores malignos.

A abordagem apresentada evidencia que o modelo logístico se obtém quando $x \ln x$ é aproximado por $x(1-x)$, resultando num efeito de retroacção mais forte. Assim, o modelo $Beta(2,2)$ está associado a um crescimento sustentado, enquanto o modelo $Betinha(2,2)$ está associado a um crescimento desregrado. É por isso do maior interesse investigar as implicações deste modelo populacional no que se refere à dinâmica da população humana, que se sabe que tem excedido consistentemente as projecções decorrentes do modelo de crescimento logístico.

Bibliografia

• Aleixo, S., Rocha, J.L., and Pestana, D. (2009). Populational Growth Model Proportional to Beta Densities with Allee Effect, in Cabada, A., Liz, E., and Nieto, J. J. (Eds.), *Mathematical Models in Engineering, Biology, and Medicine*, American Institute of Physics ISBN: 978-0-7354-0660-5, 3–12.

• Aleixo, S., Rocha, J.L., and Pestana, D. (2010). Populational Growth Model Proportional to Beta Densities, in Peixoto, M. M; Pinto, A. A.; Rand, D. A. J. (Eds.), *Dynamics, Games and Science, in honour of Maricio Peixoto and David Rand* (in print), ISBN: 978-3-642-11455-7.

• Brilhante, M. F., Mendonça, S., Pestana, D., and Sequeira, F. (2010). Using Products and Powers of Products to Test Uniformity, in Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2010*, 509–514.

• Brilhante, M. F., Pestana, D., and Sequeira, F. (2010b). Combining p-values and random p-values, in Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2010*, 515–520.

• Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Education Research*, **5**, 3–8.

• Gomes, M. I, Pestana, D., Sequeira, F., Mendonça, S., and Velosa, S. (2009). Uniformity of offsprings from uniform and non-uniform parents, in Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2009*, 243–248.

• Kagan, A. M., Linnik, Yu. V., and Rao, C. R. (1973). *Characterization Problems in Mathematical Statistics*, Wiley, New York.

• Pestana, D. (2010). Combining p-values, in M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer Verlag (in print), ISBN: 978-3-642-04897-5.

• Utts, J. (1991). Replication and meta-analysis in parapsychology, *Statistical Science* **6**, 363–403.