Nota sobre o problema da comparação das médias de dois universos normais

por M. A. Fernandes Costa

1. Designem X e Y duas variáveis independentes e da mesma natureza que nos respectivos universos $\mathcal{U}$ e $\mathcal{V}$ têm distribuições normais,

$$X \frown N(\mu_1, \sigma_1) \quad \text{e} \quad Y \frown N(\mu_2, \sigma_2)\,.\,(^1)$$

Suponhamos colhidas em $\mathcal{U}$ e $\mathcal{V}$ amostras de valores de X e Y, respectivamente $(x_1, x_2, \cdots, x_m)$ e $(y_1, y_2, \cdots, y_n)$. O problema estatístico aqui encarado é o de inferir, a partir destas amostras, conclusões sobre a plausibilidade da hipótese $H_0: \mu_1 = \mu_2$ no caso em que se desconheçam tanto estas médias como σ_1 e σ_2.

Para isso servirão as médias

$$\bar{x} = \frac{1}{m}\sum_1^m x_i \quad , \quad \bar{y} = \frac{1}{n}\sum_1^n y_i$$

e as variâncias

$$s_1^2 = \frac{1}{m}\sum_1^m (x_i - \bar{x})^2 \quad , \quad s_2^2 = \frac{1}{n}\sum_1^n (y_i - \bar{y})^2$$

das distribuições observadas nas amostras.

2. Como se sabe, o ensaio de uma hipótese fundamenta-se sempre numa «estatística» (função das observações) cuja distribuição tenha uma densidade de probabilidade que, além de teórica e numèricamente tratável, seja *independente dos parâmetros desconhecidos*.

No caso simples em que m e n são suficientemente grandes (digamos $\geqslant 30$), o problema referido resolve-se satisfatòriamente por recurso à estatística

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}\,;$$

visto que $E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_1 - \mu_2$, o numerador é uma variável normal de média nula quando seja verdadeira a hipótese H_0 (¹). Por outro lado, $Var(\bar{x} - \bar{y}) = \dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}$ e, quando m e n assumem valores elevados, $\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}$ é uma boa estimativa desta variância. Daqui resulta (¹) que se tem aproximadamente

$$u \mid \mu_1 = \mu_2 \frown N(0, 1)$$

e o ensaio de H_0 faz-se então muito simplesmente com o auxílio duma tabela da distribuição normal.

3. O caso em que m e n têm valores pequenos foi primeiro solucionado por R. A. Fisher (1925) com a restrição adicional de se supor $\sigma_1 = \sigma_2 (= \sigma)$.

Fisher observou que

$$\hat{\sigma}^2 = \frac{m\,s_1^2 + n\,s_2^2}{m + n - 2}$$

é então uma estimativa «inviciada» de σ^2 ($i \cdot e \cdot$ tal que $E(\hat{\sigma}^2) = \sigma^2$) — sendo portanto $\hat{\sigma}^2\left(\dfrac{1}{m} + \dfrac{1}{n}\right)$ uma estimativa inviciada de

---

(¹) Com $X \frown N(\mu, \sigma)$ pretende-se significar uma variável normal de média μ e variância σ^2. Num caso geral, pode escrever-se $X \frown f(x)$ para indicar que a densidade de probabilidade da distribuição de X é $f(x)$ e $X \mid H \frown f(x)$ para exprimir que isso só é verdade quando se verifique a hipótese H.

Muito se gostaria de ver generalizada tão clara e sucinta notação.

(¹) Com efeito, sabe-se que, se for $Z_1 \frown N(m_1, d_1)$ e $Z_2 \frown N(m_2, d_2)$, se tem

$$a\,Z_1 + b\,Z_2 \frown N\left(a\,m_1 + b\,m_2\,, \sqrt{a^2 d_1^2 + b^2 d_2^2}\right)$$

$Var(\bar{x} - \bar{y})$ — e utilizou em lugar de u a estatística

$$v = \frac{\bar{x} - \bar{y}}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}};$$

depois provou que, sendo verdadeira a hipótese H_0, a distribuição de v é a de «Student» com $m+n-2$ graus de liberdade (1), o que indicaremos escrevendo

$$v \mid \mu_1 = \mu_2 \frown t\,(m+n-2)\,.$$

Na resolução prática do problema pode-se assim tirar partido das tabelas da distribuição «t» de «Student».

O ponto fraco desta solução reside na premissa um tanto arbitrária de ser $\sigma_1 = \sigma_2$, o que nem sempre pode justificar-se na prática — quer por considerações apriorísticas quer por ensaio estatístico da hipótese $\sigma_1 = \sigma_2$.

4. Situa-se precisamente aqui a fonte de uma das mais interessantes e importantes controvérsias que caracterizam o desenvolvimento recente da Estatística Matemática. O problema do ensaio da hipótese $\mu_1 = \mu_2$ sem aceitação da premissa $\sigma_1 = \sigma_2$ (o qual ficou a ser conhecido por «problema de Behrens-Fisher») for também primeiro resolvido por Fisher [3] à luz da sua teoria da *fidutial inference*. Esta teoria, porém, está muito longe de ser universalmente aceite, a ela se opondo uma outra, de inspiração frequencista, criada por J. Neyman e E. S. Pearson à roda de 1930. (2)

Em muitos casos, os resultados das duas teorias coincidem, mas instâncias surgiram de profundas e graves divergências que muita tinta têm feito correr. Todavia, no caso particular do problema de Behrens-Fisher, a posição dos partidários da teoria de Neyman-Pearson era bastante incómoda: embora contestassem a validade da solução de Fisher, não sabiam opor-lhe outra. Quem conhecer a faceta polémica do grande criador da Estatística moderna não estranhará que ele se valesse desta posição de vantagem para propagandear as suas ideias e lançar comentários mordazes aos seus adversários; até que de entre estes já havia quem no seu desânimo sugerisse existirem com certeza problemas não susceptíveis de solução...

Foi só em 1947 que B. L. Welch deu a lume um tratamento definitivo da questão sem recurso à teoria da *fidutial inference*. O seu artigo [7], se não se presta a fácil leitura, é porém de importância capital; e a aplicação prática da solução de Welch ficou possibilitada com a publicação de tabelas acompanhadas de instruções práticas para o seu uso, de resto muito simples [8].

5. O objectivo desta nota é, principalmente, chamar a atenção para a memória de Welch, que não parece ter ainda tido a difusão que merece; mas aproveita-se a oportunidade para expor uma outra solução — e muito engenhosa, se bem que de menor interesse prático — que para o problema de Behrens-Fisher deu H. Scheffé [9].

Considere-se a estatística

$$w = \sqrt{m\,(m-1)}\,\frac{\bar{x} - \bar{y}}{\sqrt{\sum\limits_{1}^{m}(z_i - \bar{z})^2}},$$

onde

$$z_i = x_i - \sqrt{\frac{m}{n}}\,y_i \quad \text{e} \quad \bar{z} = \frac{1}{m}\sum_{1}^{m} z_i$$

(supõe-se $m \leqq n$, o que é evidentemente legítimo). w serve excelentemente para utensilhar um ensaio da hipótese $H_0: \mu_1 = \mu_2$, pois se tem

$$w \mid \mu_1 = \mu_2 \frown t\,(m-1)\,,$$

(1) Pode ver-se uma simples e elegante demonstração deste resultado no belo livro de Weatherburn [1], onde tão bem se introduzem os aspectos puramente matemáticos da Estatística. Veja-se também [2], pág. 56.

(2) O leitor desejoso de se pôr a par das idéias que servem de fundamento às modernas teorias do ensaio de hipóteses e da estimação encontrará exposições sucintas e elementares em [4] e [5]. Também estudará [6] com muito proveito.

sendo portanto a sua distribuição independente de σ_1 e σ_2.

Julga-se de certo interesse dar deste resultado uma demonstração elementar directa, pois SCHEFFÉ fá-lo depender de um enunciado mais geral.

Para isso, comecemos por notar que

$$Var\ (\bar{x}-\bar{y})=\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}=\tau^2$$

$$Var\ (z_i)=\sigma_1^2+\frac{m}{n}\sigma_2^2=m\,\tau^2$$

e

$$E\,(z_i)=\mu_1-\sqrt{\frac{m}{n}}\,\mu_2=0,$$

desde que para comodidade se escolha a origem de modo que o valor comum de μ_1 e μ_2 seja nulo; tem-se então

$$z_i \frown N(0\,,\sqrt{m}\,\tau)\,.$$

Ora, recorrendo a uma transformação linear ortogonal (cf. [1], págs. 164 e 169), é fácil mostrar que

$$\sum_1^m (z_i-\bar{z})^2=\sum_1^{m-1}\zeta_i^2\,,$$

onde as variáveis $\zeta_1\,,\zeta_2\,,\cdots,\zeta_{m-1}$ são independentes e

$$\zeta_i \frown N(0\,,\sqrt{m}\,\tau)\,;$$

logo,

$$\left.\sum_1^{m-1}\frac{\zeta_i^2}{m\,\tau^2}\right|\,\mu_1=\mu_2 \frown \chi^2\,(m-1)\,,$$

em que $\chi^2(m-1)$ designa a distribuição do χ^2 com $m-1$ graus de liberdade (cf. [2], pág. 54).

Por conseguinte,

$$\frac{w^2}{m-1}=\frac{(\bar{x}-\bar{y})^2/\tau^2}{\Sigma\,(z_i-\bar{z})^2/m\,\tau^2}$$

é o quociente de uma variável $\chi^2(1)$ por uma variável $\chi^2(m-1)$, e isto permite concluir o resultado enunciado ([1], pág. 187).

6. Por último, referir-se-á a curiosa solução de DARMOIS [10], a qual se baseia na construção de «regiões de confiança» para o par de parâmetros $(\mu_1\,,\mu_2)$.

Esta solução parece no entanto conter alguns pontos obscuros. Em particular, não está provado que o tipo de regiões de confiança escolhido conduza a ensaios eficientes da hipótese $\mu_1=\mu_2$. Por outro lado, não é possível fixar prèviamente um valor do coeficiente de risco, mas apenas um limite superior da probabilidade de cometer um «erro de primeira espécie».

REFERÊNCIAS

[1] C. E. WEATHERBURN, *A First Course in Mathematical Statistics*, 2.ª ed. Cambridge University Press, 1949.

[2] M. A. FERNANDES COSTA, «Sobre o Cálculo da Distribuição de uma Variável Casual», *Boletim do Instituto dos Actuários Portugueses*, n.º 8.

[3] R. A. FISHER, «The Fidutial Argument in Statistical Inference», *Ann. Eugenics*, vol. 6, 1935, págs. 391-398. (Reproduzido numa selecta dos trabalhos de Fisher: *Contributions to Mathematical Statistics,* J. Wiley & Sons, New York, 1950).

[4] C. I. K. FORESTER, «An Elementary Introduction to the Testing of Statistical Hypotheses», *Jour. Inst. of Actuaries Students' Society,* vol. VII, págs. 81-97.

[5] L. SOLOMON, «Statistical Estimation», idem, vol. VII, págs. 144-173 e 213-234.

[6] GUSTAVO DE CASTRO, «Sobre a Teoria Elementar do Ensaio de Hipóteses», *Revista de Economia*, vol. III, fasc. III.

[7] B. L. WELCH, «The Generalization of 'Student's' Problem when Several Different Population Variances Are Involved», *Biometrika*, vol XXXIV, Jan. 1947, págs. 28-35. Ver também vol. XLI, Dec. 1954.

[8] ALICE A. ASPIN, «Tables for Use in Comparisons whose Accuracy Involves Two Variances, Separately Estimated», *Biometrika*, vol. XXXVI, Dec. 1949, págs. 290-296.

[9] H. SCHEFFÉ, «On Solutions of the Behrens-Fisher Problem Based on the t-Distribution», *Ann. Math. Stat.*, vol. 14 (1943), págs. 35-44.

[10] G. DARMOIS, «Comparaison des Moyennes de Deux Populations Normales d'Ecarts-Types Inconnus et Différents», *Rev. de Statistique Appliquée*, vol. II, n.º 3, págs.37-41.