IX Congresso Anual da Sociedade Portuguesa de Estatística

Inferência Estatística

sobre a Localização usando a escala

Maria de Fátima Brilhante
Dinis Pestana
José Rocha
Maria Luísa Rocha
Sílvio Velosa

SPE
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Maria de Fátima Brilhante
Dinis Pestana
José Rocha †
Maria Luísa Rocha
Sílvio Velosa

INFERÊNCIA ESTATÍSTICA SOBRE A LOCALIZAÇÃO USANDO A ESCALA

INFERÊNCIA ESTATÍSTICA SOBRE A LOCALIZAÇÃO USANDO A ESCALA

Maria de Fátima Brilhante
Universidade dos Açores (DM) e
Centro de Estatística e Aplicações da Universidade de Lisboa

Dinis Pestana
Universidade de Lisboa (FCUL, DEIO) e
Centro de Estatística e Aplicações da Universidade de Lisboa

José Rocha †
Universidade dos Açores (DM) e
Centro de Estatística e Aplicações da Universidade de Lisboa

Maria Luísa Rocha
Universidade dos Açores (DEG) e
Centro de Estudos de Economia Aplicada do Atlântico

Sílvio Velosa
Universidade da Madeira (CCCEE) e
Centro de Estatística e Aplicações da Universidade de Lisboa

2011

Ao Professor Bento Murteira, autor de
Estatística: Inferência e Decisão,
que foi uma inspiração para todos nós.

Esta segunda edição é também dedicada
ao Professor José Carlos Andrade Rocha,
com saudade.

FCT

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia no âmbito do projecto PEst-OE/MAT/UI0006/2011.

Prefácio

A presente obra inspira-se em *Inferência Estatística sobre Localização e Escala*, de Brilhante *et al.* (2001), actualizando o texto e ampliando-o, incorporando nomeadamente novos resultados sobre *ANOSp* (*Analysis of Spacings*) em populações exponenciais, observações sobre verosimilhança, suficiência e localização e escala.

No trabalho de revisão, prévio à muitas alterações a que procedemos, tropeçámos amiúde na leitura, concluindo assim que muitas das deduções tinham sido apresentadas omitindo passos que, não sendo difíceis, não são óbvios e são trabalhosos. Acrescentámos por isso detalhes que tornam a leitura mais amena e proveitosa, para além de naturalmente termos procurado melhorar o texto, expurgando-o tão exaustivamente quanto conseguimos de erros e gralhas. Mantivemos por outro lado alguma redundância propositada, que visa simplificar a leitura de partes do texto sem constantemente obrigar a consultar o índice remissivo e folhear o livro, uma vez que actualmente a edição é muito mais fácil (e, sobretudo, menos dispendiosa) do que a antiga composição tipográfica, pelo que a concisão da escrita matemática clássica deixou de ser um imperativo económico.

Os resultados interessantes que se obtêm no caso de populações parentes gaussianas, uniformes e exponenciais não são facilmente generalizáveis a outros modelos. Pareceu-nos interessante incluir algumas razões dessa limitação. Em particular, mostramos que os referidos modelos são os únicos que, sob as condições de regularidade de Koopman em famílias exponenciais, admitem uma condensação óptima da informação de uma amostra num par de estatísticas suficiente para o par de parâmetros (localização, escala). Juntámos também a caracterização da exponencial baseada na independência dos *spacings* à caracterização da gaussiana pela independência de média e variância empíricas.

A incorporação de resultados que não apareciam no anterior curso permitiu-nos também optar por um fio condutor das matérias abordadas que nos parece, actualmente, dar uma perspectiva mais coesa e interessante, e porventura mais estimulante para quem queira identificar problemas em aberto. Escolhemos também um título mais adequado. Não é portanto a reedição de um livro que envelheceu, em muitos aspectos é de facto um livro novo. Justifica-se plenamente, no entanto, a manutenção de José Rocha como co-autor.

Fátima Brilhante — Dinis Pestana
Maria Luísa Rocha — Sílvio Velosa

Ponta Delgada, Julho/Setembro de 2011

Prefácio da 1ª edição

Agradecemos à Sociedade Portuguesa de Estatística o convite para realizarmos este curso. Não nos tendo sido imposto um modelo, decidimos optar por discutir alguns aspectos menos conhecidos do que quotidianamente ocupa tantos cientistas: comparar efeitos médios de tratamentos.

A solução no caso de populações parentes gaussianas com a mesma variância é bem conhecida. Noutras circunstâncias, em geral adoptam-se abordagens não paramétricas. O que procuramos é alertar os nossos auditores para as soluções que existem no caso de populações parentes gaussianas mas com variâncias diferentes (e confessamos já que transformar os dados não é do nosso agrado). Fisher, Jeffreys, Welch e Satterthwaite resolveram a questão de forma satisfatória, mas a polémica em que Fisher se envolveu com Bartlett e Welch levou-o, com o mau feitio que tinha, a preferir considerar o assunto uma curiosidade matemática sem relevância em análise de dados a ter que dar crédito a Welch (que nunca cita!). Obter um t de Student com um número de graus de liberdade fraccionário obrigou, naturalmente, a publicar tabelas de quantis críticos, ou a proceder a aproximações cuja qualidade não foi devidamente avaliada.

Scheffé iluminou a questão, mostrando que no caso de heterocedasticidade não é possível encontrar uma estatística que seja uma função invariante para permutações dos argumentos que tenha distribuição exacta t de Student. Por outro lado, construíu um teste aleatorizado que tem todas as boas propriedades requeridas, nomeadamente ter distribuição exacta t de Student, permitindo o recurso às tabelas usuais.

São questões que, contrariamente ao ponto de vista expresso por Fisher, nos parece terem cada vez mais relevância em análise de dados. De facto, a necessidade de proceder a grandes estudos de meta-análise, coloca na ordem do dia analisar dados recolhidos com metodologias diferentes, naturalmente com diferentes graus de precisão.

Transformar linearmente os dados contém em si mesmo a ideia que ou os dados originais não são gaussianos, ou os transformamos em dados não gaussianos. A inferência sobre localização em dados não gaussianos é complicada — apresentamos resultados para populações simétricas (é consolador verificar que no caso de parente gaussiana chegamos à densidade da t de Student!), e para populações uniformes e exponenciais, recorrendo a transformadas integrais inversas. É um caminho que merece ser explorado noutras populações, a dificuldade está, naturamente, em encontrar estimadores de localização e de escala que se prestem a usar o teorema de Basu, e conseguir inverter a transformada integral! A "análise da escala" que é possível fazer de forma elegante em populações exponenciais parece difícil de realizar noutros casos.

De facto, a comparação de localizações, mesmo no caso de duas amostras, é problema que não está resolvido. As dificuldades matemáticas que se põem justificam claramente, por en-

quanto, a opção não paramétrica.

A nossa investigação tem sido apoiada pelas nossas Universidades, pelo Centro de Estatística e Aplicações da Universidade de Lisboa, pela FCT e CITMA, instituições a quem queremos expressar a nossa gratidão.

Fátima Brilhante Dinis Pestana
José Rocha Sílvio Velosa

Índice

Lista de Variáveis Investigadas:

Studentização

1. Studentização (uma amostra gaussiana de tamanho n):

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \frown t_n$$

(Secção 1.1.1, especialmente pp. 28–32)

ou o caso especial de amostra das diferenças $x_k - y_k$ entre coordenadas de observações gaussianas emparelhadas de tamanho n:

$$t = \frac{\overline{D} - \delta}{\frac{S_D}{\sqrt{n}}} \frown t_n$$

($\delta = \mu_1 - \mu_2$)
(p. 36.)

2. Duas amostras gaussianas independentes, assumindo ho-

mocedasticidade:

$$T_{S_p} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frown t_{n_1+n_2-2}$$

onde

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(pp. 37–38.)

Mais geralmente, duas amostras gaussianas independentes, assumindo que $\sigma_1^2 = \theta\,\sigma_2^2$:

$$T_\theta^* = \frac{\overline{X}_2 - \overline{X}_1 - (\mu_2 - \mu_1)}{S_\theta \sqrt{\frac{1}{n_1} + \frac{1}{\theta n_2}}} \frown t_{n_1+n_2-2}$$

onde

$$S_\theta^2 = \frac{(n_1 - 1)S_1^2 + \theta(n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(pp. 95–103)

3. Duas amostras gaussianas independentes, sem assumir homocedasticidade

$$T_{S_1,S_2} = \frac{(\overline{X}_2 - \overline{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\,.$$

(p. 76.)

   (a) Behrens–Fisher

$$T_{S_1,S_2} = \zeta = t\sqrt{\frac{\left(1 + \frac{Mu}{\theta}\right)\left(1 + \frac{\theta}{N}\right)}{(1 + M)\left(1 + \frac{u}{N}\right)}},$$

onde $t \frown t_{n_1+n_2-2}$, $\theta = \frac{\sigma_1^2}{\sigma_2^2}$, $u = \frac{s_1^2}{s_2^2}$, $N = \frac{n_1}{n_2}$, $M = \frac{n_1-1}{n_2-1}$. — Tabelas de Fisher–Yates e tabelas de Sukhatme.
(pp. 82–83.)

(b) Welch — tabelas de Aspin e de Welch ou aproximação de Smith–Welch–Satterthwaite, $T_{\sigma_1,\sigma_2} \approx T_\nu \frown t_\nu$, com ν estimado por $\widetilde{\nu} = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$. (pp. 114–117.)

4. Estatística de Scheffé, duas amostras gaussianas independentes, sem assumir homocedasticidade, $n_1 \leq n_2$

$$T_S = \sqrt{n_1}\, \frac{\left(\overline{X}_1 - \overline{X}_2 - \delta\right)}{\sqrt{\dfrac{\sum_{k=1}^{n_1}\left(U_k - \overline{U}\right)^2}{n_1 - 1}}} \frown t_{n_1-1}.$$

onde $U_k = X_{1k} - \sqrt{\frac{n_1}{n_2}}\, X_{2k}$ e $\overline{U} = \frac{1}{n_1}\sum_{k=1}^{n_1} U_k$. (pp. 132–134.)

5. T^* auto-normalizada de Logan, Mallows, Rice e Shepp:

$$T^* = \frac{\sum_{i=1}^{n} X_i}{\left(\sum_{i=1}^{n} |X_i|^\alpha\right)^{\frac{1}{\alpha}}}$$

(p. 150.)

6. Studentização externa, interna e mixta. (p. 155.)

7. Studentização do mínimo pela amplitude (uma amostra uniforme):

$$T^*_{n-1} = \frac{X_{1:n} - \theta_1}{X_{n:n} - X_{1:n}}$$

com densidade

$$f_{T^*_{n-1}}(t) = \frac{n-1}{(1+t)^n}\, \mathrm{I}_{(0,\infty)}(t) \frown Pareto(n-1,0)$$

(pp. 172–174.)

8. Studentização obtida como função da padronizaçcão do mínimo (uma amostra uniforme):

$$\tau_{n-1} = (n+1)\sqrt{\frac{n+2}{n}}\, \frac{X_{1:n} - \theta_1}{X_{n:n} - X_{1:n}}$$

com densidade

$$f_{\tau_{n-1}}(t) = \frac{n-1}{n+1}\, \sqrt{\frac{n}{n+2}}\, \frac{1}{\left(1 + \sqrt{\frac{n}{n+2}}\, \frac{t}{n+1}\right)^n}\, \mathrm{I}_{(0,\infty)}(t),$$

(pp. 176–177.)

9. Studentização do estimador do valor médio (uma amostra uniforme):

$$\tau^*_{n-1} = \frac{\hat{\mu} - \mu}{\hat{\theta}}$$

com densidade

$$f_{\tau^*_{n-1}}(t) = \frac{n-1}{(1+2|t|)^n},\ t \in \mathbb{R}$$

(p. 178.)

10. Studentização do mínimo pelo estimador da escala (uma amostra exponencial):

$$W = \frac{X_{1:n} - \lambda}{\overline{X} - X_{1:n}} \frown Pareto(n-1, 0)$$

(p. 179–180.)

11. Studentização do mínimo pela amplitude (puma amostra exponencial):

$$\tau_{n-1} = \frac{X_{1:n} - \lambda}{X_{n:n} - X_{1:n}},$$

com densidade

$$f_{\tau_{n-1}}(t) = n\, B(n, nt) \sum_{k=1}^{n-1} \frac{nt}{(n-k)+nt},\ t > 0.$$

Tabela de quantis pp. 185–186. (pp. 180–184.)

12. Studentização usando a independência dos espaçamentos (uma amostra exponencial):

$$\tau^*_{n-1;i,k} = \frac{\overline{X}_n - \lambda}{X_{k:n} - X_{i:n}}$$

Densidade em termos de transformadas de Laplace inversas; tabela de quantis p. 194 (pp. 188–193.)

13. Studentização da diferença dos mínimos (duas amostra independentes exponenciais, homocedásticas):

$$W = \frac{Y_{1:n}^{(1)} - Y_{1:n}^{(2)} - (\lambda_1 - \lambda_2)}{\widehat{\delta}} = \frac{\dfrac{Y_{1:n}^{(1)} - \lambda_1 - (Y_{1:n}^{(2)} - \lambda_2)}{\delta}}{\dfrac{\widehat{\delta}}{\delta}},$$

com densidade

$$f_W(x) = \frac{N-2}{4\left(1+\frac{|x|}{2}\right)^{N-1}}\,\mathbb{I}_{\mathbb{R}}(x)$$

(pp. 196–198.)

14. Studentização (uma amostra de população parente simétrica, condições de regularidade):

$$T_{n-1} = \sqrt{n}\,\frac{\overline{X}_n}{\sqrt{\frac{SS_n}{n-1}}}$$

$$f_{T_{n-1}}(t) = \frac{2^{n-1}}{\sqrt{n-1}}\left[\prod_{i=3}^{n}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right]\times\int_0^{+\infty} u^{n-1}\prod_{i=1}^{n} f\!\left(\left(\frac{t}{\sqrt{n(n-1)}}+\alpha_{i,n}\right)u\right)\mathrm{d}u$$

(pp. 218–219.) No caso $n = 2$ tem-se

$$f_{T_1}(t) = 2\int_0^{+\infty} u\, f\!\left(\frac{u\,(t+1)}{\sqrt{2}}\right) f\!\left(\frac{u\,(t-1)}{\sqrt{2}}\right)\mathrm{d}u$$

(pp. 229–230, e exemplos diversos no Apêndice A.)

Análise da escala

1. ANOVA simples, ver Tabela 1.

Tabela 1: Tabela da ANOVA (simples).

Fontes de Variação	Graus de Liberdade	Quadrados Médios	razão F	p
$BSS = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{\overline{x}}\right)^2$	$k-1$	$\frac{BSS}{k-1}$	$F = \frac{\frac{BSS}{k-1}}{\frac{WSS}{N-k}}$	$\mathbb{P}\left(F_{k-1,N-k} \geq F\right)$
$WSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i\right)^2$	$N-k$ (*)	$\frac{WSS}{N-k}$		
$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{\overline{x}}\right)^2$	$N-1$			

(*) $\sum_{i=1}^{k} (n_i - 1) = N - k.$

(pp. 43–49.)

2. ANOVA simples não-paramétrica de Kruskal-Wallis:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{\nu} \frac{R_k^2}{n_k} - 3(N+1) \approx \chi^2_{\nu-1}$$

(pp. 120–122.)

3. ANOSp simples em populações exponenciais, ver Tabela 2.

(pp. 199–204.)

Tabela 2: Tabela da *ANOSp* (simples).

Soma de *Spacings*	Graus de liberdade	Estimador de δ	Valor de F
$BSSp = n\sum_{i=1}^{k}\left(X_{1:n}^{(i)} - X_{1:N}\right)$	$2(k-1)$	$\hat{\delta}_B^* = \frac{BSSp}{k-1}$	$F = \frac{\hat{\delta}_B^*}{\hat{\delta}_W^*}$
$WSSp = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - X_{1:n}^{(i)}\right)$	$2(N-k)$	$\hat{\delta}_W^* = \frac{WSSp}{N-k}$	

Introdução

Apesar de a Ciência estar a ganhar um estatuto mediático, o tempo que um resultado de facto inovador demora a migrar das revistas da especialidade para os manuais universitários é longo. Esta constatação parece incontroversa em compêndios de Matemática escritos para matemáticos, ou pelo carácter menos mediático desta ciência, ou pelo seu cariz dedutivo e profundamente hierárquico, que dificulta a exposição satisfatória de teorias que assentam em resultados especializados e difíceis. Fisher (1925, 1990), no último prefácio que escreveu, para a 13ª edição do seu célebre *Statistical Methods for Research Workers*, refere o desconforto causado por este livro — decerto um dos mais influentes na evolução da metodologia da investigação científica —, por não ser apenas um texto matemático demonstrativo, e pôr mais ênfase em resultados necessários à condução do planeamento e análise de experiências.

Por outro lado, a avaliação do que importa ensinar é caprichosa, guiada em geral por considerações pragmáticas — mas que nuns se traduzem por uma escolha do que pode ser exposto de forma coerente e rigorosa sem esforço excessivo, noutros por uma selecção do que consideram importante para aplicações, e

nem sempre pode ser plenamente justificado discursivamente.

Foi com alguma surpresa que constatámos que o problema de Behrens-Fisher, que gozou de alguma notoriedade por ter proporcionado os primeiros grandes êxitos à escola bayesiana (Jeffreys, 1940), brilha pela ausência nos manuais de Estatística para matemáticos e estatísticos. O problema de Behrens-Fisher (Behrens, 1929, que não consultámos mas que Welch considera cheio de erros; Fisher, 1935a, 1939b, 1941) é o problema da comparação das médias de duas populações gaussianas com variâncias diferentes[(1)] — dados medidos com diferentes precisões, como Fisher tão "experimentalmente" escreve — usando duas amostras independentes $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1n_1})$ e $\boldsymbol{X}_2 = (X_{21}, \ldots, X_{2n_2})$ de $X_1 \frown Gaussiana(\mu_1, \sigma_1)$ e de $X_2 \frown Gaussiana(\mu_2, \sigma_2)$, respectivamente.

Barnett (1999), por exemplo, despacha-o em meia dúzia de linhas, e com a excepção do texto um pouco mais ambicioso de Casella and Berger (2002), o típico é omitir o resultado ou, como Dixon and Masey (1969, p. 119), receitar, sem qualquer justificação ou referência, que o quociente $\dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

(1) Em rigor, quando o quociente ${\sigma_1^2}/{\sigma_2^2}$ não é conhecido. Como o caso mais usual, quando aquela razão é conhecida, é ${\sigma_1^2}/{\sigma_2^2} = 1$, é habitual fazer a dicotomia $\sigma_1^2 = \sigma_2^2$ *vs.* $\sigma_1^2 \neq \sigma_2^2$, quando em rigor se deveria fazer a dicotomia ${\sigma_1^2}/{\sigma_2^2}$ conhecido *vs.* desconhecido. Pearson (1905), a partir da palavra *skew* (assimétrico — a assimetria tem relevância na explicitação dos conceitos de elasticidade e de escala), inventou os neologismos *homocedasticidade* e *heterocedasticidade*, que actualmente se usam para expressar que as populações em estudo têm variâncias iguais, ou que nem todas têm a mesma variância, respectivamente.

tem distribuição que pode ser aproximada por uma t de Student cujo número de graus de liberdade deve ser avaliado por uma expressão que é função das variâncias das duas populações, com resultado em geral fraccionário. Desconsoladoramente, esse número de graus de liberdade, cuja expressão no caso de nada se assumir sobre o quociente das variâncias é

$$\nu = \frac{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)^2}{\dfrac{\sigma_1^4}{n_1^2(n_1-1)} + \dfrac{\sigma_2^4}{n_2^2(n_2-1)}},$$

no caso em que $\sigma_1 = \sigma_2$ não toma o valor n_1+n_2-2 que se obtém na abordagem tradicional, excepto quando também $n_1 = n_2$[(2)].

Em livros mais especializados de análise de variância e/ou planeamento de experiências a questão é em geral abordada, mas de forma muito superficial; é de crer que a maior parte dos leitores nem se aperceba de que o problema tem soluções sofisticadas, e que os seus principais construtores mereceriam mais crédito do que o que habitualmente recebem[(3)]. Behrens terá porventura

(2) Do mesmo modo, se começarmos por admitir igualdade dos erros padrões das duas amostras, isto é que $\frac{\sigma_1^2}{n_1} = \frac{\sigma_2^2}{n_2}$, $\nu = \frac{4(n_1-1)(n_2-1)}{n_1+n_2-2}$ toma o valor n_1+n_2-2 se e só se $n_1 = n_2$, e consequentemente também $\sigma_1 = \sigma_2$. Welch (1938) observa que, mais geralmente,

$$\frac{\frac{\sigma_1^2}{n_1}}{\frac{\sigma_2^2}{n_2}} = \frac{n_1-1}{n_2-1} \implies \nu = \frac{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)^2}{\dfrac{\sigma_1^4}{n_1^2(n_1-1)} + \dfrac{\sigma_2^4}{n_2^2(n_2-1)}} = n_1+n_2-2.$$

(3) Porventura é mesmo o facto de o problema ter diversas soluções, não

apenas o mérito de ter, em traços largos, dado a primeira solução ao problema, mas Fisher (1935a, 1939b, 1941, e nas sucessivas reedições de *Statistical Methods for Research Workers* e de *Statistical Methods and Scientific Inference*) e Jeffreys (1940, 1983) produziram trabalho substancial sobre a questão, o primeiro usando

---

coincidentes, que tem dificultado a sua divulgação. Os trabalhos de Welch (1938, 1939, 1947a, 1949) alinham pelas críticas que Bartlett (1936, 1937) dirigiu a Fisher, indo mesmo mais longe do que aquele, parecendo sugerir que a inferência fiducial deveria ser rejeitada em bloco, e dirige críticas semelhantes à abordagem bayesiana de Jeffreys. Mas comparado com as de Neyman (1956), as críticas de Welch são veludo...

Welch mostra além disso que a expansão assintótica que obtém é diferente da de Fisher (1941). Apesar de termos procurado repetidamente seguir a argumentação de Fisher sobre a inferência fiducial, nomeadamente em Fisher (1956b), não é fácil perceber a subtileza que distingue a escola fiducial da escola bayesiana (Pestana, 1981). Lindley (1958), no entanto, mostra que o argumento fiducial só equivale ao bayesiano se o parâmetro sobre o qual se está a fazer inferência for um parâmetro de localização, ou puder ser transformado num parâmetro de localização.

O problema de Behrens-Fisher parece central para elucidar alguns pontos das controversas questões que separam frequencistas, bayesianos e fiducialistas, cf. Welch (1939), e as abordagens divergentes de Fisher e de Neyman e Pearson à teoria dos testes de significância e dos testes de hipóteses.

Como Kendall and Stuart (1961) comentam, enquanto não se pôs o problema da comparação das médias em situação de heterocedasticidade, as diversas escolas chegavam aos mesmos resultados, e apenas parecia que estavam a cobrir o mesmo corpo de conhecimento com diversas roupagens. O problema de Behrens-Fisher veio mostrar a fissura irremediável entre as diversas abordagens à inferência.

Um comentário de Kendall and Stuart (1961) merece reflexão: como as soluções coincidem no caso de uma amostra, ou de duas amostras provenientes de populações gaussianas com iguais variâncias, e divergem no caso de duas amostras e variâncias diferentes, o esforço tem sido posto na tentativa de explicação desta diferença — quando, na opinião deles, esta diferença é que representa a situação geral, e o esforço dos investigadores deveria procurar explicar porque é que as soluções coincidem nos dois casos referidos.

a teoria da inferência fiducial que vinha desenvolvendo, o segundo uma abordagem bayesiana. Bartlett (1936, 1937) criticou a abordagem de Fisher no que ela tem de essencial, e Welch (1938, 1939, 1947a, 1951) solucionou o problema do ponto de vista frequencista. Satterthwaite (1946), cuja abordagem é semelhante à do trabalho de Welch, aparece em certo sentido como ganhador, pois os poucos livros que referem a solução em geral citam o seu nome. Tal deve-se porventura à abordagem mais pragmática que toma: estima a variância através de uma combinação de variáveis qui-quadrado, que aproxima por uma transformação linear de uma variável qui-quadrado, e o que lhe interessa é estimar o número de graus de liberdade de tal forma que a aproximação tenha valor médio e variância iguais aos da combinação linear que pretende aproximar (ao usar valores médios e variâncias procura evitar soluções negativas, possíveis usando apenas igualdade de valores médios); de facto, é uma abordagem de Cochran (1951) que dá a esse problema um enquadramento adequado, veja-se também McBean *et al.*, 1988). Casella and Berger (2002) apresentam o "estimador de Satterthwaite" como uma utilização sofisticada do método dos momentos.

Não tendo qualquer pretensão de exaustividade, procurámos citar na bibliografia os livros que encontrámos em que o problema é abordado, com a indicação das páginas (ou página!) em que é tratado. Cobb (1998) e Oehlert (2000) são os que registam mais informação; Cobb não arrola na bibliografia os trabalhos que indica, mas ocupa-se da modificação proposta por Cochran (1951) à abordagem de Satterthwaitte (1946), e Oehlert curiosamente dá crédito a Welch na p. 132 (sem o referir depois na bibliografia) e a Satterthwaitte nas pp. 262 e 274.

Enquanto se espera que qualquer estudante de Estatística

saiba testar a hipótese de igualdade de médias, com base na informação de duas amostras independentes extraídas de populações gaussianas com a mesma variância, poucos ficam a saber resolver o problema quando as variâncias são diferentes.[4]

O problema é exposto, com discussão detalhada das soluções, em Kendall and Stuart (1961), mas dificilmente se pode considerar esta obra um manual escolar. Pestana e Velosa (2010) tratam de forma elementar quer a situação $\sigma_1^2 = \theta\sigma_2^2$, com θ conhecido, quer a situação de nada se saber sobre o quociente $\frac{\sigma_1^2}{\sigma_2^2}$.

Em contrapartida, alguns livros de Bioestatística dão algum relevo à questão, porque é importante na prática testar efeitos médios de tratamentos em situações de heterocedasticidade. Zar (1996) apenas receita a solução (mas tem indicações preciosas sobre bibliografia adequada), e Samuels and Witmer (1999) usam consistentemente os resultados que em nota de rodapé atribuem a Welch e Satterthwaite, sem especifição de qualquer fonte bibliográfica.

É nosso objectivo apresentar uma panorâmica da inferência sobre localizações usando a escala como muleta, seja eliminando-a ou seja estudando o que pode revelar sobre a localização.

[4] *Nota da 2ª edição:* O teste de Welch-Satterthwaite tornou-se entretanto o teste que o *package* estatístico R executa, por defeito. Para executar o teste t "tradicional" para comparação de médias de duas gaussianas usando amostras independentes, assumindo homocedasticidade, é necessário expressamente invocar essa opção, declarando que $\sigma_X^2 = \sigma_Y^2$ tem o valor lógico $TRUE$. O *package* PASW/SPSS inclui os resultados de ambos os testes no quadro da comparação de amostras independentes e no quadro da análise de variância simples, e inclui também o teste de Brown-Forsythe, semelhante ao teste aproximado de Welch. O notável teste de Scheffé (1943) não teve ainda a divulgação que merece, e que se obtém privilegiadamente por implementação adequada.

Começamos por populações gaussianas, tratando com algum detalhe o caso menos conhecido de heterocedasticidade. Seguidamente, apresentamos resultados sobre populações não gaussianas, nas situações que proporcionam resultados interessantes:

1. populações exponenciais e uniformes, porque as famílias de localização e escala gaussiana, exponencial e uniforme são as únicas para as quais existe um par de estatísticas suficiente para a estimação do par de parâmetros (localização, escala), no contexto de famílias exponenciais de Darmois–Koopman–Pitman.

2. populações simétricas, porque um trabalho notável de Efron (1969) chama a atenção para as implicações da simetria da população parente na studentização;

E se por um lado olhamos para estes resultados com a simpatia ternurenta de quem os viu nascer, não podemos, por outro lado, deixar de apreciar a opção não paramétrica. Como dizia Saki, "*a little inaccuracy saves tons of explanations*". A outra abordagem possível — transformar os dados para os "normalizar" — não nos cativa tanto. Partilhamos do ponto de vista que um estatístico não pode olhar apenas para um modelo, tem que reflectir sobre os resíduos, e é decerto questionável, nesta perspectiva, usar transformações não lineares.

A obra seminal de Student (1908) — eliminar um parâmetro perturbador de escala quando se pretende inferir sobre a localização, técnica que genericamente designaremos "studentização" — deixou os seus contemporâneos apáticos (é pelo menos a opinião expressa por Fisher (1939a) no obituário que escreveu sobre Gosset) e o papel de Fisher para o reconhecimento da sua importância foi marcante.

Fisher (1925), ao aproveitar a variância para fazer inferência sobre a média — por a variância ser o menor momento de segunda ordem —, em lugar de lhe dar o estatuto de parâmetro perturbador e o eliminar, como Student fizera, inventou uma outra área seminal da Estatística, a análise da variância. O Capítulo 1 expõe brevemente os contributos iniciais de Student e de Fisher no que se refere a inferência sobre valores médios em populações gaussianas, na perspectiva pertinente para o seguimento da nossa exposição.

Por isso, tentamos sublinhar as ideias que os motivaram, e expomos com detalhe pouco usual uma abordagem técnica de testes de razões de verosimilhanças (por razões históricas, a controvérsia à volta do argumento fiducial e do problema de Behrens-Fisher é, nas referências que consultámos, perspectivada sempre em termos de estimação). De facto, essa análise será prosseguida no Capítulo 2 porque dela ressalta de forma inequívoca que não há solução que respeite os moldes clássicos quando o quociente entre as variâncias é desconhecido, como Scheffé (1944), por razões diversas da apresentada nesse ponto da exposição, também apontou (discutindo o resultado de Scheffé ulteriormente, no Capítulo 3).

No Capítulo 2 começamos por expor os resultados de Fisher (1935a, 1939b, 1941), e de Jeffreys (1940), recuperando um anterior trabalho de Behrens (1929), bastante imperfeito, ao que parece. Os argumentos quer de um quer de outro envolvem admitir distribuições para σ_i^2, distribuições fiduciais no caso de Fisher, distribuições *a priori* no caso de Jeffreys. O argumento fiducial de Fisher (que este retoma em *Statistical Methods and Scientific Inference*, 1956, aceitando tê-lo exposto deficientemente em 1935) originou uma polémica com Neyman e Bartlett que ainda

ecoa no livro de 1956, e em que Welch se perfila do lado de Bartlett. Nas reedições posteriores a 1938 de *Statistical Methods for Research Workers*, Fisher refere os seus trabalhos, e os de Sukhatme (1938), mas dando sempre a entender que é um trabalho de índole matemática, cujo interesse efectivo é restrito. Curiosamente, parece dar muito mais apreço ao problema, e considera-o potencialmente importante em aplicações, em *Statistical Methods and Scientific Inference*.

Expomos seguidamente, na continuação desse capítulo, a solução frequencista do "problema dos dois valores médios", por uns atribuída a Welch (1938, 1939, 1947a, 1951) e por outros a Satterthwaite (1941, 1946), que por sua vez, tal como Cochran (1951), atribui a ideia original a Smith (1936)[(5)]. De facto, Behrens–Fisher–Jeffreys e Smith–Welch–Satterthwaite ocupam-se da mesma "studentização", $T_{S_1,S_2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$, usando no entanto aproximações e métodos bem diversos, que conduzem a soluções muito distintas.

A diferença entre as abordagens de Welch e de Satterthwaite não é grande; poder-se-ia dizer que Welch, nos seus trabalhos iniciais se situa mais no campo da studentização[(6)], Satterthwaite no da análise da variância, mas ao trabalhar com duas amostras tem-se $F_{1,\nu} \equiv t_\nu^2$ e portanto as diferenças são mínimas. Welch preocupa-se mais em expor com rigor um problema matemático,

(5) Num artigo raramente citado de Satterthwaite (1941), que se pode encontrar em `http://www.springerlink.com/content/6h77h7288554h864/fulltext.pdf`, este arrola na bibliografia Welch (1938). Anote-se que o próprio Welch (1938, p. 352) indica Welch (1936) como primeira publicação da aproximação. Assim, aparentemente Smith (1936) e Welch (1936) partilham a prioridade da ideia.

(6) Mais tarde, em 1951 e em 1956, aborda também a análise da variância.

enquanto Satterthwaite, mais pragmático, dirige todo o seu esforço para estimar o número de graus de liberdade a usar num teste aproximado.

No Capítulo 3 discutimos outros desenvolvimentos do problema de Behrens-Fisher, nomeadamente aquela que consideramos ser a melhor solução[(7)] do problema, proposta por Scheffé (1943, 1944).

O Capítulo 4, breve mas importante, trata da caracterização da gaussiana pela independência de $\overline{X}$ e S^2, sugestiva de que em populações não gaussianas não é decerto compensador procurar copiar simplesmente o que se faz no caso gaussiano.

Na segunda Parte desta obra abordamos o problema da studentização em populações não gaussianas, remetendo para Hotelling (1961) e Efron (1969) para uma discussão aprofundada sobre a necessidade de obter resultados exactos neste campo, em lugar de usar acriticamente resultados aproximados, invocando o Teorema Limite Central.

Os resultados interessantes dizem respeito a uniformes e a exponenciais, pois como mostramos no Capítulo 5 estes são os únicos modelos da família exponencial de Darmois–Koopman–Pitman em que existe um par de estatísticas suficiente para o par de parâmetros (localização, escala). Também no Capítulo 5 abordamos, com a generalidade inevitável quando se foge ao modelo gaussiano, localização e escala, studentização, nomeadamente na perspectiva geral que primeiro encontrámos em David (1981), discutindo as noções de studentização interna e de

(7) No sentido restrito de não invocar tabelas especiais, ou não recorrer a uma aproximação, truncando um número de graus de liberdade fraccionário. Note-se, por outro lado, que o próprio Scheffé (1970) apodou a sua solução "an impractical solution".

studentização externa. Apresentamos ainda a caracterização da exponencial pela independência dos *spacings*, que mostra que a metodologia seguida para tratar com êxito o problema da localização em exponenciais não pode ser generalizada de forma óbvia para outras populações parentes.

Assim, nos Capítulos 6 e 7 apresentamos resultados para populações uniforme e exponencial, respectivamente, expondo os trabalhos de Brilhante, Pestana e Rocha. A questão é abordada, na perspectiva alargada de studentização interna e studentização externa, na acepção de David (1981), cf. a Definição 5.1.1. O eixo de desenvolvimento que parece mais promissor é: encontrar um par de estimadores, do parâmetro de localização sobre o qual pretendemos fazer inferência, e dos parâmetros perturbadores de que aquele depende, e que se procura eliminar; usar transformadas integrais e as suas inversas para conseguir explicitar a função densidade de probabilidade de uma variável studentizada, isto é, em que os parâmetros perturbadores foram eliminados. A ideia de usar transformadas integrais foi também explorada com sucesso por Margolin (1977). No caso de parente exponencial, apresentamos também um simile da ANOVA simples de Fisher, usando espaçamentos (que por isso siglamos ANOSp), usando o facto de o estimador natural da localização, $X_{1:n}$, ser independente do estimador $X_{n:n} - X_{1:n}$ de dispersão. Esperamos que abra o caminho a mais investigação neste campo.

O Capítulo 8 expõe uma aproximação da studentização (no sentido restrito clássico) em populações simétricas, com resultados de Pestana e Rocha, que têm aspectos positivos e negativos. Têm o aspecto fascinante de o cálculo usar os mesmo pontos, qualquer que seja a população de base (lembrando a integração de Gauss, que também usa sempre os zeros de polinómios or-

togonais, qualquer que seja a função que se pretenda integrar). Em apêndice calculamos a expressão explícita de funções densidade de probabilidade de variáveis T_1 em diversas populações, e a expressão de Perlo (1933) para T_2 no caso de parente uniforme, uma curiosidade que pretende apenas mostar com algum detalhe como T_1 em cada um dos casos que abordamos é bem diversa de $T_1 \frown Cauchy(1)$ do caso gaussiano.

Relegamos também para apêndice uma questão importante, mas pelas quais sentimos um apreço limitado: Uma forma de resolver a questão da heterogeneidade das variâncias é transformar os dados, por exemplo usando transformações de Box-Cox e estatísticas ponderadas, nomeadamente mínimos quadrados ponderados. Para além de uma discussão e remissão para literatura da especialidade no que se refere a transformações destinadas a homogeneizar as variâncias, expomos um tratamento com mínimos quadrados ponderados, baseando-nos em Box and Hill (1974), que consegue, de facto, em exemplos concretos de modelação em Química, "positivar" estimativas que com análise tradicional têm resultado negativo, o que é uma aberração do ponto de vista químico. Isto, no entanto, não altera a nossa desconfiança relativamente ao hábito, excessivamente liberalizado, de transformar dados, esquecendo que a transformação do sinal é acompanhada de transformação de ruído, e que não temos um verdadeiro controlo sobre este lado da questão, excepção feita das transformações lineares (que não alteram a forma, e não são por isso as usualmente escolhidas em análise de dados).

Parte I

Inferência sobre Valores Médios de Populações Gaussianas

Capítulo 1

Studentização e ANOVA

Como usar o estimador da variância para inferir sobre o valor médio de uma população gaussiana. Como usar estimadores da variância para comparar valores médios de populações gaussianas, assumindo que o quociente das suas variâncias é conhecido

> "*Its originator, who published anonymously under the pseudonym 'Student', possesses the remarkable distinction that, without being a professed mathematician, but a research chemist, he made early in life this revolutionary refinement of the classical theory of errors.*"
>
> R. A. Fisher, *The Design of Experiments*, p. 34.

O nosso objectivo neste capítulo é apresentar as ideias de base da studentização e da análise da variância. O objectivo comum destes dois notáveis desenvolvimentos da Estatística, inventados

no primeiro quartel do século XX, é levar a cabo inferência sobre valores médios de populações gaussianas. Mas enquanto na studentização se divide uma variável aleatória, que depende do parâmetro de localização e também do parâmetro de escala, por uma função do estimador da variância para obter uma variável fulcral adequada que não dependa deste "parâmetro perturbador", na análise da variância usa-se a informação que os estimadores da variância nos podem proporcionar sobre a localização.

A studentização foi inventada por Student (1908) para eliminar um parâmetro perturbador de escala quando o que se pretendia era estimar ou testar o valor do parâmetro de localização, com base numa amostra. Muito afortunadamente, a solução do problema da comparação de duas médias — no caso de igualdade das variâncias — é um corolário quase imediato do que sabemos sobre somas de variáveis gamas independentes, com a mesma escala, e desde a sua publicação a estatística de Student (1908) foi usada para comparar efeitos médios de dois "tratamentos".

Foi mesmo empregue para comparar, aos pares, k tratamentos, até Fisher (1925) chamar a atenção para o facto de que esta comparação aos pares levava a que o nível real e o nível nominal dos testes (ou coeficientes de confiança) eram tão coincidentes quanto a taxa nominal e a taxa real de juro quando pedimos um empréstimo ao banco ... A criação da análise da variância, em que o parâmetro de escala, em vez de ser eliminado é usado para fazer inferência sobre a localização, é um dos pontos altos da História da Estatística, que muito justamente contribuíu para tornar o livro de Fisher o de maior sucesso nesta área[(1)].

[(1)] Fisher ainda escreveu o prefácio para a 13ª edição, e preparou as remodelações para a 14ª edição antes de morrer; dessas, algumas tiveram diversas impressões, e a Oxford University Press publicou em 1990 uma

A exposição inicial é feita com base em conhecimentos elementares de Probabilidade. Seguidamente, usamos a teoria de Neyman-Pearson para dar um enquadramento à luz de testes de razões de verosimilhanças (o teorema sobre o teste de igualdade de variâncias é registado porque a questão é discutida no Apêndice B). São apenas exercício de aquecimento, mas raramente estes resultados são registados com o detalhe usado em Pestana e Velosa (2010) e que aqui usamos.

A razão deste detalhe é abordar seguindo a mesma linha, nos Capítulos 4 e 5, o problema da inferência sobre dois ou sobre $k > 2$ valores médios em populações gaussianas, na situação de variâncias desconhecidas, sendo também desconhecido o quociente $\frac{\sigma_1^2}{\sigma_2^2}$. Veremos, assim, que as soluções de Welch (1938, 1939, 1951) e de Satterthwaite (1946) não podem ser deduzidas no âmbito da teoria dos testes de razões de verosimilhanças.

1.1 Student e o 'erro de uma média'

A grande aquisição da Teoria da Probabilidade clássica é o Teorema Limite Central, que mostra que em condições muito gerais o modelo gaussiano é admissível. A família de variáveis aleatórias gaussianas é uma família de localização e escala, e em muitas circunstâncias o que o utilizador da Estatística pretende estabelecer é algum tipo de diferença nos efeitos médios de diversos tratamentos. Assim, um dos problemas centrais da estatística

publicação conjunta dos livros de Fisher, de que já houve várias edições com diversas impressões.

foi desenvolver metodologias de inferência sobre o parâmetro de localização.

Na Estatística pré-Student o problema era abordado com alguma ligeireza. Tudo indica que a questão era tratada desta forma:

Uma vez que o modelo usado é $X \frown Gaussiana(\mu, \sigma)$, a média $\overline{X}$ de uma amostra aleatória de dimensão n é $Gaussiana(\mu, \frac{\sigma}{\sqrt{n}})$, uma consequência imediata da estabilidade da gaussiana no contexto de somas, cf. Pestana e Velosa (2010, p. 829). Sendo o objectivo fazer inferência sobre μ, como o sinal de μ na amostra é $\overline{x}$ e a gaussiana centrada em 0 é simétrica,

$$\frac{\mu - \overline{X}}{\frac{\sigma}{\sqrt{n}}} \frown Gaussiana(0,1),$$

e poder-se-ia usar um intervalo de confiança para μ da forma

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right),$$

com coeficiente ou grau de confiança $(1-\alpha) \times 100\%$. (Amostra aleatória e intervalos de confiança não seriam conceitos formalizados como hoje, mas o fundamental das ideias que lhes estão associadas, e a sua utilização pragmática, estão claramente presentes nos trabalhos de Helmert e de K. Pearson, por exemplo.)

Ser μ desconhecido e σ conhecido? Como é isso possível, se na definição de $\sigma^2 = \mathbb{E}[(X-\mu)^2]$ se usa μ?! Com certeza que temos que usar o sinal de σ na amostra, e o natural é usar

$$s = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}.$$

Que influência tem isto na estandardização da gaussiana? Resposta ingénua:

$$\frac{\mu - \overline{X}}{\frac{s}{\sqrt{n}}} = \frac{\mu - \overline{X}}{\frac{\sigma}{\sqrt{n}}} \frac{\sigma}{s} \frown Gaussiana\left(0, \frac{\sigma}{s}\right),$$

e como é de admitir (pelo menos para grandes amostras) que $\frac{\sigma}{s} \approx 1$, não custa muito continuar a aceitar que o intervalo de confiança $\left(\overline{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$ pode ser considerado uma boa aproximação.

1.1.1 Média "studentizada" e inferência sobre o valor médio de uma população gaussiana

O espectacular trabalho de Student (1908) chamou a atenção para que o facto de se julgar que se estava a dividir por uma "constante errada" era, em si mesmo, um erro. A estimativa s de σ deve ser encarada como o valor observado de uma variável aleatória, o estimador S, e quando se divide $\mu - \overline{X}$ por $\frac{S}{\sqrt{n}}$ não se obtém uma gaussiana com escala não unitária, o que se obtém é uma variável aleatória com distribuição em forma de sino, como a gaussiana, mas com caudas mais pesadas (no caso de a dimensão da amostra ser $n = 2$ obtém-se a Cauchy, de caudas tão pesadas que nem tem valor médio).

A forma actual de expor a questão, que poupa alguma etapas, é notar que

- $Z_0 = \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \frown Gaussiana(0, 1)$;

- $Y_{n-1} = \dfrac{(n-1)S^2}{\sigma^2} \frown \chi^2_{n-1}$, (Pestana e Velosa, 2010, p. 974–975);
- $\overline{X}$ e S^2 são variáveis aleatórias independentes, no caso de populações gaussianas, (Pestana e Velosa, 2010, p. 979–980).

Então

$$t_{n-1} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{Z_0}{\sqrt{\frac{Y_{n-1}}{n-1}}}$$

é o quociente entre duas variáveis aleatórias independentes, uma gaussiana padrão e a raiz quadrada de um qui-quadrado dividido pelo seu valor médio $n-1$, respectivamente, a que chamamos variável aleatória t de Student com $n-1$ graus de liberdade.

Deduzir a função densidade de probabilidade deste quociente é conceptualmente simples, e para quem domine razoavelmente o uso da função gama de Euler é quase imediato. A partir daí a tabulação de quantis (e, hoje, o cálculo imediato de valores de prova p associados a valores observados de t_{n-1}) é uma questão de meios de cálculo e/ou paciência.

Teorema 1.1.1.

A função densidade de probabilidade da variável aleatória t de Student com n graus de liberdade é

$$f_{t_n}(t) = \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \frac{1}{\sqrt{n}\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} \mathrm{I}_{\mathbb{R}}(t)\,.$$

Demonstração:

Tem-se

$$\begin{aligned} F_{t_n}(t) &= \int_0^{+\infty} \mathbb{P}\left(\frac{Z_0}{\sqrt{\frac{Y_n}{n}}} \leq t \mid Y_n = y\right) f_{Y_n}(y)\ \mathrm{d}y = \\ &= \int_0^{+\infty} \mathbb{P}\left(Z_0 \leq t\sqrt{\frac{y}{n}}\right) f_{Y_n}(y)\ \mathrm{d}y = \\ &= \int_0^{+\infty} \Phi\left(t\sqrt{\frac{y}{n}}\right) f_{Y_n}(y)\ \mathrm{d}y, \end{aligned}$$

onde Φ é a função de distribuição da $Gaussiana(0,1)$. Derivando em ordem a t vem

$$f_{t_n}(t) = \int_0^{+\infty} \phi\left(t\sqrt{\frac{y}{n}}\right)\sqrt{\frac{y}{n}}\, f_{Y_n}(y)\ \mathrm{d}y$$

e como $Y_n \frown \chi^2_n$, ou seja $Y_n \frown Gama\left(\frac{n}{2}, 2\right)$, tem-se

$$f_{Y_n}(y) = \frac{y^{\frac{n}{2}-1}\,\mathrm{e}^{-\frac{y}{2}}}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)}\ \mathrm{I}_{(0,\infty)}(y)$$

donde

$$f_{t_n}(t) = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-\frac{y\,t^2}{2n}} \sqrt{\frac{y}{n}} \, \frac{y^{\frac{n}{2}-1}\mathrm{e}^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \, \mathrm{d}y =$$

$$= \frac{1}{\sqrt{2\,\pi\,n}\;2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{+\infty} y^{\frac{n-1}{2}} \mathrm{e}^{-\frac{y}{2}\left(1+\frac{t^2}{n}\right)} \, \mathrm{d}y.$$

Fazendo $\frac{y}{2}\left(1+\frac{t^2}{n}\right) = z$, $y = \frac{2z}{1+\frac{t^2}{n}} \Rightarrow \mathrm{d}y = \frac{2\;\mathrm{d}z}{1+\frac{t^2}{n}}$, vem

$$f_{t_n}(t) = \frac{1}{\sqrt{2\,\pi\,n}\;2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{+\infty} \left(\frac{2z}{1+\frac{t^2}{n}}\right)^{\frac{n-1}{2}} \mathrm{e}^{-z} \frac{2\;\mathrm{d}z}{1+\frac{t^2}{n}} =$$

$$= \frac{1}{\sqrt{n\,\pi}\;\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \int_0^{+\infty} z^{\frac{n+1}{2}-1}\mathrm{e}^{-z}\mathrm{d}z,$$

ou seja, uma vez que $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$,

$$f_{t_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\;\Gamma\left(\frac{n}{2}\right)} \, \frac{1}{\sqrt{n}\,\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \; \mathrm{I}_{\mathbb{R}}(t)\,,$$

Finalmente

$$f_{t_n}(t) = \frac{1}{B\left(\frac{1}{2},\frac{n}{2}\right)} \, \frac{1}{\sqrt{n}\,\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \; \mathrm{I}_{\mathbb{R}}(t)\,, \qquad (1.1)$$

□

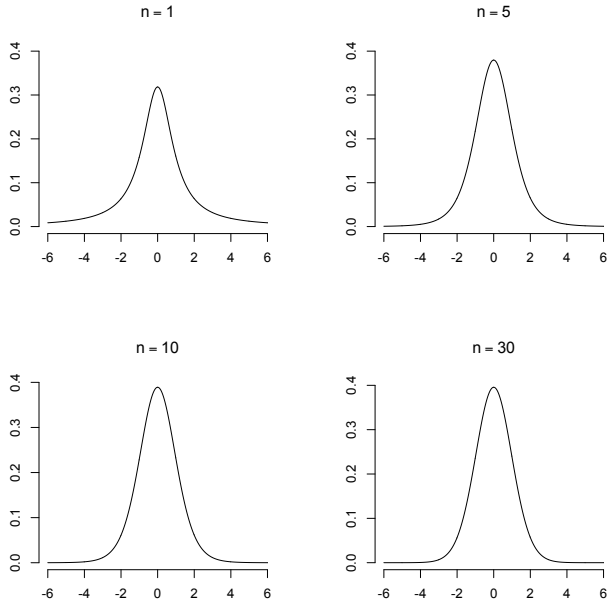

Figura 1.1: Densidades t de Student com n graus de liberdade.

Note-se que a função densidade de probabilidade de t_1 é $f_{t_1}(t) = \frac{\Gamma(1)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})} \frac{1}{\sqrt{1}\left(1+\frac{t^2}{1}\right)^{\frac{2}{2}}} \mathrm{I}_{\mathbb{R}}(t) = \frac{1}{\pi} \frac{1}{1+t^2} \mathrm{I}_{\mathbb{R}}(t)$, ou seja a t de Student com 1 grau de liberdade é a Cauchy padrão, como atrás anotáramos.

Façamos ainda alguns comentários, que apontam resultados intuitivos. Sabemos que as variáveis aleatórias gama têm

assimetria direita. Por isso, se $Z_0, Z_1, \ldots, Z_n$ forem $n+1$ variáveis aleatórias independentes e identicamente distribuídas, com $Z_k \frown Gaussiana(0,1)$, então o radicando no denominador de T_n é a média de n variáveis aleatórias $Z_k^2 \frown \chi_1^2$ e $\mathbb{P}\left[\sqrt{\frac{Z_1^2+\cdots+Z_n^2}{n}} \leq 1\right] = \mathbb{P}\left[\frac{Z_1^2+\cdots+Z_n^2}{n} \leq 1\right] > 0.5$, donde as caudas da t de Student são mais pesadas do que as caudas da gaussiana padrão. Assim, os quantis extremos da t de Student excedem os correspondentes quantis da gaussiana, uma questão que tem incidências importantes em inferência estatística.

Por outro lado, $\frac{Z_1^2+\cdots+Z_n^2}{n}$ tem valor médio 1 e desvio padrão $\sqrt{\frac{2}{n}} \xrightarrow[n\to\infty]{} 0$ e consequentemente é de esperar que com grande probabilidade se aproxime da variável aleatória degenerada em 1 quando $n \to \infty$. De facto, $\frac{Z_1^2+\cdots+Z_n^2}{n} \xrightarrow[n\to\infty]{\text{p}} 1$, quer como consequência da desigualdade de Chebycheff (cf. Pestana e Velosa, 2010, p. 430), quer como consequência imediata da Lei dos Grandes Números (cf. Pestana e Velosa, 2010, p. 1001), uma vez que aquela expressão é a média de variáveis aleatórias $Z_k^2 \frown \chi_1^2$, com valor médio 1. Então, quando o número de graus de liberdade aumenta

$$t_n \xrightarrow[n\to\infty]{\text{d}} Z \frown Gaussiana(0,1),$$

em virtude do teorema de Slutsky, cf. Pestana e Velosa (2010, p. 997–998).

A observação das tabelas da t mostra que, de facto, os quantis de probabilidade p da t convergem — decrescem, de acordo com a observação acima — para os correspondentes quantis da Gaussiana padrão. Habitualmente na última linha da tabela estão registados, para um número de graus de liberdade $n = \infty$, os

quantis de probabilidade p de $Z \frown Gaussiana(0,1)$.

Claro que o trabalho pioneiro de Student (1908) foi bem mais difícil de desenvolver do que parece vendo esta exposição feita à luz de uma compreensão mais actual do problema, como atestam as 25 páginas do seu fabuloso artigo publicado na *Biometrika*. O comentário de Welch (1947a) merece registo:

> "*It will be recalled that in Gosset's original approach to the single sample problem ('Student', 1908) his initial step was to note that the first four moments of the distribution of* s^2 *were consistent with the assumption that the distribution could be represented by a Pearson Type III curve. He was fortunate in this way to rediscover a distribution that had already been found by Helmert, as this permited him to go on to the derivation of the t-distribution.*"

Como é usual com os grandes passos em Ciência, esta notável contribuição para a teoria dos erros foi recebida com "*weighty apathy*", para usar os termos de Fisher (1939a) no obituário de Student. Em certo sentido, foi o trabalho de Fisher que chamou a atenção para a importância do trabalho pioneiro de Student, que é hoje reconhecido como a semente germinal de toda a estatística moderna.

Um olhar crítico para o trabalho de Student permite salientar os traços essenciais do seu contributo:

- Pretende usar-se a informação fornecida por uma amostra aleatória $\boldsymbol{X}$ em inferência envolvendo um parâmetro de localização μ, usando uma função $T^* = T^*(\boldsymbol{X}, \mu) = \overline{X} - \mu$;

- A função de distribuição F_{T^*} de T^* depende de um "parâmetro perturbador" de escala σ, de que $S = \Sigma(\boldsymbol{X})$ é um estimador;

- é possível deduzir a distribuição exacta de $T = \sqrt{n}\,\frac{T^*}{S}$, a qual não depende do parâmetro perturbador σ.

No rasto do trabalho de Student, quando este começou a receber o crédito que merecia, muitos procuraram desenvolver resultados análogos em populações não gaussianas. Mas a independência entre $\overline{X}$ e S^2 é característica de populações gaussianas, e os resultados sobre distribuições exactas, para populações não gaussianas, limitam-se às situações pouco relevantes de $n = 2$ ou, já com um aspecto analítico desencorajador, $n = 3$, veja-se o Apêndice A.

Claro que se $T^*(\boldsymbol{x})$ for calculada com base em $\boldsymbol{X}$ e a estimativa $S = \Sigma(\boldsymbol{y})$ for o valor observado de $S = \Sigma(\boldsymbol{Y})$, em que $\boldsymbol{Y}$ é independente de $\boldsymbol{X}$, contornamos com esta "studentização externa" o grande problema do cálculo da distribuição de $T = \frac{T^*}{S}$, que é a dependência entre numerador e denominador.

David (1981), indo directo ao âmago da questão, considera que studentização é qualquer processo de nos desembaraçarmos de um parâmetro perturbador (em geral um parâmetro de escala): Dividimos a versão centrada em 0 do estimador de um parâmetro (em geral de localização) sobre o qual queremos fazer inferência, mas cuja distribuição depende do referido parâmetro perturbador, por um estimador apropriado deste, por forma a obter uma variável fulcral (um *pivot*), cuja distribuição já não depende do parâmetro perturbador.

Se os estimadores do parâmetro de interesse e do parâmetro

perturbador forem independentes, chama-lhe studentização externa; se forem dependentes, chama-lhe studentização interna.

O interesse de David na studentização deve-se a, nesta perspectiva, já não estarem em jogo apenas valor médio e erro padrão. A amplitude studentizada, por exemplo, é um indicador importante em análise da variância, e outras funções de estatísticas ordinais, apropriadamente studentizadas, têm interesse em Estatística Aplicada.

Pestana e Rocha (1992) abordaram a questão com grande generalidade, deduzindo expressões aproximadas em populações simétricas, e estudando com inversão de transformadas integrais as populações beta e gama. Rocha (1995) apresenta com detalhe uma abordagem muito geral a populações simétricas, obtendo resultados que admitem comparação com a integração numérica de Gauss, no sentido em que os "pontos interpoladores" que usa na aplicação do teorema do valor médio generalizado são totalmente independentes da população parente. Brilhante *et al.* (1996) contém resultados diversos sobre studentização interna, enquanto Brilhante (1996a) foca a studentização externa em populações exponenciais, obtendo resultados exactos e assintóticos, quer à custa de funções especiais quer por inversão de transformadas de Laplace, fornecendo tabelas que tornam esses resultados efectivamente usáveis em inferência estatística. Pestana *et al.* (1999) e Brilhante *et al.* (2009) retomaram a questão no âmbito mais vasto da análise da escala, que adiante discutimos, a propósito da criação fundamental de Fisher.

1.1.2 Comparação dos valores médios de duas populações gaussianas

A comparação dos valores médios das margens de um par aleatório bigaussiano $(X,Y) \frown Bigaussiana(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$, partindo de uma amostra $\boldsymbol{(x,y)} = ((x_1,y_1),\ldots,(x_n,y_n))$ de observações emparelhadas é consequência imediata da teoria atrás desenvolvida para inferência sobre o valor médio de uma população gaussiana com base numa amostra — basta considerar a amostra $\boldsymbol{d} = (d_1,\ldots,d_n)$, onde $d_k = x_k - y_k,\ k = 1,\ldots,n$. Denotando $\delta = \mu_1 - \mu_2$, tem-se então, com as notações habituais,

$$\frac{\overline{D} - \delta}{\frac{S_D}{\sqrt{n}}} \frown t_{n-1}.$$

A abordagem de Student permite ainda estudar $\mu_1 - \mu_2$, com base na estimativa $\overline{x}_1 - \overline{x}_2$ calculada usando as observações de duas amostras independentes de populações gaussianas *com a mesma variância* ("*homocedasticidade*") de dimensões respectivamente n_1 e n_2.

De facto, esta hipótese apenas tem que ver com tratabilidade matemática, e é mesmo avessa à filosofia da Estatística. A soma de gamas independentes *com o mesmo parâmetro de escala* é ainda uma gama com a mesma escala, cujo índice é a soma dos índices das parcelas, e assim a soma de variáveis quis-quadrados independentes é ainda uma variável qui-quadrado, cujo número de graus de liberdade é a soma dos números de graus de liberdade das parcelas. Porém, se as escalas forem diferentes, o problema deixa de ser elementar.

Assim, admitindo homocedasticidade, o uso da variância ponderada

$$S_p^2 = \frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}$$

resolve airosamente a questão, pois

$$(n_1 + n_2 - 2)\,\frac{S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{\sigma^2} \frown \chi^2_{n_1+n_2-2}:$$

Teorema 1.1.2.
Sejam $\boldsymbol{X_1} = \left(X_{11}, \ldots, X_{1n_1}\right)$ *e* $\boldsymbol{X_2} = \left(X_{21}, \ldots, X_{2n_2}\right)$ *amostras aleatórias independentes, com* $X_{1k} \overset{\mathrm{d}}{=} X_1 \frown$ *Gaussiana* (μ_1, σ) *e* $X_{2j} \overset{\mathrm{d}}{=} X_2 \frown$ *Gaussiana* (μ_2, σ). *Então*

$$t_{S_p} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frown t_{n_1+n_2-2}.$$

Demonstração:

Comecemos por observar que

$$\left.\begin{array}{l} \overline{X}_1 \frown \textit{Gaussiana}\left(\mu_1, \frac{\sigma}{\sqrt{n_1}}\right) \\ \overline{X}_2 \frown \textit{Gaussiana}\left(\mu_2, \frac{\sigma}{\sqrt{n_2}}\right) \end{array}\right\} \implies$$

$$\implies \quad \overline{X}_1 - \overline{X}_2 \frown \textit{Gaussiana}\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \quad \implies$$

$$\Longrightarrow \quad \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frown \mathit{Gaussiana}(0,1).$$

Por outro lado, como atrás vimos

$$\frac{(n_1 - 1)\, S^2_{X_1} + (n_2 - 1)\, S^2_{X_2}}{\sigma^2} \frown \chi^2_{n_1+n_2-2}.$$

Como as populações são gaussianas, $\overline{X}_1$ e $S^2_{X_1}$ são independentes, e independentes de $\overline{X}_2$ e $S^2_{X_2}$, que por sua vez também são independentes, donde

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{e} \quad \frac{(n_1 - 1)\, S^2_{X_1} + (n_2 - 1)\, S^2_{X_2}}{\sigma^2}$$

são independentes.

Consequentemente

$$\frac{\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)\, S^2_{X_1} + (n_2-1)\, S^2_{X_2}}{(n_1+n_2-2)\,\sigma^2}}} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frown t_{n_1+n_2-2}.$$

□

(Note a semelhança de $\dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frown t_{(n_1-1)+(n_2-1)}$ com $\dfrac{\overline{X}_1 - \mu_1}{S\sqrt{\frac{1}{n_1}}} \frown t_{n_1-1}$ — no caso de uma amostra, obviamente $S_p = S$.)

Esta resolução decorre de assumir hipóteses bastante restritivas, porventura irrealistas em muitas situações. Diversas formas de relaxar essas hipóteses restritivas colocam problemas estimulantes e importantes. Nomeadamente, colocam-se as seguintes questões:

1. A extensão à comparação de $k > 2$ valores médios de populações gaussianas, assumindo homocedasticidade, cuja solução, que se deve a Fisher, foi a criação da análise da variância e a génese do planeamento de experiências. A análise da variância simples é apresentada na Secção 1.2.

2. A extensão deste resultado a gaussianas com variâncias diferentes. Este problema foi abordado por Behrens (1929), Fisher (1935b, 1939b), Jeffreys (1940), Smith (1936), Welch (1938), Satterthwaite (1946), Aspin (1948, 1949), Scheffé (1943, 1944) e Cochran (1951) entre outros, e o essencial dos resultados será apresentado nos Capítulos 3 e 4.

3. Dedução de resultados similares em populações não gaussianas. Nesse caso, a dependência entre estimadores de média e variância é uma dificuldade tal que as abordagens não paramétricas (em que apenas se admite que as populações parentes são contínuas, por forma a conceptualmente os *ranks* poderem ser definidos sem ambiguidade — ainda que na prática a discretização imposta pelos instrumentos de medida faça surgir dados empatados, e haja que atribuir *ranks* médios) se tornaram incontornáveis. Conseguem-se alguns resultados exactos, ou boas aproximações, para algumas populações não gaussianas cuja função densidade de probabilidade tem uma

expressão analítica particularmente simples, ou colocando hipóteses fortes como simetria, sobretudo num contexto mais geral de studentização, encarada como uma divisão de uma variável dependendo de um parâmetro perturbador por um estimador desse parâmetro. O assunto será tratado na Parte II.

1.2 Fisher e a comparação de k médias

Student (1908) resolveu o problema da inferência sobre o valor médio de uma população gaussiana, e subsidiariamente o problema da diferença entre os valores médios de duas populações gaussianas com a mesma variância, inventando uma forma de eliminar o parâmetro perturbador de escala.

Fisher (1925) teve a ideia de usar a influência da localização no parâmetro de escala para investigar a igualdade de valores médios de k populações gaussianas com a mesma variância.

Há duas questões elementares que queremos recordar antes de apresentar as ideias de Fisher:

1. A função $f(x) = x^2$ é a única função não trivial tal que

$$\frac{f(x+y) + f(x-y)}{2} = f(x) + f(y).$$

2. A variância é o menor dos momentos de segunda ordem.

De facto,

$$\mathbb{E}\left[(X-C)^2\right] = \mathbb{E}\left[(X-\mathbb{E}[X]+\mathbb{E}[X]-C)^2\right] = \\ = \operatorname{var}[X] + (\mathbb{E}[X]-C)^2 \geq \operatorname{var}[X],$$

com igualdade se e só se $C = \mathbb{E}[X]$.

A variância empírica goza de propriedade análoga:

$$\frac{1}{n-1}\sum_{k=1}^{n}(x_k-C)^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x}+\overline{x}-C)^2 = \\ = \frac{1}{n-1}\left\{\sum_{k=1}^{n}(x_k-\overline{x})^2 + 2(\overline{x}-C)\sum_{k=1}^{n}(x_k-\overline{x}) + n(\overline{x}-C)^2\right\} = \\ = s^2 + \frac{n}{n-1}(\overline{x}-C)^2 \geq s^2,$$

com igualdade se e só se $C = \overline{x}$.

Anote-se ainda a extensão para média e a variância ponderadas.

Sejam

$$\overline{x}_w = \sum_{j=1}^{n} w_j x_j \quad \text{e} \quad s_w^2 = \frac{1}{n-1}\sum_{j=1}^{n} n_j\left(x_j-\overline{x}_w\right)^2,$$

onde $w_j = \frac{n_j}{n} \geq 0,\ j = 1,\ldots,n$, e $\sum_{j=1}^{n} w_j = 1$. Também

$$\frac{1}{n-1}\sum_{j=1}^{n} n_j\left(x_j-C\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n} n_j\left(x_j-\overline{x}_w+\overline{x}_w-C\right)^2 = \\ = \frac{1}{n-1}\left\{\sum_{j=1}^{n} n_j\left(x_j-\overline{x}_w\right)^2 + n\left(\overline{x}_w-C\right)^2\right\} = \\ = s_w^2 + \frac{n}{n-1}\left(\overline{x}_w-C\right)^2,$$

pois

$$\sum_{j=1}^{n} n_j \left(x_j - \overline{x}_w\right) = n \sum_{j=1}^{n} \frac{n_j}{n} \left(x_j - \overline{x}_w\right) = n \left(\overline{x}_w - \overline{x}_w\right) = 0.$$

Consequentemente a média ponderada minimiza os momento empíricos centrados ponderados de ordem 2, com as mesmas ponderações,

$$\frac{1}{n-1} \sum_{j=1}^{n} n_j \left(x_j - C\right)^2 \geq s_w^2,$$

com igualdade se e só se $C = \overline{x}_w$. Gilbert (1989, p. 7) comenta: "*To use a weighted mean in an unweighted analysis of variance (or vice versa) would be misleading, and therefore wrong*".

A constatação destes dois factos elementares deve ter inspirado Fisher a decompor as somas de quadrados que entram no cálculo do estimador da variância em parcelas independentes, medindo respectivamente a variabilidade dentro das amostras e a variabilidade entre amostras. Chamou a essa decomposição *ANOVA — ANalysis Of VAriance* , onde a palavra "análise"tem a mesma acepção que em Química: decomposição em partes elementares

Apresentamos naturalmente a situação mais elementar e de notação mais legível, habitualmente denominada análise da variância simples (*one-way analysis of variance*), mas os métodos da análise da variância valem para planos experimentais muito sofisticados, com algumas questões não elementares

sobre partição das somas de quadrados e independência, e problemas de não ortogonalidade; recomendamos a leitura de Gilbert (1989), fascinante.

Considere-se então que temos k amostras aleatórias independentes

$$\begin{aligned}
\boldsymbol{X_1} &= \left(X_{11}, \ldots, X_{1n_1}\right) \\
\boldsymbol{X_2} &= \left(X_{21}, \ldots, X_{2n_2}\right) \\
&\vdots \\
\boldsymbol{X_k} &= \left(X_{k1}, \ldots, X_{kn_k}\right)
\end{aligned}$$

extraídas de populações $Gaussiana\,(\mu_i,\, \sigma)$, $i = 1, \ldots, k$. Por simplicidade de exposição, vamos supor que $\boldsymbol{X_i}$ é um vector com n_i registos dos efeitos do $i-$ésimo tratamento experimental dos k que estamos a comparar, supondo que a variância é a mesma para todos os tratamentos.

O efeito médio global pode ser avaliado por

$$\overline{\overline{X}} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

onde $N = \sum_{i=1}^{k} n_i$ é a dimensão da amostra global.

Por outro lado, a média de cada uma das amostras

$$\overline{X}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

pode ser considerada como um sinal do efeito médio de cada tratamento, isto é do correspondente valor médio μ_{i}.

A soma dos quadrados que nos permite avaliar a variância da amostra pode então ser particionada numa soma de duas parcelas:

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{\overline{X}}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\bullet} + \overline{X}_{i\bullet} - \overline{\overline{X}}\right)^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\bullet}\right)^2 + \sum_{i=1}^{k} n_i\left(\overline{X}_{i\bullet} - \overline{\overline{X}}\right)^2 = WSS + BSS,$$

uma vez que o termo rectangular, como já justificámos, é nulo.

Enquanto TSS (*Total Sum of Squares*) mede a variabilidade total, WSS (*Within Sum of Squares*) mede a variabilidade dentro das amostras, e BSS (*Between Sum of Squares*) mede a variabilidade entre as amostras.

Os "quadrados médios"

$$S_p^2 = \frac{WSS}{\sum_{i=1}^{k}(n_i - 1)} \quad \text{e} \quad \frac{BSS}{k-1}$$

são estimadores da variância.

De facto, se aquele pressuposto for verdadeiro[(2)],

$$\frac{\dfrac{BSS}{k-1}}{\dfrac{WSS}{N-k}} \frown F_{k-1,\,N-k}. \tag{1.2}$$

[(2)] Note que de $\frac{TSS}{\sigma^2} = \frac{BSS}{\sigma^2} + \frac{WSS}{\sigma^2}$, com $\frac{TSS}{\sigma^2} \frown \chi^2_{N-1}$, $\frac{BSS}{\sigma^2} \frown \chi^2_{k-1}$ e $\frac{WSS}{\sigma^2} \frown \chi^2_{N-k}$ se conclui que BSS e WSS são independentes — um caso particular do teorema de Cochran (1934).

Mas enquanto

$$\frac{WSS}{N-k}$$

é um estimador centrado de σ^2, quer $\mu_i = \mu$, $i = 1, \ldots, k$, quer não, pois

$$\frac{\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i\bullet}\right)^2}{\sigma^2} \frown \chi^2_{n_i - 1} \Longrightarrow \mathbb{E}\left[\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i\bullet}\right)^2\right] = (n_i - 1)\,\sigma^2,$$

$i = 1, \ldots, k$, e consequentemente

$$\mathbb{E}\left[\frac{WSS}{N-k}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\bullet}\right)^2}{N-k}\right] = \frac{1}{N-k}\sum_{j=1}^{k}(n_i - 1)\,\sigma^2 = \sigma^2,$$

o mesmo não acontece com $\dfrac{BSS}{k-1}$.

De facto, de

$$\overline{X}_{i\bullet} = \frac{1}{n_i}\sum_{k=1}^{n_i} X_{ij} \frown Gaussiana\left(\mu_i, \frac{\sigma}{\sqrt{n_i}}\right)$$

e

$$\overline{\overline{X}} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} \frown Gaussiana\left(\mu, \frac{\sigma}{\sqrt{N}}\right),$$

onde $N = \sum_{i=1}^{k} n_i$ é a dimensão da amostra global e $\mu = \frac{1}{N}\sum_{i=1}^{k} n_i \mu_i$

o valor médio da população global, segue-se que

$$\mathbb{E}\left[\sum_{i=1}^{k} n_{i}\left(\overline{X}_{i\bullet}-\overline{\overline{X}}\right)^{2}\right]=\mathbb{E}\left[\sum_{i=1}^{k} n_{i}\overline{X}_{i\bullet}^{2}-N\,\overline{\overline{X}}^{2}\right]=$$
$$=\sum_{i=1}^{k} n_{i}\left(\mu_{i}^{2}+\frac{\sigma^{2}}{n_{i}}\right)-N\left(\mu^{2}+\frac{\sigma^{2}}{N}\right)=$$
$$=\sum_{i=1}^{k} n_{i}\mu_{i}^{2}+k\,\sigma^{2}-N\,\mu^{2}-\sigma^{2}=(k-1)\,\sigma^{2}+\sum_{i=1}^{k} n_{i}\mu_{i}^{2}-N\,\mu^{2}.$$

Consequentemente,

$$\mathbb{E}\left[\frac{BSS}{k-1}\right]=\sigma^{2}+\frac{\sum_{i=1}^{k} n_{i}\mu_{i}^{2}-N\,\mu^{2}}{k-1}.$$

Ora, pela desigualdade de Cauchy-Schwarz,

$$N\,\mu^{2}=N\,\frac{\left(\sum_{i=1}^{k} n_{i}\mu_{i}\right)^{2}}{N^{2}}=\frac{1}{N}\left(\sum_{i=1}^{k}\sqrt{n_{i}}\,\sqrt{n_{i}}\mu_{i}\right)^{2}\leq$$
$$\leq\frac{1}{N}\sum_{i=1}^{k} n_{i}\,\sum_{i=1}^{k}\left(\sqrt{n_{i}}\mu_{i}\right)^{2}=\sum_{i=1}^{k} n_{i}\mu_{i}^{2},$$

pelo que a segunda parcela é sempre não negativa, tomando o valor zero se e só se $\mu_{i}=\mu,\ i=1,\ldots,k$.

A tabela da ANOVA (*ANalysis Of VAriance*) simples apresenta a informação fundamental:

Convém notar alguns aspectos, perspectivando agora em termos de amostra aleatória e não de amostra observada:

Tabela 1.1: Tabela da ANOVA (simples).

Fontes de Variação	Graus de Liberdade	Quadrados Médios	razão F	p
$BSS = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{\overline{x}}\right)^2$	$k-1$	$\frac{BSS}{k-1}$	$F_0 = \frac{\frac{BSS}{k-1}}{\frac{WSS}{N-k}}$	$\mathbb{P}\left(F_{k-1,N-k} \geq F_0\right)$
$WSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i\right)^2$	$N-k$ (*)	$\frac{WSS}{N-k}$		
$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{\overline{x}}\right)^2$	$N-1$			

(*) $\sum_{i=1}^{k} (n_i - 1) = N - k.$

- $\frac{BSS}{\sigma^2} \frown \chi^2_{k-1}$ e $\frac{WSS}{\sigma^2} \frown \chi^2_{N-k}$.
- Uma vez que $\frac{BSS}{\sigma^2} + \frac{WSS}{\sigma^2} = \frac{TSS}{\sigma^2} \frown \chi^2_{N-1}$, podemos concluir que as variáveis aleatórias $BSS = \sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2$ e $WSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2$ são independentes.
- Do exposto, conclui-se que a razão

$$\frac{\frac{\frac{BSS}{\sigma^2}}{k-1}}{\frac{\frac{WSS}{\sigma^2}}{N-k}} = \frac{\frac{BSS}{k-1}}{\frac{WSS}{N-k}} \frown F_{k-1,N-k},$$

  pois é o quociente de duas variáveis aleatórias independentes, respectivamente uma χ^2_{k-1} dividida pelo seu número de graus de liberdade $k-1$, e uma χ^2_{N-k} dividida pelo seu número de graus de liberdade $N-k$.
- No caso de nem todas as populações de que foram extraídas as amostras terem o mesmo valor médio, como $\frac{BSS}{k-1}$ é um estimador enviesado de σ^2, que sobre-estima, a razão F tenderá a evidenciar essa diversidade de valores médios.
- Anote-se ainda que $\frac{WSS}{N-k}$ é a generalização da variância ponderada S_p^2 que usámos na comparação de valores médios com a t de Student no caso de duas amostras independentes, assumindo homocedasticidade.

Como tínhamos referido, expusemos a situação elementar de análise da variância simples. Mas a metodologia de Fisher frutificou, e um dos problemas centrais da Estatística Aplicada é a

"decomposição de quadrados", como criticamente expõe Gilbert (1989). Não aumenta porém a compreensão do problema que nos ocupa, e por isso preferimos reperspectivar a questão em termos de razão de verosimilhanças.

1.3 Comparação de valores médios com base em duas amostras independentes homocedásticas: o teste t como teste de razão de verosimilhanças

Maximizar a função de verosimilhança é equivalente a maximizar a log-verosimilhança, uma vez que a função logaritmo é crescente.

No que se segue, fixada uma amostra $\boldsymbol{x}$ extraída de uma população gaussiana com parâmetros μ e σ, denotamos a função de verosimilhança $L(\mu, \sigma|\boldsymbol{x})$, usando notações do mesmo tipo no caso de duas amostras. Consulte-se Casella e Berger (2002, nomeadamente pp. 290 e seguintes) sobre o papel da função de verosimilhança na redução dos dados.

Começamos por registar um resultado bem conhecido no âmbito de estimação dos parâmetros de uma população gaussiana, porque vai ser a base de todo o desenvolvimento que segue:

Teorema 1.3.1.

Seja $\boldsymbol{x} = (x_1, \ldots, x_n)$ *realização de uma amostra aleatória de dimensão* n *de uma população* $X \frown Gaussiana(\mu, \sigma)$.

Sejam $\boldsymbol{x}$ *e* σ *fixos. Então* $\mu \mapsto L(\mu, \sigma \mid \boldsymbol{x})$ *é máxima para* $\hat{\mu} = \overline{x}$.

Sejam $\boldsymbol{x}$ e μ fixos. Então $\sigma \mapsto L(\mu, \sigma \mid \boldsymbol{x})$ é máxima para $\hat{\sigma}$, com $\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu)^2$.

Sejam μ e σ parâmetros desconhecidos. Então a verosimilhança é máxima para $\mu = \overline{x}$ e $\sigma = \hat{\sigma}$.

Demonstração:

A verosimilhança associada a uma amostra $\boldsymbol{x} = (x_1, \ldots, x_n)$ de uma população gaussiana é

$$L(\mu, \sigma \mid \boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[-\frac{1}{2\,\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2 \right]$$

e a função de suporte

$$f(\mu, \sigma) = \ln L(\mu, \sigma \mid \boldsymbol{x}) = -n(\ln \sqrt{2\pi} + \ln \sigma) - \frac{1}{2\,\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2$$

Tem-se $\frac{\mathrm{d}}{\mathrm{d}\mu} \ln L(\mu, \sigma \mid \boldsymbol{x}) = \frac{1}{\sigma^2} \sum_{k=1}^{n} (x_k - \mu)$ e aquela expressão anula-se, passando de positiva a negativa, apenas em $\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k = \overline{x}$.

Por outro lado, $\frac{\mathrm{d}}{\mathrm{d}\sigma} L(\mu, \sigma \mid \boldsymbol{x}) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{k=1}^{n} (x_k - \mu)^2$ e esta expressão anula-se, passando de positiva a negativa, apenas em $\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu)^2$.

No caso de μ e σ serem ambos parâmetros desconhecidos,

observe-se que a verosimilhança é contínua em $\mathbb{R} \times \mathbb{R}^+$ e

$$\begin{aligned}\lim_{\sigma \to 0} L(\mu, \sigma \mid \boldsymbol{x}) &= \lim_{\frac{1}{\sigma} \to +\infty} \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n}{\exp\left[\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2\right]} = 0 \\ &= \lim_{\mu \to \pm\infty} L(\mu, \sigma \mid \boldsymbol{x}),\end{aligned}$$

portanto atinge o máximo no interior do seu domínio. Derivando,

$$\begin{cases} \dfrac{\partial f}{\partial \mu}(\mu, \sigma) = \dfrac{n}{\sigma^2}(\overline{x} - \mu) = 0 \\ \\ \dfrac{\partial f}{\partial \sigma}(\mu, \sigma) = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3} \displaystyle\sum_{k=1}^{n} (x_k - \mu)^2 = 0 \end{cases},$$

concluindo-se assim que

$$\begin{cases} \hat{\mu} = \overline{x} \\ \hat{\sigma}^2 = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} (x_k - \overline{x})^2 \end{cases}$$

Pode-se justificar que a solução $(\overline{x}, \hat{\sigma})$ encontrada para o sistema de estacionaridade corresponde ao máximo neste caso notando que no expoente da verosimilhança se tem

$$-\sum_{k=1}^{n} (x_k - \mu)^2 = -\sum_{k=1}^{n} (x_k - \overline{x})^2 - n(\overline{x} - \mu)^2 \leq -\sum_{k=1}^{n} (x_k - \overline{x})^2,$$

com igualdade se e só se $\mu = \overline{x}$ (de acordo com o que foi visto na secção anterior). Logo, para cada valor de σ a verosimilhança

é máxima quando $\mu = \overline{x}$, e então basta achar o maximizante de $f(\sigma) = L(\overline{x}, \sigma \mid \boldsymbol{x})$ — que já antes mostrámos que é $\hat{\sigma}$. No que se segue deixaremos ao cuidado do leitor a verificação de que os pontos de estacionaridade encontrados são efectivamente maximizantes, que em geral segue as mesmas linhas.[(3)]

□

Abordemos agora os problemas de comparação de parâmetros populacionais com base na observação de duas amostras independentes.

Teorema 1.3.2.

Sejam $\boldsymbol{x}_1 = \left(x_{11}, \ldots, x_{1n_1}\right)$, $\boldsymbol{x}_2 = \left(x_{21}, \ldots, x_{2n_2}\right)$ *realizações de duas amostras aleatórias independentes, de populações* $X_1 \frown \mathit{Gaussiana}\,(\mu_1, \sigma_1)$ *e* $X_2 \frown \mathit{Gaussiana}\,(\mu_2, \sigma_2)$.

A função de verosimilhança

$$L\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right)$$

é máxima para

$$\hat{\mu}_1 = \overline{x}_1 \qquad\qquad \hat{\mu}_2 = \overline{x}_2$$

$$\hat{\sigma}_1^{\,2} = \frac{1}{n_1} \sum_{k=1}^{n_1} \left(x_{1k} - \overline{x}_1\right)^2 \qquad\qquad \hat{\sigma}_2^{\,2} = \frac{1}{n_2} \sum_{j=1}^{n_2} \left(x_{2j} - \overline{x}_2\right)^2$$

[(3)] Outra justificação possível para funções de duas variáveis é verificar que f é de classe C^2 em todo o domínio, um subconjunto aberto de $\mathbb{R}^2$, e satisfaz as duas condições de segunda ordem: (I) $f_{\mu\mu}(\overline{x}, \hat{\sigma}) = -n/\hat{\sigma}^2 < 0$ e $f_{\sigma\sigma}(\overline{x}, \hat{\sigma}) = -2n/\hat{\sigma}^2 < 0$ (basta uma destas segundas derivadas ser negativa); (II) o discriminante é $f_{\mu\mu}(\overline{x}, \hat{\sigma}) f_{\sigma\sigma}(\overline{x}, \hat{\sigma}) - f_{\mu\sigma}^{\,2}(\overline{x}, \hat{\sigma}) = 2n^2/\hat{\sigma}^4 > 0$. Como há só um máximo local, trata-se do máximo absoluto.

sendo o máximo igual a

$$L\left(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) = \left(2\pi \mathrm{e} \hat{\sigma}_1^{\,2}\right)^{-\frac{n_1}{2}} \left(2\pi \mathrm{e} \hat{\sigma}_2^{\,2}\right)^{-\frac{n_2}{2}}.$$

Demonstração:

Como estamos a trabalhar com amostras independentes de Gaussianas,

$$L\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) = L\left(\mu_1, \sigma_1 \mid \boldsymbol{x}_1\right) L\left(\mu_2, \sigma_2 \mid \boldsymbol{x}_2\right).$$

sendo ambos os factores do segundo membro positivos, e portanto os valores que maximizam $L\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right)$ são os que maximizam $L\left(\mu_1, \sigma_1 \mid \boldsymbol{x}_1\right)$ e $L\left(\mu_2, \sigma_2 \mid \boldsymbol{x}_2\right)$. Face aos resultados do teorema anterior:

$$\hat{\mu}_1 = \overline{x}_1 \qquad\qquad \hat{\mu}_2 = \overline{x}_2$$

$$\hat{\sigma}_1^{\,2} = \frac{1}{n_1} \sum_{k=1}^{n_1} \left(x_{1k} - \overline{x}_1\right)^2 \qquad \hat{\sigma}_2^{\,2} = \frac{1}{n_2} \sum_{j=1}^{n_2} \left(x_{2j} - \overline{x}_2\right)^2.$$

Por substituição, obtém-se o máximo.

Teorema 1.3.3.

Sejam $\boldsymbol{x}_1 = \left(x_{11}, \ldots, x_{1n_1}\right)$ *e* $\boldsymbol{x}_2 = \left(x_{21}, \ldots, x_{2n_2}\right)$ *as realizações de duas amostras independentes, de populações* $X_1 \frown$ *Gaussiana* (μ_1, σ) *e* $X_2 \frown$ *Gaussiana* (μ_2, σ). *Com esta*

hipótese adicional de homocedasticidade, $L(\mu_1, \mu_2, \sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$ *é máxima para*

$$\hat{\mu}_1 = \overline{x}_1 \qquad \hat{\mu}_2 = \overline{x}_2 \qquad \hat{\sigma}_p^{\,2} = \frac{n_1 \hat{\sigma}_1^{\,2} + n_2 \hat{\sigma}_2^{\,2}}{n_1 + n_2},$$

e o valor máximo da função de verosimilhança é $\left(2\pi \mathrm{e} \hat{\sigma}_p^{\,2}\right)^{-\frac{n_1+n_2}{2}}$

onde

$$\hat{\sigma}_p^{\,2} = \frac{\sum_{k=1}^{n_1} \left(x_{1k} - \overline{x}_1\right)^2 + \sum_{j=1}^{n_2} \left(x_{2j} - \overline{x}_2\right)^2}{n_1 + n_2} = \frac{n_1 \hat{\sigma}_1^{\,2} + n_2 \hat{\sigma}_2^{\,2}}{n_1 + n_2}$$

é a estimativa ponderada de máxima verosimilhança da variância comum, com base na informação das duas amostras.

Demonstração:

Do resultado anterior, sabemos $\hat{\mu}_1 = \overline{x}_1$ e $\hat{\mu}_2 = \overline{x}_2$; há agora que determinar o valor de σ que maximiza

$$L(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{n_1+n_2} \times$$
$$\times \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n_1}\left(x_{1k} - \overline{x}_1\right)^2 + \sum_{j=1}^{n_2}\left(x_{2j} - \overline{x}_2\right)^2\right)\right].$$

Logaritmizando,

$$\ln L(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = -\frac{1}{2}\ (n_1+n_2)\ \ln(2\pi) -$$
$$-\ (n_1+n_2)\ \ln\sigma - \frac{1}{2\sigma^2}\left(\sum_{k=1}^{n_1}(x_{1k}-\overline{x}_1)^2 + \sum_{j=1}^{n_2}\left(x_{2j}-\overline{x}_2\right)^2\right)$$

e derivando em ordem a σ e igualando a 0, vem

$$\sigma^2 = \frac{\sum_{k=1}^{n_1}(x_{1k}-\overline{x}_1)^2 + \sum_{j=1}^{n_2}\left(x_{2j}-\overline{x}_2\right)^2}{n_1+n_2} = \frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{n_1+n_2}\,.$$

O máximo de $L(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$ obtém-se por substituição. □

Estamos então em condições de abordar o teste da razão verosimilhanças para testar a igualdade de variâncias com base na observação de duas amostras. Como habitualmente, denotando $\boldsymbol{\Theta}$ o espaço do parâmetro $\boldsymbol{\theta}$, de que $\boldsymbol{\Theta_0}$ é um subespaço

$$\Lambda(\boldsymbol{x}) = \frac{\sup\limits_{\boldsymbol{\Theta_0}} L(\boldsymbol{\theta} \mid \boldsymbol{x})}{\sup\limits_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \boldsymbol{x})} = \frac{L_0(\boldsymbol{\theta} \mid \boldsymbol{x})}{L_1(\boldsymbol{\theta} \mid \boldsymbol{x})}$$

é a estatística do teste de razão de verosimilhanças para testar

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta_0} \ \textit{vs.}\ H_A : \boldsymbol{\theta} \in \boldsymbol{\Theta} - \boldsymbol{\Theta_0}. \qquad \square$$

(Anote-se que escrevemos $\sup\limits_{\boldsymbol{\Theta_0}} L(\boldsymbol{\theta} \mid \boldsymbol{x}) = L_0(\boldsymbol{\theta} \mid \boldsymbol{x})$ e $\sup\limits_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \boldsymbol{x}) = L(\boldsymbol{\theta} \mid \boldsymbol{x})$ em vez de usar as notações, porventura mais correctas, $\sup\limits_{\boldsymbol{\Theta_0}} L(\boldsymbol{\theta} \mid \boldsymbol{x}) = L(\boldsymbol{\theta}_0 \mid \boldsymbol{x})$ e $\sup\limits_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \boldsymbol{x}) =$

$L(\boldsymbol{\theta}_1 \mid \boldsymbol{x})$, usadas por exemplo por Casella and Berger (2002, veja-se p. 375), por a convenção que adoptamos simplificar substancialmente as notações na exposição que se segue.)

No caso de duas amostras, usamos as extensões naturais desta notação. Mais uma vez recomendamos a lúcida exposição de Casella and Berger (2002, pp. 374 e seguintes) sobre os testes de razão de verosimilhanças e as suas boas propriedades.

Teorema 1.3.4.

Sejam $\boldsymbol{x}_1 = \left(x_{11}, \ldots, x_{1n_1}\right)$ *e* $\boldsymbol{x}_2 = \left(x_{21}, \ldots, x_{2n_2}\right)$ *as realizações de duas amostras aleatórias independentes, extraídas de populações* $X_1 \frown \mathit{Gaussiana}\,(\mu_1, \sigma_1)$ *e* $X_2 \frown \mathit{Gaussiana}\,(\mu_2, \sigma_2)$*, respectivamente.*

Para testar $H_0 : \sigma_1 = \sigma_2 = \sigma$ *vs.* $H_1 : \sigma_1 \neq \sigma_2$*, a razão de verosimilhanças é*

$$\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) = \frac{\sup\limits_{\sigma_1=\sigma_2=\sigma} L\left(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right)}{\sup L\left(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right)} =$$

$$= \frac{\left(\dfrac{\hat{\sigma}_2^{\,2}}{\hat{\sigma}_1^{\,2}}\right)^{\frac{n_2}{2}}}{\left(n_1 + n_2 \dfrac{\hat{\sigma}_2^{\,2}}{\hat{\sigma}_1^{\,2}}\right)^{\frac{n_1+n_2}{2}}}$$

e a região crítica é definida por

$$\frac{s_2^{\,2}}{s_1^{\,2}} \leq F_{1-\frac{\alpha}{2}\,;\,n_2-1,n_1-1} \;\; ou \;\; \frac{s_2^{\,2}}{s_1^{\,2}} \geq F_{\frac{\alpha}{2}\,;\,n_2-1,n_1-1},$$

onde

$$s_1^2 = \frac{n_1}{n_1 - 1}\,\hat{\sigma}_1^2 \quad e \quad s_2^2 = \frac{n_2}{n_2 - 1}\,\hat{\sigma}_2^2$$

são as estimativas das variâncias populacionais fornecidas pelos estimadores centrados usuais.

Demonstração:

Sob a validade de H_0, estamos nas condições do teorema anterior, e nas condições do Teorema 1.3.2, sob validade de H_1. Então

$$\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\left(2\pi e \hat{\sigma}_p^2\right)^{-\frac{n_1+n_2}{2}}}{\left(2\pi e \hat{\sigma}_1^2\right)^{-\frac{n_1}{2}} \left(2\pi e \hat{\sigma}_2^2\right)^{-\frac{n_2}{2}}} =$$

$$= \frac{(n_1 + n_1)^{\frac{n_1+n_2}{2}} \left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right)^{\frac{n_2}{2}}}{\left(n_1 + n_2 \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right)^{\frac{n_1+n_2}{2}}}.$$

Consequentemente $\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2)$ é uma função de $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$.

Assim, a região de rejeição é definida por $\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \delta$, o que equivale a

$$\frac{\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right)^{\frac{n_2}{2}}}{\left(n_1 + n_2 \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right)^{\frac{n_1+n_2}{2}}} \leq \delta_0$$

com $\delta_0 = \delta(n_1+n_2)^{-\frac{n_1+n_2}{2}}$.

Considere-se $t = \dfrac{\hat{\sigma}_2^{\,2}}{\hat{\sigma}_1^{\,2}}$. A função

$$\varphi(t) = \frac{t^{\frac{n_2}{2}}}{(n_1+n_2 t)^{\frac{n_1+n_2}{2}}}$$

cujo gráfico é mostrado na Figura 2.2, é positiva, contínua e unimodal; consequentemente a região crítica é

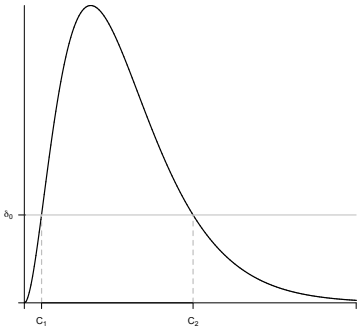

Figura 1.2: Forma da região crítica.

$$\mathcal{C} = \left\{(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathbb{R}^{n_1+n_2} : \frac{\hat{\sigma}_2^{\,2}}{\hat{\sigma}_1^{\,2}} \leq C_1\right\} \cup \left\{(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathbb{R}^{n_1+n_2} : \frac{\hat{\sigma}_2^{\,2}}{\hat{\sigma}_1^{\,2}} \geq C_2\right\}.$$

□

Como estatística de teste é mais conveniente usar o valor

$$T = \frac{(n_1 - 1)n_2\hat{\sigma}_2^2}{(n_2 - 1)n_1\hat{\sigma}_1^2} = \frac{S_2^2}{S_1^2} \quad \underset{|H_0 \text{ verd.}}{\frown} \quad F_{n_2-1,n_1-1},$$

proporcional a $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$, onde S_1^2 e S_1^2 são os estimadores centrados usuais das variâncias.

Recordamos que $T^* = \frac{1}{T} = \frac{S_1^2}{S_2^2} \quad \underset{|H_0 \text{ verd.}}{\frown} \quad F_{n_1-1,n_2-1}$ pode também, naturalmente, ser usada como estatística de teste. Escolhemos a que tornar mais simples a consulta da tabela dos quantis críticos da F de Fisher-Snedecor. (Na prática é mais usual fazer o teste de caudas equilibradas e não o da razão de verosimilhanças.)

Vamos agora abordar a comparação de médias de duas populações Gaussianas, no caso de as variâncias serem desconhecidas, mas admitindo homocedasticidade (e, como veremos adiante, admitir que a razão das variâncias é conhecida é uma variante simples deste resultado).

Teorema 1.3.5.

Sejam $\boldsymbol{x}_1 = \left(x_{11}, \ldots, x_{1n_1}\right)$ *e* $\boldsymbol{x}_2 = \left(x_{21}, \ldots, x_{2n_2}\right)$ *observações de amostras aleatórias independentes extraídas de populações Gaussianas com a mesma variância* σ^2*, desconhecida,* $X_1 \frown$ *Gaussiana* (μ_1, σ) *e* $X_2 \frown$ *Gaussiana* (μ_2, σ)*, respectivamente.*

Sob a restrição $\mu_2 - \mu_1 = \Delta$*, o máximo de* $L(\mu_1, \mu_2, \sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$*, para* $\boldsymbol{x}_1$ *e* $\boldsymbol{x}_2$ *fixos, corresponde ao ponto* $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_0)$*, onde*

$$\hat{\mu}_1 = \frac{n_1 \overline{x}_1 + n_2 (\overline{x}_2 - \Delta)}{n_1 + n_2},$$

$$\hat{\mu}_2 = \frac{n_1 (\overline{x}_1 + \Delta) + n_2 \overline{x}_2}{n_1 + n_2}$$

e

$$\hat{\sigma}_0^{\,2} = \hat{\sigma}_p^{\,2} + n_1 n_2 \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1 + n_2} \right)^2,$$

sendo o máximo da verosimilhança $\left(2\pi \mathrm{e} \hat{\sigma}_0^{\,2}\right)^{-\frac{n_1+n_2}{2}}$.

Demonstração:

Queremos maximizar

$$f(\mu_1, \mu_2, \sigma) = \ln L(\mu_1, \mu_2, \sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = -(n_1 + n_2) \ln(\sqrt{2\pi}) -$$
$$- (n_1 + n_2) \ln \sigma - \frac{1}{2\sigma^2} \left(\sum_{k=1}^{n_1} (x_{1k} - \mu_1)^2 + \sum_{j=1}^{n_2} (x_{2j} - \mu_2)^2 \right)$$

sob a restrição $\mu_2 - \mu_1 = \Delta$. Escreva-se $g(\mu_1, \mu_2, \sigma) = \mu_2 - \mu_1 - \Delta$.

Usando o método dos multiplicadores de Lagrange, tal máximo, se existir, é atingido no ponto $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_0)$ para o qual existe algum escalar λ tal que

$$\nabla f(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_0) = \lambda \nabla g(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_0).$$

Então

$$\left(\frac{1}{\hat{\sigma}_0^{\,2}} \sum_{k=1}^{n_1} (x_{1k} - \hat{\mu}_1)\,, \frac{1}{\hat{\sigma}_0^{\,2}} \sum_{j=1}^{n_2} \left(x_{2j} - \hat{\mu}_2\right), -\frac{n_1+n_2}{\hat{\sigma}_0} + \right.$$

$$\left. + \frac{1}{\hat{\sigma}_0^{\,3}} \left[\sum_{k=1}^{n_1} (x_{1k} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} \left(x_{2j} - \hat{\mu}_2\right)^2 \right] \right) = \lambda(-1, 1, 0)$$

o que quer dizer que

$$\begin{cases} \hat{\sigma}_0^{\,2} \;=\; \dfrac{1}{n_1+n_2} \left[\displaystyle\sum_{k=1}^{n_1} (x_{1k} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} \left(x_{2j} - \hat{\mu}_2\right)^2 \right] \\ \displaystyle\sum_{k=1}^{n_1} (x_{1k} - \hat{\mu}_1) \;=\; -\sum_{j=1}^{n_2} \left(x_{2j} - \hat{\mu}_2\right) \end{cases}$$

Desenvolvendo a última expressão, vem

$$n_1 \overline{x}_1 - n_1 \hat{\mu}_1 = -n_2 \overline{x}_2 + n_2 \hat{\mu}_2$$

donde, por $\hat{\mu}_2 - \hat{\mu}_1 = \Delta$, temos então[(4)]

$$\hat{\mu}_1 = \frac{n_1 \overline{x}_1 + n_2 \left(\overline{x}_2 - \Delta\right)}{n_1+n_2} \quad \text{e} \quad \hat{\mu}_2 = \frac{n_1 \left(\overline{x}_1 + \Delta\right) + n_2 \overline{x}_2}{n_1+n_2}\,.$$

Podemos por outro lado reescrever $\hat{\sigma}_0^{\,2}$ notando que

[(4)] Para verificar que $\hat{\mu}_1$ e $\hat{\mu}_2$ maximizam a verosimilhança restrita note que sob a hipótese nula $\mu_2 - \mu_1 = \hat{\mu}_2 - \hat{\mu}_1 = \Delta$, donde a igualdade $\sum (x_{1k} - \hat{\mu}_1) = -\sum \left(x_{2j} - \hat{\mu}_2\right)$ implica que $\sum (x_{1k} - \mu_1)^2 + \sum (x_{2j} - \mu_2)^2 = \sum (x_{1k} - \hat{\mu}_1)^2 + \sum (x_{2j} - \hat{\mu}_2)^2 + n_1 (\hat{\mu}_1 - \mu_1)^2 + n_2 (\hat{\mu}_2 - \mu_2)^2$.

$$\begin{aligned}
\hat{\mu}_1 - \overline{x}_1 &= \frac{n_2(\overline{x}_2 - \overline{x}_1 - \Delta)}{n_1 + n_2}; \\
\sum_{k=1}^{n_1}(x_{1k} - \hat{\mu}_1)^2 &= \sum_{k=1}^{n_1}(x_{1k} - \overline{x}_1)^2 + n_1(\hat{\mu}_1 - \overline{x}_1)^2 \\
&= \sum_{k=1}^{n_1}(x_{1k} - \overline{x}_1)^2 + n_1 n_2^2 \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1 + n_2}\right)^2
\end{aligned}$$

e da mesma forma

$$\sum_{j=1}^{n_2}(x_{2j} - \hat{\mu}_2)^2 = \sum_{j=1}^{n_2}(x_{2j} - \overline{x}_2)^2 + n_2 n_1^2 \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1 + n_2}\right)^2,$$

donde vem então

$$\hat{\sigma}_0^2 = \hat{\sigma}_p^2 + n_1 n_2 \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1 + n_2}\right)^2.$$

□

Podemos então estabelecer facilmente o teorema que se segue:

Teorema 1.3.6.

Sejam $\boldsymbol{x}_1 = \left(x_{11}, \ldots, x_{1n_1}\right)$ *e* $\boldsymbol{x}_2 = \left(x_{21}, \ldots, x_{2n_2}\right)$ *realizações de amostras aleatórias independentes extraídas de populações*

$X_1 \frown Gaussiana\,(\mu_1, \sigma)$ *e* $X_2 \frown Gaussiana\,(\mu_2, \sigma)$, *em que* σ *é desconhecido.*

Para testar $H_0 : \mu_2 - \mu_1 = \Delta$ *vs.* $H_1 : \mu_2 - \mu_1 \neq \Delta$, *a razão de verosimilhanças é*

$$\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[1 + \frac{1}{n_1 + n_2 - 2} \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)^2 \right]^{-\frac{n_1+n_2}{2}},$$

e a região crítica correspondente é dada por

$$\frac{|\overline{x}_2 - \overline{x}_1 - \Delta|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{\frac{\alpha}{2};\, n_1+n_2-2},$$

onde

$$s_p^2 = \frac{n_1 + n_2}{n_1 + n_2 - 2} \hat{\sigma}_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

é a estimativa da variância comum baseada no estimador ponderado centrado usual S_p^2.

Demonstração:

O teorema 1.3.5 permite encontrar a expressão para $L_0\,(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$ no cálculo de $\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2)$, para testar $H_0 : \mu_2 - \mu_1 = \Delta$ *vs.* $H_1 : \mu_2 - \mu_1 \neq \Delta$. Nessas condições,

$$L_0\,(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \left[2\pi \mathrm{e} \hat{\sigma}_p^2 + 2\pi \mathrm{e} n_1 n_2 \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1 + n_2} \right)^2 \right]^{-\frac{n_1+n_2}{2}}.$$

No que respeita $L_1(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$, que resulta de maximizar sem restrições a expressão

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n_1+n_2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{n_1}\left(x_{1k}-\mu_1\right)^2+\sum_{j=1}^{n_2}\left(x_{2j}-\mu_2\right)^2\right]\right\},$$

vimos no teorema 1.3.3 que $L_1(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \left(2\pi \mathrm{e}\hat{\sigma}_p^2\right)^{-\frac{n_1+n_2}{2}}$.

Assim, a razão de verosimilhanças é:

$$\begin{aligned}
\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \frac{L_0(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)}{L_1(\sigma \mid \boldsymbol{x}_1, \boldsymbol{x}_2)} \\
&= \left[1+\frac{n_1 n_2}{\hat{\sigma}_p^2}\left(\frac{\overline{x}_2-\overline{x}_1-\Delta}{n_1+n_2}\right)^2\right]^{-\frac{n_1+n_2}{2}} \\
&= \left[1+\frac{1}{n_1+n_2-2}\left(\frac{\overline{x}_2-\overline{x}_1-\Delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)^2\right]^{-\frac{n_1+n_2}{2}}
\end{aligned}$$

□

A região crítica é definida por

$$\left[1+\frac{1}{n_1+n_2-2}\left(\frac{\overline{x}_2-\overline{x}_1-\Delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)^2\right]^{-\frac{n_1+n_2}{2}} \leq \delta \iff \frac{|\overline{x}_2-\overline{x}_1-\Delta|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq c$$

onde $c = \sqrt{(n_1 + n_2 - 2)(\delta^{-\frac{2}{n_1+n_2}} - 1)}$.

A expressão $T_{S_p} = \dfrac{\overline{X}_2 - \overline{X}_1 - \Delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad |H_0 \overset{\frown}{verd.} \quad t_{n_1+n_2-2}$ é, naturalmente, a estatística de teste.

Pode estabelecer-se um resultado análogo quando as variâncias são diferentes, mas o quociente entre elas é conhecido; apresentámos o caso particular de esse quociente ser 1, diferindo para mais tarde o resultado geral. Veja-se a dedução completa na Secção 3 do Capítulo 3.

Não detalhamos a comparação de valores médios de duas populações gaussianas X e Y com base em amostras emparelhadas, uma vez que esse caso se reduz a inferência sobre o valor médio da variável aleatória gaussiana $D = X - Y$. Uma exposição detalhada pode ser consultada em Pestana e Velosa (2010, Teorema 5.9.12).

1.4 Comparação de valores médios a partir de k amostras independentes homocedásticas: o teste F como teste de razão de verosimilhanças

A estatística F do modelo linear (análise de variância e regressão linear) pode também ser derivada com base no critério da razão de verosimilhanças. Por exemplo, para o modelo da análise de variância simples, $Y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, 2, ...k$, $j = 1, 2, ...n_i$, em

que os μ_i são os valores médios para cada grupo e ϵ_{ij} são erros aleatórios gaussianos com valor médio nulo e a mesma variância desconhecida, σ^2.

Teorema 1.4.1.

Sejam $Y_{ij} = \mu_i + \epsilon_{ij}$, *com* $i = 1, 2, \ldots, k$ *e* $j = 1, 2, \ldots, n_i$, k *amostras aleatórias de dimensões* $n_1, n_2, \ldots, n_k$, *respectivamente,* $N = n_1 + n_2 + \cdots + n_k$ *a dimensão da amostra global resultante de combinar as diversas subamostras,* μ_i *constantes e* $\epsilon_{ij} \frown \mathit{Gaussiana}\,(0, \sigma)$ *erros aleatórios independentes.*

Então a verosimilhança das realizações $\boldsymbol{y}_i$ *das* k *amostras é máxima para* $\hat{\mu}_i = \overline{y}_{i\bullet}$, *em que* $\overline{y}_{i\bullet}$ *é a média da amostra proveniente do* i*-ésimo grupo, e*

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{i\bullet}\right)^2.$$

O máximo absoluto da verosimilhança é

$$L_1(\boldsymbol{y}) = \left(2\pi \text{ e } \widehat{\sigma}^2\right)^{-\frac{N}{2}}.$$

Demonstração:

Dadas as k amostras $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k)$ a verosimilhança é

$$L(\mu_1, \mu_2, ..., \mu_k, \sigma \mid \boldsymbol{y}) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^N} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \mu_i\right)^2\right]$$

e portanto

$$f(\mu_1, \mu_2, ..., \mu_k, \sigma) = \ln\, L\,(\mu_1, \mu_2, ..., \mu_k, \sigma \mid \boldsymbol{y}) =$$
$$= -N\, \ln\, \sqrt{2\pi} - N\, \ln\, \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \mu_i\right)^2 .$$

Derivando e igualando a zero, obtemos as estimativas de máxima verosimilhança para os valores médios e a variância. As condições de estacionaridade são

$$\frac{\partial f}{\partial \mu_i} = \frac{1}{\sigma^2} \sum_{j=1}^{n_i} \left(y_{ij} - \mu_i\right) = 0$$

e

$$\frac{\partial f}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \mu_i\right)^2 = 0.$$

Procedendo do mesmo modo do que no caso de duas populações com variâncias iguais, chega-se aos valores

$$\widehat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \overline{y}_{i\bullet} \qquad \widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{i\bullet}\right)^2 .$$

Atendendo à forma de f, é fácil concluir que estes correspondem de facto ao máximo da verosimilhança.

O máximo de $L\,(\mu_1, \mu_2, ..., \mu_k, \sigma \mid \boldsymbol{y})$ é obtido por substituição:

$$L\,(\widehat{\mu}_1, \ldots, \widehat{\mu}_k, \widehat{\sigma} \mid \boldsymbol{y}) = \frac{1}{\left(\sqrt{2\pi}\ \widehat{\sigma}\right)^N} \exp\left(-\frac{N}{2\widehat{\sigma}^2}\ \widehat{\sigma}^2\right) = \left(2\pi\, \mathrm{e}\, \widehat{\sigma}^2\right)^{-\frac{N}{2}} .$$

□

Teorema 1.4.2.

Nas mesmas condições do que o resultado anterior, suponha-se que se verifica a hipótese $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

Com essa restrição adicional, a verosimilhança é máxima para $\hat{\mu}_1 = \hat{\mu}_2 = \cdots = \hat{\mu}_k = \overline{\overline{y}}$, *a média global nas amostras combinadas, e*

$$\widehat{\sigma}_0^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{\overline{y}} \right)^2 ,$$

tomando o valor

$$\left(2\pi \text{ e } \widehat{\sigma}_0^2 \right)^{-\frac{N}{2}} .$$

Demonstração:

Se H_0 for verdadeira, podemos escrever $\mu = \mu_1 = \cdots = \mu_k$, e a verosimilhança reduz-se a

$$L\left(\mu, \sigma \mid \boldsymbol{y}\right) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^N} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \mu \right)^2 \right] .$$

Nessas condições,

$$\frac{\partial f}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \mu \right) \quad \text{e} \quad \frac{\partial f}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \mu \right)^2 ,$$

e portanto as estimativas de máxima verosimilhança sujeitas à restrição H_0 são:

$$\widehat{\mu}_0 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \overline{\overline{y}} \quad \text{e} \quad \widehat{\sigma}_0^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{\overline{y}} \right)^2 .$$

O máximo de $L(\mu, \sigma \mid \boldsymbol{y})$ obtém-se com facilidade por substituição:

$$L(\widehat{\mu}_0, \widehat{\sigma}_0 \mid \boldsymbol{y}) = \frac{1}{\left(\sqrt{2\pi}\ \widehat{\sigma}_0^2\right)^N} \exp\left(-\frac{N}{2\widehat{\sigma}_0^2}\ \widehat{\sigma}_0^2\right) = \left(2\pi \text{ e } \widehat{\sigma}_0^2\right)^{-\frac{N}{2}}.$$

□

Teorema 1.4.3.

Sejam $Y_{ij} = \mu_i + \epsilon_{ij}$, *com* $i = 1, 2, \ldots, k$ *e* $j = 1, 2, \ldots, n_i$, k *amostras aleatórias de dimensões* $n_1, n_2, \ldots, n_k$, *respectivamente,* $N = n_1 + n_2 + \cdots + n_k$ *a dimensão da amostra global resultante de combinar as diversas subamostras,* μ_i *constantes e* $\epsilon_{ij} \frown$ *Gaussiana* $(0, \sigma)$ *erros aleatórios independentes.*

A razão de verosimilhanças para testar $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ *vs.* H_1 : *"Os valores médios não são todos iguais" é dada por*

$$\Lambda(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k) = \left(1 + \frac{k-1}{n-k}\ F\right)^{-\frac{N}{2}},$$

em que a estatística

$$F = \frac{\dfrac{BSS}{k-1}}{\dfrac{WSS}{n-k}}$$

tem distribuição F *de Fisher-Snedecor com* $(k-1, N-k)$ *graus de liberdade.*

A região crítica é dada por $F \geq F_{k-1, N-k;\, 1-\alpha}$.

Demonstração:

Atendendo aos resultados anteriores, a razão de verosimilhanças é dada por

$$\Lambda(\boldsymbol{y}) = \frac{\sup\limits_{\Theta_0} L(\boldsymbol{y})}{\sup\limits_{\Theta} L(\boldsymbol{y})} = \left(\frac{\widehat{\sigma}_0^{\,2}}{\widehat{\sigma}^2}\right)^{-\frac{N}{2}} = \left[\frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(y_{ij}-\overline{\overline{y}}\right)^2}{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{i.}\right)^2}\right]^{-\frac{N}{2}} =$$

$$= \left(\frac{TSS}{WSS}\right)^{-\frac{N}{2}} = \left(1+\frac{TSS-WSS}{WSS}\right)^{-\frac{N}{2}} = \left(1+\frac{k-1}{N-k}F\right)^{-\frac{N}{2}}.$$

O teste da razão de verosimilhanças consiste portanto em rejeitar H_0 para

$$\left(1+\frac{k-1}{N-k}F\right)^{-\frac{N}{2}} \leq \delta \Longleftrightarrow F \geq c,$$

com $c = \frac{N-k}{k-1}\left(\delta^{-\frac{2}{N}} - 1\right)$. □

A estatística de teste é portanto a razão

$$F = \frac{\frac{BSS}{k-1}}{\frac{WSS}{N-k}} \quad \underset{verd.}{\overset{\frown}{|H_0}} \quad F_{k-1,N-k}.$$

A dedução do teste da razão de verosimilhanças para outros modelos de análise de variância é análoga, bem como para

modelos de regressão linear. Rohatgi (1976, p. 502–503) faz a dedução do teste da razão de verosimilhanças para o modelo linear genérico $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e}$.

Capítulo 2

Comparação de Valores Médios de Gaussianas com Quociente das Variâncias Desconhecido

2.1 Behrens, Fisher e Jeffreys, uma solução à procura do problema

"*It has been repeatedly stated, perhaps through a misreading of the last paragraph, that our method involves the "assumption" that the two variances are equal. This is an incorrect form of the statement; the equality of the variances is a necessary part of the hypothesis to be tested, namely that the two samples*

> *are drawn from the same normal population. The validity of the t-test as a test of hypothesis, is therefore absolute, and requires no assumption whatever. It would of course be legitimate to make a different test of significance appropriate to the question: Might these samples have been drawn from different normal populations having the same mean? This problem has, in fact, been solved, but in relation to the real situations arising in biological research, the question it answers appears to be somewhat academic.*"
>
> R. A. Fisher, *Statistical Methods for Research Workers*, p. 124-125.

> "*The mathematical problem of the comparison of means of samples, not only small in size, but for which there is no reason a priori to dismiss the largest imaginable differences in precision, was of mathematical interest, and potential experimental importance, though it is not common to find realistic data which present this problem, partly because they are rarely sought.*"
>
> R. A. Fisher, *Statistical Methods and Scientific Inference*, p. 97.

Para efectuar testes de hipóteses sobre o valor médio de uma gaussiana cuja variância é desconhecida com base nos valores observados de uma amostra aleatória, a estatística de teste apropriada é

$$t = \frac{\overline{X} - \mu_0}{S\sqrt{\frac{1}{n}}},$$

que tem distribuição t de Student com $n-1$ graus de liberdade caso $H_0 : \mu = \mu_0$ seja verdadeira (e a variável fulcral $t = \frac{\overline{X}-\mu}{S\sqrt{\frac{1}{n}}}$, é apropriada para construir intervalos de confiaça para μ).

De forma semelhante, para fazer inferências sobre a diferença entre os valores médios de duas gaussianas com a mesma variância desconhecida, com base em observações de amostras independentes extraídas de cada uma das populações, usamos

$$T_{S_p} = \frac{\overline{X}_2 - \overline{X}_1 - \delta}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

(onde a variância ponderada S_p^2 é o estimador adequado da variância comum desconhecida σ^2), que tem distribuição t com $n_1 + n_2 - 2$ graus de liberdade caso $H_0 : \mu_2 - \mu_1 = \delta$ seja verdadeira.

Em ambos os casos, a estatística usada é um quociente cujo numerador é um estimador com valor médio nulo (caso a hipótese nula seja verdadeira) e cujo denominador é a raiz quadrada de um estimador centrado da variância do numerador — um estimador do desvio padrão, portanto.

É então natural tentar seguir o mesmo caminho para estudar a diferença entre os valores médios de duas gaussianas independentes com variâncias desconhecidas mas possivelmente distintas. Faz todo o sentido usar de novo $\overline{X}_2 - \overline{X}_1 - \delta$ como indicador da diferença real entre os valores médios, e a variância deste estimador é

$$\text{var}\left[\overline{X}_2 - \overline{X}_1 - \delta\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

o que nos levaria a usar a estatística

$$T_{S_1,S_2} = \frac{\overline{X}_2 - \overline{X}_1 - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\,. \tag{2.1}$$

Infelizmente, a distribuição de T_{S_1,S_2} não é a de uma t de Student. Não apenas não é uma t de Student como é uma distribuição pouco tratável, cuja função densidade de probabilidade não pode ser explicitada em forma analítica simples. De resto, se as variâncias populacionais forem distintas não existe nenhuma função $v = f(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2)$ — mais geralmente ainda, nenhuma função invariante para permutações de cada uma das amostras aleatórias — que permita chegar a uma distribuição t de Student exacta (Scheffé, 1944, cf. introdução do Capítulo 5).

Esta dificuldade é por vezes ultrapassada admitindo que se dispõe de grandes amostras e fazendo uma aproximação pela distribuição normal. Na base desse procedimento está a convicção, ou esperança, de que S_1^2 e S_2^2 sejam estimativas suficientemente precisas de σ_1^2 e σ_2^2 para que se possa actuar como se as variâncias fossem conhecidas, sem grande prejuízo. Mas claro que isto não é completamente satisfatório.

Outra atitude, por vezes mesmo adoptada por ignorância, consiste em usar o teste t válido apenas para situações de homocedasticidade. Mas é óbvio que não é o mais indicado, e se as duas populações tiverem variâncias consideravelmente diferentes o desempenho do teste t pode ser bastante mau em termos de potência — as diferenças entre as variâncias "mascaram" possíveis diferenças entre os valores médios, baixando o nível de significância efectivo e levando a que uma hipótese nula falsa seja rejeitada com menos frequência do que devia.

Várias soluções mais eficientes têm sido propostas para fazer inferência sobre a diferença entre dois valores médios de populações gaussianas quando estas têm variâncias diferentes, com base em amostras extraídas de amostras aleatórias independentes, muitas vezes denominado problema de Behrens-Fisher.

A ausência de uma solução única que seja consensual parece dever-se, por um lado, às dificuldades computacionais envolvidas, que levam a procurar aproximações mais pragmáticas, por outro à ausência de critérios de optimalidade que nos guiem na escolha do teste (veremos por exemplo mais adiante que o critério da razão de verosimilhanças não indica uma função única dos dados para ser usada como estatística de teste), e por outro lado ainda, segundo parece, à própria controvérsia que cerca algumas das soluções, que levam a repensar os fundamentos da inferência estatística, e por isso não são aceites por todos.

Neste capítulo, começamos por apresentar a abordagem de Fisher e a controvérsia que ela suscitou; apresentamos também brevemente a forma como Jeffreys solucionou a questão. A similitude das duas abordagens quase leva a crer que Fisher seria igualmente um bayesiano, embora um trabalho posterior de Lindley (1958) mostre que as duas metodologias conduzem à mesma solução apenas quando o problema é equivalente — o parâmetro de interesse não precisa de ser de localização, mas tem de poder ser transformado noutro que o seja, veja-se o final da p. 103 do artigo do Lindley (1958) — à estimação de um parâmetro de localização.

Ainda neste capítulo, retomamos a abordagem dos testes de razões de verosimilhanças, o que permite verificar que só existe uma solução analiticamente tratável nos moldes clássicos se a razão entre as variâncias, $\theta = \sigma_1^2/\sigma_2^2$, for conhecida. Este facto

indicia a inevitabilidade das conclusões de Scheffé (1944), que expomos no Capítulo 3.

2.1.1 Os desenvolvimentos de Fisher

Behrens (1929) foi o primeiro a propor uma solução para o problema da comparação dos valores médios de duas populações gaussianas com variâncias possivelmente diferentes. Tal como outros que lhe sucederam, considerou a estatística T_{S_1,S_2} definida em (2.1).

O trabalho de Behrens (1929), porém, parece estar longe de ser perfeito; no dizer de Welch (1947a): "[...] *a solution which they ascribe initially to W. U. Behrens (1929). Looked at from one point of view, Behrens's paper appears to contain certain gross algebraical errors. Fisher and Jeffreys, however, develop lines of argument by means of which they claim that Behrens's solution is quite justified*".

No entanto, anos depois Fisher retomou a solução de Behrens, justificando-a com uma argumentação mais cuidada e publicando tabelas. A justificação de Fisher, porém, fazia intervir o seu conceito de "probabilidade fiducial", tão controverso como difícil de compreender.

Até certo ponto, estas primeiras soluções do problema de Behrens-Fisher, bem como as que se lhes seguiram e com elas competiram, apareceram num tempo em que os princípios da inferência estatística estavam ainda em génese (antes de Neyman e Pearson a terem "matematizado", como lamentaria mais tarde o prório Fisher), e eram por vezes usadas como equivalentes algumas noções que mais tarde vieram a ser separadas.

As polémicas em que foram envolvidas as soluções para o problema de Behrens-Fisher tiveram o mérito de contribuir para tornar claras as diferenças entre as várias abordagens da inferência estatística. Antes do seu aparecimento, as soluções produzidas pelas escolas frequencista, bayesiana e fiducialista para os problemas de inferência estatística, embora partindo de pressupostos diferentes e tendo interpretações distintas, eram formalmente idênticas. Isso tornava razoável considerar que todas eram formas alternativas de dizer o mesmo. Era comum, por exemplo, usar o termo "intervalo fiducial" com o sentido que hoje reservamos para "intervalo de confiança". Mas este problema veio mostrar que as várias abordagens não eram equivalentes, visto que as soluções produzidas pelos defensores das diversas escolas não coincidiam.

Tentemos então apresentar de forma simples a dedução de Fisher:

Dadas duas amostras aleatórias $\boldsymbol{X}_1 = \left(X_1, \ldots, X_{n_1}\right)$ e $\boldsymbol{X}_2 = \left(X_1, \ldots, X_{n_2}\right)$, extraídas de populações gaussianas com parâmetros $\mu_1,\ \sigma_1$ e $\mu_2,\ \sigma_2$, respectivamente, e denotando $\overline{X}_1 = \dfrac{1}{n_1}\sum_{k=1}^{n_1} X_{1k},\ S_1^2 = \dfrac{1}{n_1 - 1}\sum_{k=1}^{n_1}\left(X_{1k} - \overline{X}_1\right)^2,\ \overline{X}_2 = \dfrac{1}{n_2}\sum_{j=1}^{n_2} X_{2j},$ $S_2^2 = \dfrac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2$, como em populações gaussianas $\overline{X}$ e S^2 são independentes, a densidade conjunta das médias e variâncias empíricas é

$$\mathrm{d}F\left(\overline{x}_1, \overline{x}_2, s_1, s_2\right) \propto$$

$$\propto \frac{1}{\sigma_1 \sigma_2}\ \mathrm{e}^{-\frac{n_1}{2\sigma_1^2}\left(\overline{x}_1 - \mu_1\right)^2 - \frac{n_2}{2\sigma_2^2}\left(\overline{x}_2 - \mu_2\right)^2}\ \mathrm{d}\overline{x}_1\ \mathrm{d}\overline{x}_2 \times$$

$$\times \frac{s_1^{n_1-1}}{\sigma_1^{n_1-1}} \frac{s_2^{n_2-1}}{\sigma_2^{n_2-1}} \, \mathrm{e}^{-\frac{n_1-1}{2}\frac{s_1^2}{\sigma_1^2}-\frac{n_2-1}{2}\frac{s_2^2}{\sigma_2^2}} \frac{\mathrm{d}s_1}{s_1}\frac{\mathrm{d}s_2}{s_2}.$$

(Como $\frac{(n_k-1)\,S_k^2}{\sigma_k^2} \frown \chi^2_{n_k-1}$, e consequentemente se tem $S_k^2 \frown Gama\left(\frac{n_k-1}{2}, \frac{2\sigma_k^2}{n_k-1}\right)$, segue-se, com a substituição $y_k = s_k^2$, que

$$f_{S_k^2}(s_k^2) = \frac{(s_k^2)^{\frac{n_k-1}{2}-1} \, \mathrm{e}^{-\frac{n_k-1}{2}\frac{s_k^2}{\sigma_k^2}} 2s_k}{(\frac{2}{n_k-1}\sigma_k^2)^{\frac{n_k-1}{2}} \, \Gamma(\frac{n_k-1}{2})} \propto \left(\frac{s_k}{\sigma_k}\right)^{n_k-1} \mathrm{e}^{-\frac{n_k-1}{2}\frac{s_k^2}{\sigma_k^2}} \frac{1}{s_k},$$

para $k = 1, 2$)

O argumento fiducial leva a reinterpretar a expressão acima como distribuição fiduciária conjunta de μ_1, μ_2, σ_1 e σ_2, para $\overline{x}_1$, $\overline{x}_2$, s_1^2 e s_2^2 constantes (a proporcionalidade "absorve" as potências de s_1^2 e s_2^2). De acordo com este argumento, substituímos $\mathrm{d}\overline{x}_1$ e $\mathrm{d}\overline{x}_2$ por $\mathrm{d}\mu_1$ e $\mathrm{d}\mu_2$, $\frac{\mathrm{d}s_1}{s_1}$ e $\frac{\mathrm{d}s_2}{s_2}$ por $\frac{\mathrm{d}\sigma_1}{\sigma_1}$ e $\frac{\mathrm{d}\sigma_2}{\sigma_2}$.

$$\mathrm{d}F(\mu_1, \mu_2, \sigma_1, \sigma_2) \propto$$

$$\propto \frac{1}{\sigma_1^{n_1+1}\sigma_2^{n_2+1}} \, \mathrm{e}^{-\frac{n_1}{2\sigma_1^2}(\overline{x}_1-\mu_1)^2-\frac{n_2}{2\sigma_2^2}(\overline{x}_2-\mu_2)^2} \, \mathrm{d}\mu_1 \, \mathrm{d}\mu_2 \times$$

$$\times \, \mathrm{e}^{-\frac{n_1-1}{2}\frac{s_1^2}{\sigma_1^2}-\frac{n_2-1}{2}\frac{s_2^2}{\sigma_2^2}} \, \mathrm{d}\sigma_1 \, \mathrm{d}\sigma_2.$$

Então, fazendo a mudança de variável $\sigma_k^2 = \dfrac{1}{w}$ em

$$\int_0^{+\infty} \mathrm{e}^{-\frac{n_k}{2\sigma_k^2}(\overline{x}_k-\mu_k)^2-\frac{n_k-1}{2}\frac{s_k^2}{\sigma_k^2}} \frac{\mathrm{d}\sigma_k}{\sigma_k^{n_k+1}}$$

obtém-se

$$\int_0^{+\infty} w^{\frac{n_k+1}{2}} \mathrm{e}^{-w\left[\frac{n_k}{2}(\overline{x}_k-\mu_k)^2+\frac{n_k-1}{2}s_k^2\right]} \frac{1}{2} w^{-\frac{3}{2}} \mathrm{d}w =$$

$$\frac{1}{2}\left(\frac{1}{\frac{n_k}{2}(\overline{x}_k-\mu_k)^2+\frac{n_k-1}{2}s_k^2}\right)^{\frac{n_k}{2}} \int_0^{+\infty} y^{\frac{n_k}{2}-1}\mathrm{e}^{-y}\,\mathrm{d}y.$$

Escrevendo

$$t_1 = \frac{\mu_1-\overline{x}_1}{\frac{s_1}{\sqrt{n_1}}} \quad \text{e} \quad t_2 = \frac{\mu_2-\overline{x}_2}{\frac{s_2}{\sqrt{n_2}}}$$

obtém-se a função densidade de probabilidade fiduciária conjunta de μ_1 e de μ_2:

$$\mathrm{d}F(t_1,t_2) \propto \frac{\mathrm{d}t_1\,\mathrm{d}t_2}{\left[1+\frac{t_1^2}{n_1-1}\right]^{\frac{n_1}{2}}\left[1+\frac{t_2^2}{n_2-1}\right]^{\frac{n_2}{2}}}.$$

Notando que, se escrevermos $d = \overline{x}_1 - \overline{x}_2$ e $\delta = \mu_1 - \mu_2$,

$$(\mu_1-\overline{x}_1)-(\mu_2-\overline{x}_2) = \delta - d = t_1\frac{s_1}{\sqrt{n_1}} - t_2\frac{s_2}{\sqrt{n_2}}$$

estamos então em condições de perfazer inferência sobre $\delta = \mu_1 - \mu_2$, usando a distribuição fiduciária acima.

Fisher (1935a), inspirando-se em Behrens (1929), escolheu a estatística

$$\zeta = \frac{d - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

que exprime ζ como uma combinação linear conveniente de t_1 e de t_2,

$$\zeta = t_1 \cos\psi - t_2 \sin\psi \quad \text{onde} \quad \cos\,\psi = \frac{s_1}{\sqrt{n_1}} \text{ e } \sin\,\psi = \frac{s_2}{\sqrt{n_2}}.$$

Assim, como t_1 e t_2 têm distribuições t independentes, a distribuição de ζ é a de uma combinação linear de variáveis t de Student. Fisher, e antes dele Behrens, determinaram tabelas de quantis desta distribuição para os níveis de significância usuais e alguns valores de ψ. Fisher and Yates (1938) é a fonte mais acessível no que respeita consulta de tabelas para uso da estatística de Behrens-Fisher, mas consulte-se também Sukhatme (1938). Cochran propôs uma forma aproximada para os quantis críticos para cada nível α,

$$\widehat{d}_{1-\alpha} = \frac{\frac{s_1^2}{n_1}\, t_{n_1-1,\,1-\alpha} + \frac{s_2^2}{n_2}\, t_{n_2-1,\,1-\alpha}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

apresentando um estudo exaustivo da qualidade da aproximação em Cochran (1964), veja-se também McBean *et al.* (1988). Como ele próprio refere, esta aproximação é adequada para a o teste de Behrens-Fisher, mas não é adequada para o teste de Welch-Satterthwaite que se discute no capítulo seguinte, ainda que formalmente usem a mesma estatística de teste.

A fragilidade da abordagem de Fisher, como imediatamente Bartlett (1936) notou, é o papel especial, diferente do admitido em inferência clássica, que os valores observados s_1^2 e s_2^2 desempenham neste raciocínio.

Bartlett (1936) foi além disso provavelmente o primeiro a notar que o intervalo de confiança "fiducial" de Fisher não tem a mesma interpretação que os intervalos de confiança frequencistas. Se determinarmos para uma dada probabilidade $1-\alpha$ os quantis $\zeta_{\frac{\alpha}{2}}$ e $\zeta_{1-\frac{\alpha}{2}}$ correspondentes — e Fisher (1941) calculou e incentivou a tabulação de quantis, que apareceram em Fisher and Yates (1938) e Sukhatme (1938) — obtém-se um intervalo

$$\left(\overline{x}_1 - \overline{x}_2 - \zeta_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\, \overline{x}_1 - \overline{x}_2 + \zeta_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$$

que não contém o verdadeiro valor do parâmetro em $(1-\alpha)\times 100\%$ dos casos, a longo termo, como é óbvio observando que com $\theta = \frac{\sigma_1^2}{\sigma_2^2}$, $u = \frac{s_1^2}{s_2^2}$, $N = \frac{n_1}{n_2}$, $M = \frac{n_1-1}{n_2-1}$,

$$\zeta = t\sqrt{\frac{\frac{(n_1-1)\,s_1^2}{\sigma_1^2}+\frac{(n_2-1)\,s_2^2}{\sigma_2^2}}{n_1+n_2-2}}\,\frac{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = t\sqrt{\frac{\left(1+\frac{Mu}{\theta}\right)\left(1+\frac{\theta}{N}\right)}{(1+M)\left(1+\frac{u}{N}\right)}},$$

sendo

$$t = \frac{\frac{d-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{\frac{(n_1-1)\,s_1^2}{\sigma_1^2}+\frac{(n_2-1)\,s_2^2}{\sigma_2^2}}{n_1+n_2-2}}} \frown t_{n_1+n_2-2}\ .$$

De facto, por um lado,

$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2}}{n_1+n_2-2}=\frac{\frac{(n_2-1)S_2^2}{\sigma_2^2}\left[1+\frac{(n_1-1)\sigma_2^2S_1^2}{(n_2-1)\sigma_1^2S_2^2}\right]}{n_1+n_2-1}=$$

$$=\frac{\frac{(n_2-1)S_2^2}{\sigma_2^2}\left(1+\frac{Mu}{\theta}\right)}{(n_1-1)+(n_2-1)}=\frac{S_2^2\left(1+\frac{Mu}{\theta}\right)}{\sigma_2^2\,(1+M)}\,. \tag{2.2}$$

e por outro lado,

$$\frac{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}=\frac{\frac{\sigma_2^2}{n_2}\left(1+\frac{n_2\sigma_1^2}{n_1\sigma_2^2}\right)}{\frac{S_2^2}{n_2}\left(1+\frac{n_2S_1^2}{n_1S_2^2}\right)}=\frac{\sigma_2^2\left(1+\frac{\theta}{N}\right)}{S_2^2\left(1+\frac{u}{N}\right)} \tag{2.3}$$

Multiplicando (2.2) e (2.3) vem

$$\zeta=t\sqrt{\frac{\left(1+\frac{Mu}{\theta}\right)\left(1+\frac{\theta}{N}\right)}{(1+M)\left(1+\frac{u}{N}\right)}}.$$

Isto evidencia que a distribuição de ζ depende do quociente θ das variâncias desconhecidas, enquanto a distribuição de t não depende desse quociente.

Fisher começou por abordar o problema em 1935 (a, p. 396), envolvendo-se em polémica com Bartlett (1936, 1937) sobre a

sua "inferência fiducial" (Fisher, 1937, em especial p. 374), retomando a questão em 1939 e em 1941, e a ela voltando já no fim da sua carreira (Fisher and Healy, 1956). Fisher refere ainda no *Statistical Methods and Scientific Inference*, p. 97-100, que o ponto de discussão era menor e sem sentido, "[...] *but was eagerly seized upon by M. S. Bartlett, as though if it were a defect in the test of significance of a composite or comprehensive hypothesis, that in special cases the criterion of rejection is less frequently attained by chance than in others. On reflexion I do not think one should expect anything else, and it was perhaps only because at this time Bartlett was confidently putting forward an alternative attempt to solve the same problem that he made so much of a circumstance which is, indeed, generally to be expected.*".

As reedições de *Statistical Methods for Research Workers* foram incoporando os novos resultados sobre o problema, mas curiosamente sem nunca referirem o trabalho de Welch ou de Satterthwaite sobre a mesma questão. O próprio pensamento de Fisher sobre a relevância do problema parece ter evoluído muito: em *Statistical Methods for Research Workers* parece considerar a questão uma curiosidade matemática, sem reflexos efectivos em investigação experimental, enquanto em *Statistical Methods and Scientific Inference* já a classifica de potencialmente importante.

Lamentamos não saber discutir apropriadamente as questões filosóficas da inferência estatística, sentindo que com isso estamos a perder uma parte substancial da riqueza deste assunto, que alimentou quer a polémica bayesianos *vs.* frequencistas, quer a polémica fiducialistas *vs.* bayesianos e fiducialistas *vs.* frequencistas. Remetemos para Murteira (1988), ainda que o cerne do livro não seja o pensamento fiducial.

Fisher repudiou de uma forma inequívoca as ideias bayesia-

nas, mas claramente, na sua reexposição do "argumento fiducial" que faz em *Statistical Methods and Scientific Inference*, p. 54-60, admite que do ponto de vista fiducial o parâmetro ganha uma distribuição de probabilidade quando se recolhe uma amostra: "*By contrast, the fiducial argument uses the observations only to change the logical status of the parameter from one in which nothing is known of it, and no probability statement about it can be made, to the status of a random variable having a well-defined distribution.*"

2.1.2 A dedução de Jeffreys

Curiosamente, Jeffreys (1940; 1983, p. 280) resolveu de forma elegante o problema de Behrens-Fisher por métodos bayesianos, admitindo distribuições *a priori* para os parâmetros. Começa por postular — o que é natural, mas não trivial — que

$$p\left(\mathrm{d}t_1, \mathrm{d}t_2 \mid \overline{x}_1, \overline{x}_2, s_1, s_2, H\right) = f_1(t_1)\, f_2(t_2)\ \mathrm{d}t_1 \mathrm{d}t_2$$

e que a densidade *a priori* de $\mathrm{d}\mu_1 \mathrm{d}\mu_2 \mathrm{d}\sigma_1 \mathrm{d}\sigma_2 \mid H$ é

$$p\left(\mathrm{d}\mu_1 \mathrm{d}\mu_2 \mathrm{d}\sigma_1 \mathrm{d}\sigma_2 \mid H\right) \propto \frac{\mathrm{d}\mu_1 \mathrm{d}\mu_2 \mathrm{d}\sigma_1 \mathrm{d}\sigma_2}{\sigma_1 \sigma_2}\,.$$

Por outro lado, denotando por D os dados, a verosimilhança é

$$p\left(D \mid \mu_1, \mu_2, \sigma_1, \sigma_2, H\right) \propto \frac{1}{\sigma_1^{n_1} \sigma_2^{n_2}} \times$$

$$\times\ \mathrm{e}^{-\frac{1}{2\sigma_1^2}\left\{n_1\left(\mu_1 - \overline{x}_1\right)^2 + (n_1 - 1)s_1^2\right\} - \frac{1}{2\sigma_2^2}\left\{n_2\left(\mu_2 - \overline{x}_2\right)^2 + (n_2 - 1)s_2^2\right\}}\,.$$

Consequentemente,

$$p\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid D, H\right) = \frac{1}{\sigma_1^{n_1+1}\sigma_2^{n_2+1}} \times$$

$$\times \ \mathrm{e}^{-\frac{1}{2\sigma_1^2}\left\{n_1\left(\mu_1-\overline{x}_1\right)^2+(n_1-1)s_1^2\right\}-\frac{1}{2\sigma_2^2}\left\{n_2\left(\mu_2-\overline{x}_2\right)^2+(n_2-1)s_2^2\right\}}.$$

Integrando em ordem a σ_1 e a σ_2 para obter a distribuição *a posteriori*, reencontra-se a expressão deduzida por Fisher (1935a). As críticas de Bartlett (1936) poderiam ser tão certeiramente dirigidas à solução de Jeffreys quanto à de Fisher. É de crer que nenhum dos autores se tenha preocupado com isso, uma vez que o objectivo não era encontrar intervalos de confiança que admitissem uma interpretação frequencista.

A solução frequencista do problema das duas médias só viria a ser deduzida por Welch (1938, 1947a, 1947b, 1956a), que sempre comparou cuidadosamente os resultados do seu trabalho com os resultados de Fisher, naturalmente por ter ideias diferentes sobre a inferência estatística. Que o problema de Behrens-Fisher está no âmago das discordância entre as várias escolas é patente no seu trabalho mais centrado sobre a filosofia da inferência (Welch, 1939). A sua discordância em relação ao argumento fiducial de Fisher tenta-o a uma rejeição em bloco. Citamos, de *The generalization of "Student's" problem when several different population variances are involved* (Welch, 1947a, p. 34):

> *We may, however, permit ourselves one observation about the development according to Fisher and Jeffreys.*

Both these writers, at some stage, limit the field of their probability inferences to a subset in which the s_i^2 *are regarded as fixed. In order to solve the problem on these lines Jeffreys introduces an a priori distribution function for the unknown* σ_i*, following his general philosophy for dealing with such questions. Fisher, on the other hand, arrives at the same answer by a special utilization of what he terms the fiducial distribution of* σ_i*.*

Jeffreys approach here does not raise any new issues to those who are familiar with the general body of his researches on statistical inference. Fisher's justification of Behrens's solution is perhaps of more immediate interest as it raises controversial points which are important more specifically in relation to our present topic of discussion. For although Fisher's approach has been very much criticized by a number of writers, starting with M. S. Bartlett (1936), the critics have not wished to throw doubt on the whole body of results which Fisher includes under the heading of fiducial inference. The criticism has been for the most part selective, directed mainly at the way in which so-called simultaneous fiducial distributions of several parameters have been defined and manipulated.

I have, myself, quite definitive views on these questions (particularly on the usage of the word 'fiducial') but do not feel that I need express them at any great length here. I disagree with Fisher, but this divergence of opinion must already have become apparent in the way I have defined the field within which

> *I made my probability inferences about* η*. It appears to me to be quite artificial to restrict our view to one which, even in a limited sense, fixes* s_i^2*. It is true that, in the two sample problem, we have to draw our inferences from the unique pair of samples observed, or, more precisely, from the statistics* $\overline{x}_1$*,* $\overline{x}_2$*,* s_1^2 *and* s_2^2 *which they provide. These statistics are the only data for the purpose of making inferences, but we add something to these data in the interpretation when we regard the samples as being drawn randomly from hypothetical normal population. Once having embarked on this method of interpretation, we should stick to it consistently throughout. The sampling variations of the* s_i^2 *should be taken into account only by a direct use of the probability distributions as given in our equation* (1)*, and not by any inversion such as is involved in Fisher's conception of the fiducial distribution of* σ_i^2*. As we have seen, it is quite possible to make probability statements about the difference between the population means without making any reference whatever either to inverse probability or to fiducial distributions.*

Fisher, com o bom feitio com que Nosso Senhor o dotou, e pelo qual merece o maior crédito, ignorou os trabalhos de Welch, até na verrinosa polémica de 1956 publicada no *J. Royal Statist. Soc.*

Porventura a opção concordante com a opinião, que manteve até ao fim, de que a sua abordagem era a única válida.

2.2 A razão de verosimilhanças no caso de heterogeneidade de variâncias

Uma das causas das controvérsias acerca do problema de Behrens-Fisher é sem dúvida o facto de não serem aplicáveis critérios de optimalidade que definam um teste que possa ser considerado "melhor".

Atrás deduzimos o teste da razão de verosimilhanças sob homocedasticidade. Vamos agora observar as dificuldades que traz considerar que as variâncias são diferentes. Começamos pelo caso em que a razão entre as variâncias não é conhecida, e abordamos seguidamente a situação em que essa razão é conhecida.

A razão de verosimilhanças na investigação da igualdade de valores médios

Sejam $\boldsymbol{x}_1 = \left(x_{11}, \ldots, x_{1n_1}\right)$, $\boldsymbol{x}_2 = \left(x_{21}, \ldots, x_{2n_2}\right)$ realizações de duas amostras independentes, de $X_1 \frown \mathit{Gaussiana}\,(\mu_1, \sigma_1)$ e de $X_2 \frown \mathit{Gaussiana}\,(\mu_2, \sigma_2)$, respectivamente. Desejamos testar

$$H_0 : \mu_2 - \mu_1 = \Delta \quad vs. \quad H_1 : \mu_2 - \mu_1 \neq \Delta.$$

Consideremos sem perda de generalidade $\Delta = 0$ e $\mu = \mu_1 = \mu_2$ o valor médio comum sob H_0. A razão de verosimilhanças é

$$\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) = \frac{\sup\limits_{\mu_1 = \mu_2} L\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right)}{\sup L\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right)}.$$

O máximo absoluto da verosimilhança, de acordo com o Teorema 1.3.2 é atingido para

$$\hat{\mu}_1 = \overline{x}_1 \qquad\qquad \hat{\mu}_2 = \overline{x}_2$$

$$\hat{\sigma}_1^{\,2} = \tfrac{1}{n_1} \sum_{k=1}^{n_1} \left(x_{1k} - \overline{x}_1\right)^2 \qquad \hat{\sigma}_2^{\,2} = \tfrac{1}{n_2} \sum_{j=1}^{n_2} \left(x_{2j} - \overline{x}_2\right)^2$$

sendo $\sup L\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) = \left(2\pi\, \mathrm{e}\, \widehat{\sigma}_1^{\,2}\right)^{-\frac{n_1}{2}} \left(2\pi\, \mathrm{e}\, \widehat{\sigma}_2^{\,2}\right)^{-\frac{n_2}{2}}$.

Encontrar o máximo da verosimilhança restrita a H_0 equivale a maximizar a função de suporte

$$f\left(\mu, \sigma_1, \sigma_2\right) = \ln\, L\left(\mu, \mu, \sigma_1, \sigma_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) =$$

$$= -n_1 \ln(\sqrt{2\pi}\sigma_1) - n_2 \ln(\sqrt{2\pi}\sigma_2) - \\ - \frac{1}{2\sigma_1^2} \sum_{k=1}^{n_1} \left(x_{1k} - \mu\right)^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{n_2} \left(x_{2j} - \mu\right)^2.$$

Igualando a zero as derivadas parciais

$$\frac{\partial f}{\partial \sigma_1} = -\frac{n_1}{\sigma_1} + \frac{1}{\sigma_1^3} \sum_{k=1}^{n_1} \left(x_{1k} - \mu\right)^2,$$

$$\frac{\partial f}{\partial \sigma_2} = -\frac{n_2}{\sigma_2} + \frac{1}{\sigma_2^3} \sum_{j=1}^{n_2} \left(x_{2j} - \mu\right)^2,$$

$$\frac{\partial f}{\partial \mu} = \frac{1}{\sigma_1^2} \sum_{k=1}^{n_1} \left(x_{1k} - \mu\right) + \frac{1}{\sigma_2^2} \sum_{j=1}^{n_2} \left(x_{2j} - \mu\right) =$$

$$= \frac{n_1(\overline{x}_1 - \mu)}{\sigma_1^2} + \frac{n_2(\overline{x}_2 - \mu)}{\sigma_2^2},$$

obtêm-se as condições

$$\sigma_1^2 = \frac{1}{n_1}\sum_{k=1}^{n_1}(x_{1k} - \mu)^2 = \widehat{\sigma}_1^2 + (\overline{x}_1 - \mu)^2 \qquad (2.4)$$

$$\sigma_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(x_{2j} - \mu)^2 = \widehat{\sigma}_2^2 + (\overline{x}_2 - \mu)^2 \qquad (2.5)$$

$$\frac{n_1(\overline{x}_1 - \mu)}{\widehat{\sigma}_1^2 + (\overline{x}_1 - \mu)^2} + \frac{n_2(\overline{x}_2 - \mu)}{\widehat{\sigma}_2^2 + (\overline{x}_2 - \mu)^2} = 0$$

A última destas é uma equação do terceiro grau em μ que se pode escrever na forma $\mu^3 - a\mu^2 + b\mu - c = 0$ com $a, b, c \geq 0$. É complicada de resolver algebricamente, mas a solução pode ser aproximada por métodos numéricos. De (2.2) obtém-se também

$$\mu = \frac{n_1\overline{x}_1\sigma_2^2 + n_2\overline{x}_2\sigma_1^2}{n_1\sigma_2^2 + n_2\sigma_1^2}, \qquad (2.6)$$

que sugere de imediato uma iteração para resolver o sistema de estacionaridade:

0. Atribuir valores iniciais às variâncias, por exemplo as estimativas irrestritas $\sigma_1^{2(0)} = \widehat{\sigma}_1^2$ e $\sigma_2^{2(0)} = \widehat{\sigma}_2^2$.

1. Estimar o valor médio $\mu^{(i)}$ usando a equação (2.6).

2. Actualizar as estimativas das variâncias, $\sigma_1^{2(i+1)}$ e $\sigma_2^{2(i+1)}$ usando as equações (2.4) e (2.5), e voltar ao passo 1 até atingir a precisão desejada.

Designando por $(\widehat{\mu}_0, \widehat{\sigma}_{01}, \widehat{\sigma}_{02})$ o maximizante da verosimilhança restrita a H_0, o máximo é $L\left(\widehat{\mu}_0, \widehat{\mu}_0, \widehat{\sigma}_{01}, \widehat{\sigma}_{02} \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) = \left(2\pi\,\mathrm{e}\,\widehat{\sigma}_{01}^2\right)^{-\frac{n_1}{2}} \left(2\pi\,\mathrm{e}\,\widehat{\sigma}_{02}^2\right)^{-\frac{n_2}{2}}$ e a razão de verosimilhanças

$$\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) = \left[1 + \left(\frac{\overline{x}_1 - \widehat{\mu}_0}{\widehat{\sigma}_1}\right)^2\right]^{-\frac{n_1}{2}} \left[1 + \left(\frac{\overline{x}_2 - \widehat{\mu}_0}{\widehat{\sigma}_2}\right)^2\right]^{-\frac{n_2}{2}}.$$

A região crítica corresponde então ao conjunto dos pares de amostras $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ tais que $\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) \leq \delta_0$, mas não parece possível tentar extrair daqui uma estatística de teste univariada como se faz no caso de homogeneidade de variâncias. No entanto, nas condições do teorema de Wilks a própria razão de verosimilhanças pode servir de estatística de teste: assintoticamente, $-2\ln\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) = \sum n_i \ln\widehat{\sigma}_{0i}^2 - \sum n_i \ln\widehat{\sigma}_i^2 \simeq \chi_\nu^2$, com $\nu = \dim\Theta - \dim\Theta_0 = 1$.

A busca de algoritmos de maximização da verosimilhança tem atraído bastante interesse na última década. O algoritmo simples que esboçamos acima não é necessariamente o mais eficiente. Drton and Eichler (2006) mostraram que converge para um dos pontos de estacionaridade, mas que pode não ser o máximo absoluto. Belloni and Didier (2008) desenvolveram um algoritmo com convergência garantida para o máximo absoluto maximizando directamente a log-verosimilhança em vez de partirem do sistema de estacionaridade.

Com efeito, a verosimilhança restrita a H_0 no problema de Behrens-Fisher pode ter mais de um ponto de estacionaridade. Bozdogan and Ramirez (1987) dão exemplos e notam que o sistema de estacionaridade pode ser instável, embora acrescentem que o nível de significância do teste da razão de verosimilhanças se mantém controlado.

Sugiura and Gupta(1987) provaram que (quase certamente) a equação cúbica para $\widehat{\mu}_0$ ou tem uma única solução real ou três, correspondendo o segundo caso a dois máximos locais da verosimilhança perto das médias amostrais com um mínimo local entre eles. As raízes múltiplas tendem a ocorrer quando a diferença entre as médias amostrais é grande comparada com as variâncias. Quando a diferença entre as médias amostrais é pequena ou as amostras são grandes a solução do sistema tende a ser única (Drton, 2008) e estável (Bozdogan and Ramirez, 1987). Simulações sugerem que a situação mais comum sob H_0 é a raiz ser única, e a probabilidade de ocorrerem raízes múltiplas é positiva mas pequena (Sugiura and Gupta, 1987). Sundberg (2010) argumenta que a ocorrência de pontos de estacionaridade múltiplos na verosimilhança se deve tomar como sinal de inadequação do modelo.

Buot *et al.* (2007) generalizaram estes resultados ao caso multivariado e recordam, citando Linnik (1967), que o problema de Behrens-Fisher foi dos primeiros exemplos de um membro da família exponencial com estatísticas suficientes que, no espaço dos parâmetros restrito à hipótese nula, não são completas.

O teste da razão de verosimilhanças quando a razão entre as variâncias é conhecida

Ainda é possível, apesar disso, encontrar uma estatística de teste com distribuição t de Student para a comparação de populações gaussianas independentes com variâncias desconhecidas e diferentes, se a razão entre elas for conhecida.

Vamos abordar o mesmo problema, mas supondo agora que as variâncias são desconhecidas, mas que a sua razão $\theta = \frac{\sigma_1^2}{\sigma_2^2}$ é conhecida (de que o caso de homocedasticidade, $\theta = 1$, é apenas um caso particular).

Teorema 2.2.1.
Sejam $\boldsymbol{X}_1 = \left(X_1, \ldots, X_{n_1}\right)$ *e* $\boldsymbol{X}_2 = \left(X_1, \ldots, X_{n_2}\right)$ *amostras aleatórias independentes, com* $X_{1k} \frown \mathit{Gaussiana}\,(\mu_1, \sigma_1)$, $k = 1, \ldots, n_1$ $X_{2j} \frown \mathit{Gaussiana}\,(\mu_2, \sigma_2)$, $j = 1, \ldots, n_2$, *em que* σ_1^2 *e* σ_2^2 *são desconhecidas, mas a razão entre as variâncias* $\theta = \frac{\sigma_1^2}{\sigma_2^2}$ *é conhecida.*

Então a verosimilhança é máxima para

$$\widehat{\mu}_1 = \overline{x}_1, \quad \widehat{\mu}_2 = \overline{x}_2,$$

$$\widehat{\sigma}_1^2 = \frac{\sum_{k=1}^{n_1} \left(x_{1k} - \overline{x}_1\right)^2 + \theta \sum_{j=1}^{n_2} \left(x_{2j} - \overline{x}_2\right)^2}{n_1 + n_2} = \frac{Q\left(\overline{x}_1, \overline{x}_2\right)}{n_1 + n_2}$$

sendo o seu máximo

$$L(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = (2\pi \mathrm{e})^{-\frac{n_1+n_2}{2}}\, \theta^{\frac{n_2}{2}} \left(\frac{n_1 + n_2 - 2}{n_1 + n_2}\, s_\theta^2\right)^{-\frac{n_1+n_2}{2}},$$

onde

$$S_{\theta}^{2} = \frac{(n_1 - 1)S_1^2 + \theta(n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

é um estimador ponderado centrado de σ_1^2.

Demonstração:

Com a condição $\theta = \frac{\sigma_1^2}{\sigma_2^2} \iff \sigma_2^2 = \frac{1}{\theta}\,\sigma_1^2$, a verosimilhança é

$$L(\mu_1, \mu_2, \sigma_1 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\theta^{\frac{n_2}{2}}}{\left(\sqrt{2\pi}\sigma_1\right)^{n_1+n_2}} \times$$
$$\times \exp\left\{-\frac{1}{2\sigma_1^2}\left[\sum_{k=1}^{n_1}\left(x_{1k} - \mu_1\right)^2 + \theta \sum_{j=1}^{n_2}\left(x_{2j} - \mu_2\right)^2\right]\right\}$$

donde

$$\ln L(\mu_1, \mu_2, \sigma_1 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{n_2}{2}\,\ln\,\theta - (n_1 + n_2)\ln\,\sqrt{2\pi} -$$
$$-\,(n_1 + n_2)\,\ln\,\sigma_1 - \frac{1}{2\sigma_1^2}\left(\sum_{k=1}^{n_1}\left(x_{1k} - \mu_1\right)^2 + \theta \sum_{j=1}^{n_2}\left(x_{2j} - \mu_2\right)^2\right).$$

Podemos então concluir que

$$\frac{\partial}{\partial \mu_1}\ln\,L(\mu_1, \mu_2, \sigma_1 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = 0 \Longrightarrow \widehat{\mu}_1 = \overline{x}_1;$$

$$\frac{\partial}{\partial \mu_2}\ln\,L(\mu_1, \mu_2, \sigma_1 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = 0 \Longrightarrow \widehat{\mu}_2 = \overline{x}_2;$$

$$\frac{\partial}{\partial\sigma_1}\ln L(\mu_1,\mu_2,\sigma_1 \mid \boldsymbol{x}_1,\boldsymbol{x}_2) =$$

$$= -\frac{n_1+n_2}{\sigma_1} + \frac{1}{\sigma_1^3}\left[\sum_{k=1}^{n_1}\left(x_{1k}-\mu_1\right)^2 + \theta\sum_{j=1}^{n_2}\left(x_{2j}-\mu_2\right)^2\right] =$$

$$= -\frac{n_1+n_2}{\sigma_1} + \frac{1}{\sigma_1^3}Q\left(\mu_1,\mu_2\right)$$

donde

$$\frac{\partial}{\partial\sigma_1}\ln L(\mu_1,\mu_2,\sigma_1 \mid \boldsymbol{x}_1,\boldsymbol{x}_2) = 0 \Longrightarrow \widehat{\sigma}_1^2 = \frac{Q\left(\hat{\mu}_1,\hat{\mu}_2\right)}{n_1+n_2}.$$

É de notar que

$$\frac{Q\left(\overline{x}_1,\overline{x}_2\right)}{n_1+n_2} = \frac{n_1+n_2-2}{n_1+n_2}\,\frac{(n_1-1)s_1^2+\theta(n_2-1)s_2^2}{n_1+n_2-2} = \frac{n_1+n_2-2}{n_1+n_2}\,s_\theta^2,$$

e é fácil ver que S_θ^2 é um estimador centrado de σ_1^2.

Por outro lado, substituindo,

$$L(\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1 \mid \boldsymbol{x}_1,\boldsymbol{x}_2) =$$

$$= \frac{\theta^{\frac{n_2}{2}}}{\left(\sqrt{2\pi}\right)^{n_1+n_2}}\left[\frac{Q(\overline{x}_1,\overline{x}_2)}{n_1+n_2}\right]^{-\frac{n_1+n_2}{2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{Q(\overline{x}_1,\overline{x}_2)}{n_1+n_2}\right)^{-1}Q(\overline{x}_1,\overline{x}_2)} =$$

$$= (2\pi\mathrm{e})^{-\frac{n_1+n_2}{2}}\,\theta^{\frac{n_2}{2}}\left[\frac{Q(\overline{x}_1,\overline{x}_2)}{n_1+n_2}\right]^{-\frac{n_1+n_2}{2}}. \qquad \square$$

Teorema 2.2.2.

Sejam $\boldsymbol{X}_1 = \left(X_1, \ldots, X_{n_1}\right)$ *e* $\boldsymbol{X}_2 = \left(X_1, \ldots, X_{n_2}\right)$ *amostras aleatórias independentes, com* $X_{1k} \frown \mathit{Gaussiana}\,(\mu_1, \sigma_1)$, $k = 1, \ldots, n_1$ $X_{2j} \frown \mathit{Gaussiana}\,(\mu_2, \sigma_2)$, $j = 1, \ldots, n_2$, *em que* σ_1^2 *e* σ_2^2 *são desconhecidas, mas a razão entre as variâncias* $\theta = \frac{\sigma_1^2}{\sigma_2^2}$ *é conhecida. Se além disso exigirmos* $\mu_2 - \mu_1 = \Delta$, *a verosimilhança é máxima para*

$$\widehat{\mu}_1 = \frac{n_1 \overline{x}_1 + \theta n_2 (\overline{x}_2 - \Delta)}{n_1 + \theta n_2},$$

$$\widehat{\mu}_2 = \frac{n_1 (\overline{x}_1 + \Delta) + \theta n_2 \overline{x}_2}{n_1 + \theta n_2},$$

$$\widehat{\sigma}_\theta^2 = \frac{Q\left(\widehat{\mu}_1, \widehat{\mu}_2\right)}{n_1 + n_2} = \frac{\sum_{k=1}^{n_1} \left(x_{1k} - \widehat{\mu}_1\right)^2 + \theta \sum_{j=1}^{n_2} \left(x_{2j} - \widehat{\mu}_2\right)^2}{n_1 + n_2}$$

sendo o seu máximo

$$(2\pi \mathrm{e})^{-\frac{n_1+n_2}{2}}\, \theta^{\frac{n_2}{2}}\, \widehat{\sigma}_\theta^{-\frac{n_1+n_2}{2}}.$$

Demonstração:

Queremos maximizar

$$f(\mu_1, \mu_2, \sigma_1) = -(n_1 + n_2) \ln \sqrt{2\pi} + \frac{n_2}{2} \ln \theta - (n_1 + n_2) \ln \sigma_1 -$$

$$- \frac{1}{2\sigma_1^2} \left[\sum_{k=1}^{n_1} \left(x_{1k} - \mu_1\right)^2 + \sum_{j=1}^{n_2} \left(x_{2j} - \mu_2\right)^2 \right]$$

com a restrição $\mu_2 - \mu_1 = \Delta \iff g(\mu_1, \mu_2, \sigma_1) = 0$.

Usando o método dos multiplicadores de Lagrange,

$$\nabla f(\mu_1, \mu_2, \sigma_1) = \lambda \nabla g(\mu_1, \mu_2, \sigma_1), \qquad \nabla g = (-1, 1, 0),$$

donde

$$\frac{1}{\sigma_1^2} \sum_{k=1}^{n_1} (x_{1k} - \mu_1) = -\lambda, \qquad \frac{\theta}{\sigma_1^2} \sum_{j=1}^{n_2} \left(x_{2j} - \mu_2\right) = \lambda,$$

segue-se que

$$\sum_{k=1}^{n_1} (x_{1k} - \mu_1) = -\theta \sum_{j=1}^{n_2} \left(x_{2j} - \mu_2\right) \iff n_1 \overline{x}_1 - n_1 \mu_1 = \theta(n_2 \mu_2 - n_2 \overline{x}_2).$$

Como $\mu_2 - \mu_1 = \Delta \iff \mu_2 = \mu_1 + \Delta$, substitui-se, obtendo

$$n_1 \overline{x}_1 - n_1 \mu_1 = \theta(n_2 \mu_1 + n_2 \Delta - n_2 \overline{x}_2) \iff$$

$$\iff (\theta n_2 + n_1)\mu_1 = n_1 \overline{x}_1 - \theta n_2 (\Delta - \overline{x}_2)$$

e assim

$$\hat{\mu}_1 = \frac{n_1 \overline{x}_1 + \theta n_2 (\overline{x}_2 - \Delta)}{n_1 + \theta n_2}\,.$$

Por outro lado, de $\mu_1 = \mu_2 - \Delta$,

$$n_1 \overline{x}_1 - n_1(\mu_2 - \Delta) = \theta(n_2 \mu_2 - n_2 \overline{x}_2) \iff (\theta n_2 + n_1)\mu_2 = n_1(\overline{x}_1 + \Delta) + \theta n_2 \overline{x}_2$$

e assim

$$\hat{\mu}_2 = \frac{n_1 (\overline{x}_1 + \Delta) + \theta n_2 \overline{x}_2}{n_1 + \theta n_2}\,.$$

Quanto a σ_1, é imediato que

$$\hat{\sigma}_1^{\,2} = \frac{Q(\widehat{\mu}_1, \widehat{\mu}_2)}{n_1 + n_2}\,.$$

Portanto,

$$\sup_{\mu_2-\mu_1=\Delta} L(\mu_1,\mu_2,\sigma_1 \mid \boldsymbol{x}_1,\boldsymbol{x}_2) =$$

$$= (2\pi)^{-\frac{n_1+n_2}{2}}\ \theta^{\frac{n_2}{2}}\ \widehat{\sigma}_\theta^{-\frac{n_1+n_2}{2}}\ \exp\left[-\frac{1}{2\widehat{\sigma}_\theta^2}Q(\widehat{\mu}_1,\widehat{\mu}_2)\right] =$$

$$= (2\pi)^{-\frac{n_1+n_2}{2}}\ \theta^{\frac{n_2}{2}}\ \widehat{\sigma}_\theta^{-\frac{n_1+n_2}{2}}\ \exp\left[-\frac{n_1+n_2}{2\,Q(\widehat{\mu}_1,\widehat{\mu}_2)}Q(\widehat{\mu}_1,\widehat{\mu}_2)\right] =$$

$$= (2\pi\mathrm{e})^{-\frac{n_1+n_2}{2}}\ \theta^{\frac{n_2}{2}}\ \widehat{\sigma}_\theta^{-\frac{n_1+n_2}{2}}$$

é o máximo da verosimilhança naquelas condições.

□

Teorema 2.2.3.

Sejam $\boldsymbol{X}_1 = \left(X_1,\ldots,X_{n_1}\right)$ *e* $\boldsymbol{X}_2 = \left(X_1,\ldots,X_{n_2}\right)$ *amostras aleatórias independentes, com* $X_{1k} \frown \mathit{Gaussiana}\,(\mu_1,\sigma_1)$, $k = 1,\ldots,n_1$, $X_{2j} \frown \mathit{Gaussiana}\,(\mu_2,\sigma_2)$, $j = 1,\ldots,n_2$, *em que* σ_1^2 *e* σ_2^2 *são desconhecidas, mas a razão entre as variâncias* $\theta = \frac{\sigma_1^2}{\sigma_2^2}$ *é conhecida.*

A razão de verosimilhanças para testar

$$H_0 : \mu_2 - \mu_1 = \Delta \quad vs. \quad H_1 : \mu_2 - \mu_1 \neq \Delta$$

é

$$\Lambda(\boldsymbol{x}_1,\boldsymbol{x}_2) = \left[1 + \frac{1}{n_1+n_2-2}\left(\frac{\overline{x}_2-\overline{x}_1-\Delta}{s_\theta\sqrt{\frac{1}{n_1}+\frac{1}{\theta n_2}}}\right)^2\right]^{-\frac{n_1+n_2}{2}}.$$

Demonstração:

Comecemos por estabelecer que

$$\widehat{\sigma}_{\theta}^{2} = \frac{n_1 + n_2 - 2}{n_1 + n_2}\, s_{\theta}^{2} + \theta\, \frac{n_1 n_2}{n_1 + n_2}\, \frac{\left(\overline{x}_2 - \overline{x}_1 - \Delta\right)^2}{n_1 + \theta n_2}\,.$$

Com efeito

$$\widehat{\sigma}_{\theta}^{2} = \frac{Q(\widehat{\mu}_1, \widehat{\mu}_2)}{n_1 + n_2} =$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{k=1}^{n_1}\left(x_{1k} - \overline{x}_1 + \overline{x}_1 - \frac{n_1 + \theta n_2(\overline{x}_2 - \Delta)}{n_1 + \theta n_2}\right)^2 + \right.$$

$$\left. + \theta \sum_{j=1}^{n_2}\left(x_{2j} - \overline{x}_2 + \overline{x}_2 - \frac{n_1(\overline{x}_1 + \Delta) + \theta n_2 \overline{x}_2}{n_1 + \theta n_2}\right)^2\right] =$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{k=1}^{n_1}\left(x_{1k} - \overline{x}_1\right)^2 + \theta \sum_{j=1}^{n_2}\left(x_{2j} - \overline{x}_2\right)^2\right] +$$

$$+ \frac{n_1}{n_1 + n_2}\left[\frac{\theta n_2(\overline{x}_1 - \overline{x}_2 + \Delta)}{n_1 + \theta n_2}\right]^2 + \frac{\theta n_2}{n_1 + n_2}\left[\frac{n_1\left(\overline{x}_2 - \overline{x}_1 - \Delta\right)}{n_1 + \theta n_2}\right]^2 =$$

$$= \frac{n_1 + n_2 - 2}{n_1 + n_2}\, s_{\theta}^{2} + \theta\, \frac{n_1 n_2^{2}\theta + n_1^{2} n_2}{n_1 + n_2}\left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1 + \theta n_2}\right)^2 =$$

$$= \frac{n_1+n_2-2}{n_1+n_2}\, s_\theta^{\,2} + \theta\, \frac{n_1 n_2 (n_1+\theta n_2)}{n_1+n_2} \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{n_1+\theta n_2} \right)^2 =$$

$$= \frac{n_1+n_2-2}{n_1+n_2}\, s_\theta^{\,2} + \theta\, \frac{n_1 n_2}{n_1+n_2}\, \frac{(\overline{x}_2 - \overline{x}_1 - \Delta)^2}{n_1+\theta n_2}\,.$$

Portanto

$$\Lambda(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[1 + \theta\, \frac{n_1 n_2}{n_1+n_2-2}\, \frac{(\overline{x}_2 - \overline{x}_1 - \Delta)^2}{s_\theta^2 (n_1+\theta n_2)} \right]^{-\frac{n_1+n_2}{2}} =$$

$$= \left[1 + \frac{1}{n_1+n_2-2}\, \frac{(\overline{x}_2 - \overline{x}_1 - \Delta)^2}{s_\theta^2 \left(\frac{1}{\theta n_2} + \frac{1}{n_1} \right)} \right]^{-\frac{n_1+n_2}{2}} =$$

$$= \left[1 + \frac{1}{n_1+n_2-2} \left(\frac{\overline{x}_2 - \overline{x}_1 - \Delta}{s_\theta \sqrt{\frac{1}{n_1} + \frac{1}{\theta n_2}}} \right)^2 \right]^{-\frac{n_1+n_2}{2}} . \quad \square$$

Note-se que

$$S_\theta^2 \left(\frac{1}{n_1} + \frac{1}{\theta n_2} \right)$$

é um estimador centrado da variância de $\overline{X}_1 - \overline{X}_2$.

De facto, $\sigma_2^2 = \frac{1}{\theta}\, \sigma_1^2$, donde

$$\operatorname{var}\left[\overline{X}_1 - \overline{X}_2 \right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{\theta n_2} = \sigma_1^2 \left(\frac{1}{n_1} + \frac{1}{\theta n_2} \right).$$

Fica também claro que o que simplifica o problema neste caso particular é que a variância de $\overline{X}_1 - \overline{X}_2$ pode ser escrita como função apenas da variância de uma das gaussianas (neste caso σ_1^2) — e por isso pode ser estimada com base em um único estimador ponderado, cuja distribuição é facilmente reportável ao qui-quadrado.

Com efeito, $\frac{(n_1+n_2-2)\,S_\theta^2}{\sigma_1^2}$ é um qui-quadrado com n_1+n_2-2 graus de liberdade, e por essa razão

$$T_\theta^* = \frac{\dfrac{\overline{X}_2 - \overline{X}_1 - (\mu_2-\mu_1)}{\sigma_1\sqrt{\frac{1}{n_1}+\frac{1}{\theta n_2}}}}{\sqrt{\dfrac{\frac{(n_1+n_2-2)\,S_\theta^2}{\sigma_1^2}}{n_1+n_2-2}}} = \frac{\overline{X}_2 - \overline{X}_1 - (\mu_2-\mu_1)}{S_\theta\sqrt{\frac{1}{n_1}+\frac{1}{\theta n_2}}}$$

tem distribuição t de Student com n_1+n_2-2 graus de liberdade. Note-se ainda que S_θ^2 estima apenas uma das variâncias, como faz sentido nas condições do problema.

2.3 Welch e Satterthwaite, e a solução frequencista

No Capítulo 3 ficou de certo modo claro que Welch propôs uma dedução alternativa ao problema de Behrens-Fisher, concordando com Bartlett (1936) na crítica a Fisher por este admitir uma distribuição fiducial para os σ_i^2. A solução a que chega é mais geral, e não coincide com a de Fisher (1935a, 1939b, 1941),

como o próprio Welch comenta no seu trabalho de 1951, em que explicitamente aponta termos diferentes na expansão de Fisher e na sua.

Além disso — o que não deixa de ser um pouco perturbador — o tratamento de Welch leva a um teste t para duas amostras com um número de graus de liberdade fraccionário, que mesmo que $\sigma_1^2 = \sigma_2^2$ não coincide com o número de graus de liberdade $n_1 + n_2 - 2$ do tratamento clássico. De facto, no teste de Welch, o número de graus de liberdade é $n_1 + n_2 - 2$ se e só se os erros padrões das duas amostras forem iguais, e portanto no caso de variâncias populacionais iguais o número de graus de liberdade é $n_1 + n_2 - 2 = 2(n-1)$ se e só se $n_1 = n_2 = n$, como já antes observámos. Veja-se também Scheffé (1970, p. 1502).

O trabalho inicial de Welch reclama-se da herança de Student, e nesse aspecto trata de studentização. Satterthwaite (1946) propôs um estimador do número de graus de liberdade que redescobre resultados de Welch, mas que parece ter tido mais aceitação junto de especialistas e utilizadores de análise da variância, prestando-se naturalmente a estimar variâncias em contrastes e em combinações lineares mais elaboradas. Welch (1947a, 1947b, 1951, 1956a) publicou extensões mostrando que o seu trabalho serve para a comparação de k médias em situação de heterocedasticidade.

Já vimos as dificuldades que surgem tentando deduzir uma estatística de teste no âmbito da teoria de Neyman-Pearson, e seguidamente abordamos os desenvolvimentos teóricos de Welch. Apresentamos seguidamente a construção usualmente atribuída a Satterthwaite (1946), mas que é de facto, na essência, de Welch.

Enquanto Welch aborda a questão na perspectiva mais sofisticada de aproximações obtidas a partir de uma expansão em série,

Satterthwaite parte de uma aproximação de $\sum_{k=1}^{n} \alpha_k X_k$, onde os X_k são variáveis de qui-quadrado independentes, por $\frac{X}{\nu}$, onde X é uma variável de qui-quadrado com um número apropriado ν de graus de liberdade, equacionando momentos de primeira e de segunda ordem (se apenas equacionarmos momentos de primeira ordem é possível obter um número de graus de liberdade negativo!). Pareceu-nos interessante deduzir, na esteira de Mosteller e Rourke (1993), a "análise da variância simples não paramétrica" de Kruskal-Wallis, para mostrar a simplicidade e unidade essencial da metodologia estatística: neste caso deduz-se uma expressão para a estatística D, distância ponderada do *rank* médio do grupo (na amostra combinada) ao *rank* médio global $\frac{N+1}{2}$. Como a soma de *ranks*, que sob a hipótese nula de homogeneidade de efeitos devem ser variáveis aleatórias permutáveis, é aproximável por uma gaussiana (Teorema Limite Central para variáveis permutáveis , cf. Pestana e Velosa (2010), Teorema 8.4.1), a referida soma de quadrados deve ser aproximada por uma transformação linear — os quadrados são de variáveis centradas, mas com variância diferente de 1 — de uma variável qui-quadrado com $k-1$ graus de liberdade, e por isso calcula-se o valor médio $\mathbb{E}[D]$ da expressão, e divide-se D por $\dfrac{\mathbb{E}[D]}{k-1}$ para obter a estatística de teste H.

2.3.1 A abordagem de Welch

Na sequência das críticas de Bartlett à solução fiducial de Fisher para o problema das duas amostras, Welch tentou calcular valores críticos para a estatística

$$T_{S_1,S_2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

sem recorrer a argumentos fiduciais.

O trabalho de Welch tem muito em comum com o de Student (1908) acerca de inferências sobre o valor médio de uma gaussiana quando a variância real é desconhecida. No seu trabalho original, Student, depois de se ter apercebido de que uma variância estimada devia ser encarada como valor observado de uma variável aleatória, determinou os primeiros momentos da sua distribuição e notou que correspondiam a uma curva de Pearson de tipo III. Procurou então qual das curvas dessa família era compatível com os momentos calculados, a partir do que conseguiu determinar a distribuição exacta de T_{S_1,S_2}.

Também Welch começou por calcular os momentos do quadrado do denominador de T_{S_1,S_2} e viu que se coadunavam com os de uma Pearson de tipo III; a menos de uma transformação de escala, um qui-quadrado com índice fraccionário na nossa linguagem actual. Assim, a distribuição de T_{S_1,S_2} podia ser aproximada por uma t com número de graus de liberdade fraccionário. A distribuição exacta não tem neste caso forma analítica simples, mas a aproximação encontrada é suficientemente boa para estudar as propriedades de T_{S_1,S_2} e construir testes de hipóteses e intervalos de confiança eficientes.[(1)] Esta solução aproximada mais pragmática será desenvolvida adiante, quando se falar no

[(1)] A mesma técnica pode ser usada para aproximar a distribuição da estatística de Student T quando falha a hipótese de homocedasticidade, que em geral deixa de ser uma t. Foi nesse contexto que Welch (1938) a desenvolveu.

estimador de Satterthwaite. Anos depois, Welch propôs também a solução exacta que passamos a explicar.

A inferência sobre o valor médio μ de uma gaussiana com desvio padrão desconhecido σ, e a comparação de valores médios de duas gaussianas quando se conhece o quociente das respectivas variâncias, são simples porque a distribuição de t não envolve auqeles parâmetros populacionais, nem sequer as suas estimativas, e portanto o problema resume-se em achar a constante t_{α}, dependente apenas do número de graus de liberdade que é uma função simples da dimensão (ou das dimensões) da(s) amostra(s), tal que $\mathbb{P}[t_n < t_{\alpha}] = \alpha$.

A generalização natural a dois grupos independentes com variâncias distintas depara-se com a dificuldade de que a distribuição frequencista de T_{S_1,S_2}, referida ao universo de todos valores possíveis para as variâncias estimadas S_1^2 e S_2^2, depende dos parâmetros σ_1^2 e σ_2^2, neste caso constantes desconhecidas. A solução de Welch equivale a achar um valor crítico *aleatório* $T_{\alpha} = T_{\alpha}(S_1^2, S_2^2)$ que satisfaça $\mathbb{P}\left[T_{S_1,S_2} < T_{\alpha}\right] = \alpha$ para todos os valores das variâncias populacionais.

Welch parte de um quadro mais geral que o problema de Behrens-Fisher clássico, a comparação de k grupos independentes com heterocedasticidade. Embora o nosso foco seja a comparação de duas amostras, seguiremos de perto a sua notação original, que além de apontar de imediato para aplicações à análise de variância e à análise de regressão é difícil de superar em concisão.

Seja η um parâmetro populacional estimado por uma variável aleatória $Y \frown \mathit{Gaussiana}(\eta, \sigma_Y)$. Suponha-se que $\sigma_Y^2 = \sum \lambda_i \sigma_i^2$,

onde $i = 1, \ldots, k$, os λ_i são constantes positivas conhecidas e os σ_i^2 são variâncias estimadas por variáveis aleatórias S_i^2 tais que $\frac{f_i S_i^2}{\sigma_i^2} \frown \chi_{f_i}^2$, portanto com densidades

$$p_i(w_i) = \left(\frac{f_i}{2\sigma_i^2}\right) \frac{1}{\Gamma\left(\frac{f_i}{2}\right)} \left(\frac{f_i w_i}{2\sigma_i^2}\right)^{\frac{f_i}{2}-1} \exp\left(-\frac{f_i w_i}{2\sigma_i^2}\right),$$

independentes umas das outras e de Y. No problema de Behrens-Fisher, $k = 2$, $\eta = \mu_1 - \mu_2$, $Y = \overline{X}_1 - \overline{X}_2$, $f_1 = n_1 - 1$, $f_2 = n_2 - 1$, $\lambda_1 = \frac{1}{n_1}$, $\lambda_2 = \frac{1}{n_2}$, e S_1^2, S_2^2 são os estimadores centrados usuais das variâncias com base em cada amostra.

Procuremos uma variável aleatória H_α tal que $\mathbb{P}[Y - \eta < H_\alpha] = \alpha$. Como as estatísticas Y e $S_1^2, S_2^2, \ldots, S_k^2$ são conjuntamente suficientes para os parâmetros $(\eta, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$, é natural admitir que $H_\alpha = h_\alpha(S_1^2, S_2^2, \ldots, S_k^2)$ para alguma função determinista $h_\alpha : \mathbb{R}^k \to \mathbb{R}$ (dependente também das dimensões das amostras, ou seja dos f_i).

Para aligeirar a notação, escrevemos $\boldsymbol{w} = (w_1, w_2, \ldots, w_k)$. Então a probabilidade da desigualdade anterior, condicional aos valores observados das variâncias empíricas, é

$$\mathbb{P}\left[Y - \eta < H_\alpha \mid S_1^2 = w_1, \ldots, S_k^2 = w_k\right] = \Phi\left(\frac{h_\alpha(\boldsymbol{w})}{\sigma_Y}\right) \equiv j\boldsymbol{w}), \quad (2.7)$$

onde Φ é a função de distribuição da gaussiana padrão, e do teorema da probabilidade total para densidades vem que

$$\mathbb{P}\left[Y-\eta<H_{\alpha}\right]=\int_{0}^{\infty}\cdots\int_{0}^{\infty}j(\boldsymbol{w})\prod_{i}p_{i}(w_{i})\mathrm{d}w_{i}\equiv\Theta j=\alpha, \qquad (2.8)$$

onde $\boldsymbol{\Theta}$ representa o operador integral assim definido. Invertendo esta equação funcional em j, achamos a solução genérica $h_{\alpha}(w_{1},w_{2},\ldots,w_{k})$.

Para tal, Welch desenvolve j em série de Taylor multivariada em torno de $\boldsymbol{s}^{2}=(s_{1}^{2},s_{2}^{2},\ldots,s_{k}^{2})$, formalmente

$$j\exp\left(\sum_{i}\left[(w_{i}-s_{i}^{2})\partial_{i}\right]j\right),$$

onde as potências do operador diferencial múltiplo $\boldsymbol{\partial}$ se devem interpretar de acordo com $(\partial_{i}^{r}j)(\mathbf{w})=(\partial_{i}^{r}j)/(\partial w_{i}^{r})(\boldsymbol{s}^{2})$. Este desenvolvimento simbólico permite factorizar o integral e achar uma série formal para o operador $\boldsymbol{\Theta}$.

Por outro lado, substituindo Φ pela sua série de Taylor em torno de z_{α}, o quantil ascendente de nível α da gaussiana padrão, na definição (2.6) de j obtém-se também um desenvolvimento para o argumento de $\boldsymbol{\Theta}$. Compondo as duas séries, pode resolver-se a equação funcional (2.8) por aproximações sucessivas.

Neste ponto convém escrever

$$v_{\alpha}\equiv\frac{h_{\alpha}(s_{1}^{2},s_{2}^{2},\ldots,s_{k}^{2})}{\sqrt{\sum\lambda_{i}s_{i}^{2}}}\sim v_{0}+v_{1}+v_{2}+\ldots,$$

onde v_{r} contém os termos da ordem de $\frac{1}{f_{i}^{r}}$ ($r=0,1,2,\ldots$; $i=1,2,\ldots,k$). A primeira parcela corresponde à aproximação

à gaussiana em grandes amostras, $v_0 = z_\alpha$, e desprezando os termos de ordem superior a $\frac{1}{f_i^2}$ obtém-se[(2)]

$$v_\alpha \sim z_\alpha \left[1 + \frac{1+z_\alpha^2}{4}\sum \frac{c_i^2}{f_i} - \frac{1+z_\alpha^2}{2}\sum \frac{c_i^2}{f_i^2} + \right.$$
$$\left. + \frac{3+5z_\alpha^2+z_\alpha^4}{3}\sum \frac{c_i^3}{f_i^2} - \frac{15+32z_\alpha^2+9z_\alpha^4}{32}\left(\sum \frac{c_i^2}{f_i}\right)^2\right],$$

onde $c_i = \dfrac{\lambda_i s_i^2}{\sum \lambda_i s_i^2}$.

Welch observa que quando $k = 1$ a fórmula se reduz a um desenvolvimento conhecido para os quantis da t de Student:

$$t_\alpha \sim z_\alpha \left(1 + \frac{1+z_\alpha^2}{4f} + \frac{3+16z_\alpha^2+5z_\alpha^4}{96f^2} + \dots\right),$$

e comparando a sua solução com a de Behrens ($k = 2$), para a qual Fisher (1941) obtivera, até a ordem $1/f_i$, o desenvolvimento

$$z_\alpha \left[1 + \frac{1+z_\alpha^2}{4}\left(\frac{c_1^2}{f_1} + \frac{c_2^2}{f_2}\right) + c_1 c_2 \left(\frac{1}{f_1} + \frac{1}{f_2}\right)\right],$$

[(2)] Os detalhes da composição e inversão das séries são demasiado fastidiosos para reproduzir aqui, mas podem-se encontrar no artigo original Welch (1947a), que merece bem ser lido por inteiro e está facilmente acessível em `http://pds9.egloos.com/pds/200804/26/44/2332510.pdf`. O artigo de Welch (1951), que aplica a mesma técnica a um problema semelhante, usando funções geradoras de momentos, é também interessante e pode encontrar-se em `http://www.soph.uab.edu/Statgenetics/People/MBeasley/Courses/Welch1951.pdf`.

nota que coincidem quando as amostras são suficientemente grandes para se aproximar pela gaussiana, mas diferem a partir do termo de primeira ordem ou seja, nas suas palavras, "assim que se começa a dar importância à studentização". A fórmula de Welch leva ademais a intervalos de confiança mais apertados e testes de hipóteses mais sensíveis do que a de Fisher-Behrens .

O desenvolvimento de Welch foi estendido até termos da ordem de $\frac{1}{f_i^4}$ pela sua colaboradora Aspin (1948, 1949), que publicou tabelas de valores críticos na comparação de duas médias para algumas combinações de $n_1, n_2 \geq 7$ e $c_1 = 1 - c_2 = 0\,(0.1)\,1$. Mas já no seu artigo original de 1947 Welch reconhecera que a solução exacta era pouco prática, e sugerira que em vez disso se aproximassem os valores críticos pelos de uma t de Student com graus de liberdade aleatórios dados por $f = \frac{1}{\sum c_i^2/f_i}$ [(3)].

Observe-se que substituindo este valor na expressão dos quantis da t se obtém

$$z_\alpha \left[1 + \frac{1 + z_\alpha^2}{4} \sum \frac{c_i^2}{f_i} + \frac{3 + 16 z_\alpha^2 + 5 z_\alpha^4}{96} \left(\sum \frac{c_i^2}{f_i} \right)^2 + \dots \right],$$

que concorda com a expansão da solução exacta até o termo de primeira ordem. Estudos posteriores têm confirmado que esta aproximação é satisfatória e, nas palavras de Scheffé (1970), "não requer mais que as omnipresentes tabelas da t". Na próxima secção desenvolvemos esta solução aproximada.

[(3)] No artigo de 1947 Welch defende uma ligeira modificação a esta fórmula, mas mais tarde retirou a recomendação (Welch, 1949).

2.3.2 O estimador de Welch–Satterthwaite

As tabelas publicadas por Welch e Aspin são inconvenientes sobretudo por dependerem de três parâmetros, as dimensões de ambas as amostras e a razão entre as variâncias empíricas, o que complica a tabulação dos valores críticos, enquanto por exemplo para consultar as tabelas da t basta conhecer a dimensão total das amostras e o nível de significância pretendido. Por outro lado, os resultados de Welch são uma aproximação numérica dos valores reais com base num desenvolvimento em série que pode ser bastante oneroso de calcular.

Outra abordagem ao problema, em alternativa a calcular aproximações numéricas dos valores críticos exactos, era usar uma distribuição aproximada mas cujos quantis fossem mais fáceis de determinar. A distribuição t de Student é uma candidata natural, e pode-se mesmo mostrar que a distribuição exacta de T_{S_1,S_2} está compreendida entre a t com $\min\{n_1, n_2\} - 1$ graus de liberdade e a t com $n_1 + n_2 - 2$ graus de liberdade usada em situações de homocedasticidade.

Seguindo as linhas de raciocínio de Student, Welch procurou então aproximar o radicando no denominador desta estatística, que é uma combinação linear de quis-quadrados, por uma curva de Pearson de tipo III apropriada — o que corresponde a uma gama, ou seja um múltiplo de um qui-quadrado com graus de liberdade ν possivelmente não inteiros. Com efeito, de $\chi^2_\nu \equiv Gama\left(\frac{\nu}{2}, 2\right)$ e $Y \frown Gama(\alpha, \beta) \iff cY \frown Gama(\alpha, c\beta)$ para $c > 0$ decorre que $Gama(\alpha, \beta) \equiv \sigma^2 \frac{\chi^2_\nu}{\nu}$, com $\sigma^2 = \alpha\beta$ e $\nu = 2\alpha$. Deste modo, a distribuição do quociente podia ser aproximada por uma t com o mesmo número de graus de liberdade.

Welch usou este procedimento como confirmação dos resultados a que tinha chegado com o desenvolvimento em série, e também como forma de estudar o comportamento do teste, visto que a distribuição exacta era difícil de calcular. Mais tarde generalizou o método à situação em que estavam envolvidas diversas variâncias, abrindo o caminho para aplicações à análise de variância com grupos de variâncias heterogéneas.

No entanto, a divulgação deste método de um ponto de vista mais prático parece ter-se devido a Satterthwaite, que posteriormente publicou artigos de análise de variância em que deduzia a mesma aproximação de modo bastante mais simples e directo que o argumento de Welch, e talvez seja por isso que o nome deste é referido com tanta frequência a respeito da técnica de aproximação pelas distribuições t ou F aplicada quando as variâncias envolvidas são diferentes. A abordagem que a seguir apresentamos segue de perto Satterthwaite (1941, 1946).

Comecemos por notar que se extrairmos amostras aleatórias independentes $\boldsymbol{X}_1 = \left(X_{11}, \ldots, X_{1n_1}\right)$ e $\boldsymbol{X}_2 = \left(X_{21}, \ldots, X_{2n_2}\right)$ de populações respectivamente $X_1 \frown Gaussiana\left(\mu_1, \sigma_1\right)$ e $X_2 \frown Gaussiana\left(\mu_2, \sigma_2\right)$, então $\overline{X}_1 \frown Gaussiana\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$ e $\overline{X}_2 \frown Gaussiana\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$.

Se queremos comparar os valores médios das duas populações, o natural é usar

$$\overline{X}_1 - \overline{X}_2 \frown Gaussiana\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Estandardizando temos

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \frown \textit{Gaussiana}\,(0,1)\,.$$

Como σ_1 e σ_2 são desconhecidos, é natural estimá-los por S_1 e S_2, e usar a "studentização"

$$T_{S_1,S_2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \tag{2.9}$$

O problema reduz-se então a procurar uma boa aproximação para a distribuição daquela expressão.

O resultado em que se baseia a comparação de valores médios de populações gaussianas com diferentes variâncias é o seguinte:

Teorema 2.3.1.

Se $\boldsymbol{X}_1 = \left(X_{11}, \ldots, X_{1n_1}\right)$ *com os* $X_{1k} \frown \textit{Gaussiana}\,(\mu_1, \sigma_1)$ *independentes,* $k = 1, \ldots, n_1$*, e* $\boldsymbol{X}_2 = \left(X_{21}, \ldots, X_{2n_2}\right)$*, com os* $X_{2j} \frown \textit{Gaussiana}\,(\mu_2, \sigma_2)$ *independentes,* $j = 1, \ldots, n_2$*, forem amostras aleatórias independentes, então*

$$T_{S_1,S_2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \simeq t_\nu$$

onde o número de graus de liberdade ν é estimado por

$$\widetilde{\nu} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2\,(n_1-1)} + \frac{S_2^4}{n_2^2\,(n_2-1)}}.$$

Demonstração:

Se a finalidade é achar uma distribuição aproximada para V, uma ideia que intuitivamente se impõe é aproximar o quadrado $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ do denominador por $\sigma^2 \frac{Y_\nu}{\nu}$, onde $Y_\nu \frown \chi^2_\nu$, e ν e σ^2 são escolhidos de tal forma que

$$\mathbb{E}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) = \mathbb{E}\left(\sigma^2\,\frac{Y_\nu}{\nu}\right) = \sigma^2$$

e

$$\operatorname{var}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) = \operatorname{var}\left(\sigma^2\,\frac{Y_\nu}{\nu}\right) = \frac{2\,\sigma^4}{\nu}.$$

Ora

$$\mathbb{E}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

e de $\frac{(n_1-1)\,S_1^2}{\sigma_1^2} \frown \chi^2_{n_1-1}$ e $\frac{(n_2-1)\,S_2^2}{\sigma_2^2} \frown \chi^2_{n_2-1}$ segue-se que

$$\operatorname{var}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) = \frac{2\sigma_1^4}{n_1^2\,(n_1-1)} + \frac{2\sigma_2^4}{n_2^2\,(n_2-1)}.$$

Assim, a aproximação que estamos a fazer usa uma variável com o mesmo valor esperado e a mesma variância da que vai substituir se e só se

$$\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2} = \sigma^{\,2}$$

e

$$\nu = \frac{\left(\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}\right)^2}{\frac{\sigma_1^{\,4}}{n_1^2(n_1-1)} + \frac{\sigma_2^{\,4}}{n_2^2(n_2-1)}}\,.$$

Consequentemente, procedemos à aproximação

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \simeq \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\,\frac{Y_\nu}{\nu}}} =$$

$$= \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\frac{\sigma}{\sqrt{\frac{Y_\nu}{\nu}}}}\,.$$

Como $\sigma^{\,2} = \frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}$, no numerador temos uma variável aleatória Gaussiana padrão, e no denominador a raiz de uma variável qui-quadrado dividida pelo seu número de graus de liberdade, sendo as referidas variáveis mutuamente independentes. Quer isto dizer que

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \simeq t_\nu\,,$$

onde ν pode ser estimado substituindo as variâncias populacionais por variâncias amostrais,

$$\widetilde{\nu} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

□

Na prática, se não tivermos à mão meios de cálculo adequados e estivermos limitados a consulta de tabelas, aproximamos por um número de graus de liberdade natural.

A técnica que acabamos de usar no caso simples do problema das duas amostras generaliza-se sem novidade à aproximação de uma combinação linear qualquer de quis-quadrados independentes com coeficientes positivos por outro qui-quadrado. Para $Y_k \frown \chi^2_{\nu_k}$ independentes, $k = 1, \ldots, n$, tem-se

$$W = \sum_{k=1}^{n} a_k Y_k \simeq \sigma^2 \frac{Y_\nu}{\nu}$$

onde $Y_\nu \frown \chi^2_\nu$, $\sigma^2 = \sum_{k=1}^{n} a_k \nu_k$ e os graus de liberdade são estimados pela *aproximação de Satterthwaite*:

$$\widetilde{\nu} = \frac{\left(\sum_{k=1}^{n} a_k Y_k\right)^2}{\sum_{k=1}^{n} \frac{a_k^2}{\nu_k} Y_k^2}.$$

Anote-se também que para contrastes usamos

$$T_{S_1,\ldots,S_g} = \frac{\sum_{k=1}^{g} w_k \overline{X}_{k\bullet}}{\sqrt{\sum_{k=1}^{g} \frac{w_k^2 S_k^2}{n_k}}},$$

que tem distribuição t aproximada com ν graus de liberdade, estimados por

$$\widetilde{\nu} = \frac{\left(\sum_{k=1}^{g} w_k^2 S_k^2\right)^2}{\sum_{k=1}^{g} \frac{w_k^4 S_k^4}{n_k^2 (n_k - 1)}},$$

cf. Oehlert, 2000, p. 132.

Não deixa de ser interessante apontar que os poucos livros que referem este resultado citam, em geral, apenas o trabalho de Satterthwaite (1946), o qual atribui a ideia original a Smith (1936).[(4)] Kendall and Stuart (1961), pelo contrário, atribuem todo o crédito a Welch (1936, 1938), anterior e com sofisticação matemática superior à do trabalho de Satterthwaite. Samuels and Witmer (1999), que consistentemente abordam a comparação de médias nesta perspectiva mais alargada, creditam a Welch e a Satterthwaite os métodos que empregam, sem no entanto darem qualquer referência bibliográfica precisa. Davenport

[(4)] Num trabalho menos referenciado, apesar de disponível em `http://www.springerlink.com/content/6h77h7288554h864/fulltext.pdf`, Satterthwaite (1941) cita Welch (1938). Anote-se também que Welch (1938) afirma que utilizou anteriormente a mesma aproximação em Welch (1936).

and Webster (1975) serão dos poucos a citar ambos os autores e Smith.

A importância deste trabalho reside em fornecer uma fórmula para estimar o número de graus de liberdade, aparentemente pouco prática, e obtida equacionando momentos de primeira e de segunda ordem, mas que leva a um resultado sempre positivo. Casella and Berger (1990, 2002), um pouco mais prolixos sobre este resultado do que os outros bons livros de Estatística que consultámos, apresentam o estimador de Welch-Satterthwaite como um exemplo elaborado do método dos momentos.

Welch voltou ao assunto mais tarde, mostrando como a sua abordagem podia ser reformulada para comparar $k \geq 2$ médias, e em Welch (1951) trata de um modelo de análise de variância ponderada, com pesos inversamente proporcionais às variâncias empíricas (portanto aleatórios), usando um método semelhante ao desenvolvido em Welch (1947a).

2.4 Breve nota sobre a abordagem não paramétrica à localização de k amostras

Não é nosso objectivo ocuparmo-nos de abordagens não paramétricas para a comparação da localização de duas populações, nomeadamente populações não gaussianas[(5)].

[(5)] Chamamos no entanto a atenção do leitor para a relevância da questão, reconhecida pelo *International Statistical Institute* com a atribuição em 2011 de um prémio Jan Timbergen a Kavitha Mehendale e Manjula Kalluraya pelo seu inovador trabalho sobre testes não paramétricos no problema de Behrens-Fisher.

É interessante porém referir com algum detalhe a forma como Mosteller e Rourke (1993, p. 242) apresentam a análise de variância simples não paramétrica de Kruskal-Wallis, pois evidencia a unidade fundamental dos raciocínios estatísticos — como depois comentamos, quer Welch e Satterthwaitte quer Kruskal e Wallis fazem aproximaçõs a variáveis qui-quadrado, mas os dois primeiros autores equacionam momentos à cabeça, enquanto Kruskal e Wallis não reduzem as gaussianas que aproximam as somas de *ranks* a escala 1 antes de as elevar ao quadrado, pelo que é na fase final que equacionam momentos.

Sejam $\boldsymbol{X}_1 = \{X_{1j}\}_{j=1}^{n_1}$, ..., $\boldsymbol{X}_\nu = \{X_{\nu j}\}_{j=1}^{n_\nu}$ amostras independentes, $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\nu)$ a amostra combinada, r_{ij}, $i = 1, \ldots, \nu$; $j = 1, \ldots, n_i$ o *rank* de X_{ij} na amostra combinada.

Usualmente requere-se que a distribuição subjacente seja contínua, para evitar igualdade de valores; no entanto a precisão finita de qualquer instrumento de medição leva a que haja observações empatadas, a que se atribui o *rank ajustado*, isto é o *rank* médio dos *ranks* que lhes seriam atribuídos no caso de ordenar esses elementos sequencialmente — ao atribuir-lhes depois o *rank* médio, a arbitrariedade dessa seriação torna-se irrelevante; o uso de *ranks* ajustados não traz alterações ao cálculo de médias, mas altera as variâncias, devendo usar-se factores de correcção apropriados (Pestana e Velosa 2010, p. 647, Sprent and Smeeton, 2001).

Definam-se ainda as soma dos *ranks* em cada grupo $R_k = \sum_{j=1}^{n_k} r_{kj}$, $k = 1, \ldots, \nu$. Seja

$$D = \sum_{k=1}^{\nu} n_k \left(\frac{R_k}{n_k} - \frac{N+1}{2} \right)^2 .$$

Naquela expressão, $\frac{R_k}{n_k}$ é o *rank* médio no k-ésimo grupo. Como o *rank* médio da amostra combinada, que tem $N = \sum_{k=1}^{\nu} n_k$ elementos, é $\frac{N+1}{2}$, a estatística D mede globalmente os desvios quadrados (e ponderados pelo número de elementos em cada grupo, como convém) do *rank* médio de cada grupo ao *rank* médio global.

Também nesta situação há a convicção de que a estatística D, sendo uma soma de quadrados de variáveis que, devido ao teorema limite central para somas de parcelas permutáveis, podem ser consideradas aproximadamente gaussianas, é bem aproximada por uma variável proporcional a um qui-quadrado com $\nu - 1$ graus de liberdade. No entanto, como não começámos por estandardizar as variáveis, há que calcular o valor médio $\mathbb{E}[D]$ de D, e usar $H = \dfrac{\nu - 1}{\mathbb{E}[D]} D$ que tem valor médio $\nu - 1$ — por outras palavras, usar uma variável H cujo valor médio é o valor médio da $\chi^2_{\nu-1}$.

A álgebra deste problema é bastante elementar: Por um lado, desenvolvendo os quadrados e efectuando os somatórios,

$$D = \sum_{k=1}^{\nu} \frac{R_k^2}{n_k} - \frac{N(N+1)^2}{4}.$$

Por outro lado,

$$\mathbb{E}[D] = \sum_{k=1}^{\nu} n_k \mathbb{E}\left[\left(\frac{R_k}{n_k} - \frac{N+1}{2}\right)^2\right] = \sum_{k=1}^{\nu} n_k \mathbb{E}\left[\left(\frac{R_k}{n_k} - \mathbb{E}\left[\frac{R_k}{n_k}\right]\right)^2\right],$$

e assim

$$\mathbb{E}\left[D\right] = \sum_{k=1}^{\nu} n_{k} \mathrm{var}\left[\frac{R_{k}}{n_{k}}\right].$$

Ora os r_{kj} (vamos assumir que estamos na situação ideal de não haver empates, pois se houver empates há que usar correcções adequadas, como já foi referido) são inteiros consecutivos de 1 a N, dispostos aleatoriamente. Temos assim observações de uma variável aleatória uniforme discreta $X = \left\{ \begin{array}{ll} i & i = 1, \ldots, N \\ \frac{1}{N} & \end{array} \right.$, pelo que $\mathbb{E}[X] = \sum_{1}^{N} i \, \frac{1}{N} = \frac{N+1}{2}$ e $\mathbb{E}[X^{2}] = \sum_{1}^{N} i^{2} \, \frac{1}{N} = \frac{(N+1)(2N+1)}{6}$, e consequentemente a variância de N inteiros consecutivos é $\frac{N^{2}-1}{12}$.

Como a atribuição dos *ranks* em cada grupo $k = 1, \ldots, \nu$ é uma extracção de n_{k} desses números de 1 a N, sem reposição, no cálculo da variância há que usar o factor de correcçã0 $\frac{N-n_{k}}{N-1}$ e segue-se que $\mathrm{var}\left[\frac{R_{k}}{n_{k}}\right] = \frac{1}{n_{k}^{2}} \sum_{j=1}^{n_{k}} \frac{N^{2}-1}{12} \frac{N-n_{k}}{N-1}$, e portanto

$$\mathbb{E}\left[D\right] = \frac{N+1}{12} \sum_{k=1}^{\nu} \left(N - n_{k}\right) = \frac{\left(N+1\right) N \left(\nu - 1\right)}{12}$$

e consequentemente a estatística de Kruskal–Wallis

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{\nu} \frac{R_{k}^{2}}{n_{k}} - 3(N+1)$$

tem uma distribuição que é bem aproximada pela da variável $\chi_{\nu-1}^{2}$.

Note-se que é, na sua essência, um modo de pensar muito semelhante ao usado por Satterthwaite, apenas a fase em que se usa uma aproximação do(s) momento(s) é diferente. Kruskal e Wallis começam por fazer os desenvolvimentos algébricos, usando quadrados de gaussianas não reduzidas, e é depois disso que "ajeitam"o primeiro momento, por forma a aproximar por um qui-quadrado com o número adequado de graus de liberdade. Quer Welch quer Satterthwaite comeąm pela ponta oposta, isto é começam por forçar a variável que usam como aproximação a ter valor médio e variância iguais aos da variável que aproximam, e só depois disso psosseguem para os desenvolvimentos algébricos do problema.

Note-se aliás que Welch e Satterthwaite se inscrevem na tradição que vem de Helmert (1875) e Student (1908) de procurar aproximações para o denominador usando os momentos como "instrumento de diagnóstico".

Capítulo 3

O Resultado Exacto de Scheffé

e outros desenvolvimentos do problema de Behrens-Fisher

Quer Jeffreys, quer Behrens e Fisher, quer Welch, Satterthwaite e Aspin procuraram resolver o problema da comparação de valores médios de duas populações gaussianas com variâncias diferentes e desconhecidas encontrando (com base em princípios diferentes) a distribuição de

$$T_{S_1,S_2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

em que $\overline{X}_1, S_1^2$ e $\overline{X}_2, S_2^2$ provêm de amostras aleatórias independentes extraídas de cada uma das populações cujos valores

médios se pretende comparar.

Mas, como foi visto, qualquer das soluções exactas leva a distribuições bastante complexas, de que não há tabelas facilmente acessíveis dos quantis que seriam relevantes para tomar decisões, de cálculo difícil, sugerindo mesmo imediatamente a Welch a conveniência de aproximar a distribuição exacta por uma t de Student apropriada.

Têm sido propostas outras variáveis ligeiramente diferentes, sempre com o objectivo de melhorar a aplicação prática, e em geral numa tentativa de ter melhor aproximação a uma distribuição t conveniente.

Scheffé (1943, 1944) considerou a questão de outro ponto de vista, chegando de certo modo a uma resposta definitiva ao problema, se bem que inesperada. A sua ideia foi quase óbvia, mas também brilhante: Em vez de estudar a distribuição da variável fulcral T_{S_1,S_2}, porque não procurar uma outra função das amostras que *tenha efectivamente distribuição exacta* t de Student?

Se o objectivo é fazer inferência sobre $\delta = \mu_1 - \mu_2$ usando a distribuição t de Student, devemos procurar uma função linear L das amostras aleatórias (ainda no âmbito restrito de gaussianidade e indepedência, claro), e uma função quadrática Q, tais que, para quaisquer valores de σ_1^2 e de σ_2^2,

- L e Q sejam independentes;
- $\mathbb{E}[L] = \delta$ e $\mathrm{var}[L] = V$;
- $\dfrac{Q}{V}$ seja uma variável qui-quadrado com ν graus de liberdade.

Nestas circunstâncias

$$t = \frac{L - \delta}{\sqrt{\dfrac{Q}{\nu}}}$$

tem distribuição t com ν graus de liberdade.

Scheffé (1944) mostrou que os anteriores investigadores do "problema dos dois valores médios" tinham estado a olhar para o lado errado, no sentido em que tinham cedido à tentação natural de trabalhar com funções invariantes para permutações de cada uma das amostras aleatórias.

De facto, suponha-se que t é uma função simétrica nos argumentos, no sentido em que se mantém invariante para permutações de qualquer uma das duas amostras, e que por isso

$$\begin{cases} L = c_1 \displaystyle\sum_{k=1}^{n_1} X_{1k} + c_2 \sum_{j=1}^{n_2} X_{2j} \\ \\ Q = c_3 \displaystyle\sum_{k=1}^{n_1} X_{1k}^2 + c_4 \sum_{k=1}^{n_1} \sum_{\substack{i=1 \\ i \neq k}}^{n_1} X_{1k} X_{1i} + \\ \qquad + c_5 \displaystyle\sum_{j=1}^{n_2} X_{2j}^2 + c_6 \sum_{j=1}^{n_2} \sum_{\substack{\ell=1 \\ \ell \neq j}}^{n_2} X_{2j} X_{2\ell} + c_7 \sum_{k=1}^{n_1} \sum_{j=1}^{n_2} X_{1k} X_{2j} \end{cases}$$

Então

$$\delta = \mu_1 - \mu_2 = \mathbb{E}\,[L] = c_1 n_1 \mu_1 + c_2 n_2 \mu_2$$

e consequentemente

$$c_1 = \frac{1}{n_1} \quad \text{e} \quad c_2 = -\frac{1}{n_2}\,.$$

Mas daí segue-se que

$$L = \overline{X}_1 - \overline{X}_2,$$

e

$$V = \mathrm{var}\,[L] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Como $\dfrac{Q}{V} \frown \chi_\nu^2$, segue-se que

$$\mathbb{E}\,[Q] = \nu \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right).$$

Por outro lado, usando a representação de Q como função simétrica das amostras,

$$\begin{aligned}\mathbb{E}\,[Q] = {} & c_3 n_1 \left(\sigma_1^2 + \mu_1^2\right) + c_4 n_1 (n_1 - 1)\mu_1^2 + \\ & + c_5 n_2 \left(\sigma_2^2 + \mu_2^2\right) + c_6 n_2 (n_2 - 1)\mu_2^2 + c_7 n_1 n_2 \mu_1 \mu_2.\end{aligned}$$

Segue-se então que

$$\begin{cases} c_3 = \frac{\nu}{n_1^2} \\ c_4 = -\frac{\nu}{n_1^2 (n_1 - 1)} \\ c_5 = \frac{\nu}{n_2^2} \\ c_6 = -\frac{\nu}{n_2^2 (n_2 - 1)} \\ c_7 = 0 \end{cases}$$

e consequentemente

$$Q = \frac{\nu}{n_1}\left(\frac{\sum_{k=1}^{n_1} X_{1k}^2}{n_1} - \frac{\sum_{k=1}^{n_1}\sum_{\substack{i=1\\ i\neq k}}^{n_1} X_{1k}X_{1i}}{n_1(n_1-1)}\right) + \frac{\nu}{n_2}\left(\frac{\sum_{j=1}^{n_2} X_{2j}^2}{n_2} - \frac{\sum_{j=1}^{n_2}\sum_{\substack{\ell=1\\ \ell\neq j}}^{n_2} X_{2j}X_{2\ell}}{n_2(n_2-1)}\right) =$$

$$= \nu\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right).$$

Obviamente $\dfrac{Q}{V} = \nu\,\dfrac{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ não pode ser um qui-quadrado com ν graus de liberdade para todos os valores possíveis de n_1, n_2, σ_1, σ_2.

3.1 A solução de Scheffé

Face à impossibilidade de encontrar funções simétricas das amostras que levem ao resultado pretendido, procurem-se soluções adequadas noutra classe de funções.

Suponha-se que $n_1 \leq n_2$, e que usamos

$$\begin{cases} L = \dfrac{1}{n_1}\displaystyle\sum_{k=1}^{n_1} D_k \\[2ex] Q = \dfrac{1}{n_1}\displaystyle\sum_{k=1}^{n_1} (D_k - L)^2 \end{cases}$$

onde os D_k são variáveis aleatórias i.i.d., gaussianas, com

$$\mathbb{E}[D_k] = \delta \quad \text{e} \quad \text{var}[D_k] = \sigma^2, \quad k = 1, \ldots, n_1.$$

A variável aleatória

$$t = \frac{L - \delta}{\sqrt{\frac{Q}{n_1 - 1}}}$$

tem então distribuição t de Student com $n_1 - 1$ graus de liberdade.

Escolham-se

$$D_k = X_{1k} - \sum_{j=1}^{n_2} c_{kj} X_{2j}.$$

As condições necessárias e suficientes para se ter, como pretendido, $\mathbb{E}[D_k] = \delta$, e $\text{var}[D_k] = \sigma^2$, para $k = 1, \ldots, n_1$, são

$$\begin{cases} \sum_{k=1}^{n_2} c_{kj} = 1 \\ \sum_{k=1}^{n_2} c_{kj}^2 = c^2 \\ \sum_{k=1}^{n_2} c_{kj} c_{ij} = 0 \quad (i \neq k) \end{cases}$$

e assim o intervalo de confiança para δ, com coeficiente de confiança $1 - \alpha$, é

$$\left(L - t_{n_1 - 1,\, 1 - \frac{\alpha}{2}} \sqrt{\frac{Q}{n_1 - 1}},\; L + t_{n_1 - 1,\, 1 - \frac{\alpha}{2}} \sqrt{\frac{Q}{n_1 - 1}} \right)$$

e o seu comprimento $\mathcal{L}$ tem valor esperado

$$\mathbb{E}\left[\mathcal{L}\right] = 2\, t_{n_1-1,\,1-\frac{\alpha}{2}}\ \frac{\sigma}{n_1-1}\ \mathbb{E}\left[\sqrt{\frac{(n_1-1)\,Q}{\sigma^2}}\ \right] =$$

$$= 2\, t_{n_1-1,\,1-\frac{\alpha}{2}}\ \frac{\sigma}{n_1-1}\ \frac{\sqrt{2}\,\Gamma\left(\frac{n_1}{2}\right)}{\Gamma\left(\frac{n_1-1}{2}\right)}\,.$$

Para minimizar aquele comprimento esperado há então que minimizar

$$\sigma = \sqrt{\sigma_1^2 + c^2\sigma_2^2},$$

ou seja, há que minimizar c^2.

Reescrevendo a condição necessária e suficiente sobre os c_{kj} sob forma matricial,

$$\begin{cases} \boldsymbol{c}_k\,\boldsymbol{u}' = 1 \\ \boldsymbol{c}_k\,\boldsymbol{c}_j' = c^2 & j = k \\ \boldsymbol{c}_k\,\boldsymbol{c}_j' = 0 & j \neq k \end{cases}$$

onde $\boldsymbol{c}_k$ é o vector da k-ésima linha da matriz $\left[c_{kj}\right]$ e $\boldsymbol{u}$ é o vector linha com elemento genérico 1, e considerando ainda $n_2 - n_1$ vectores satisfazendo as condições de ortogonalidade acima expressas e tais que os n_2 vectores $\boldsymbol{c}_k$ constituam uma base do espaço vectorial de dimensão n_2, podemos exprimir $\boldsymbol{u}$ como combinação linear dos vectores da base:

$$\boldsymbol{u} = \sum_{j=1}^{n_2} \alpha_j\,\boldsymbol{c}_j$$

e de

$$1 = \boldsymbol{c}_k\,\boldsymbol{u}' = \boldsymbol{c}_k \sum_{j=1}^{n_2} \alpha_j\,\boldsymbol{c}_j' = \sum_{j=1}^{n_2} \alpha_j\,\boldsymbol{c}_k\,\boldsymbol{c}_j' = g_k\,c^2$$

segue-se que

$$g_k = \frac{1}{c^2}, \quad k = 1, \ldots, n_1.$$

Por outro lado, como $\boldsymbol{u}$ é o vector linha de unidades,

$$\boldsymbol{u}\boldsymbol{u}' = \left(\sum_k g_k \boldsymbol{c}_k\right)\left(\sum_k g_k \boldsymbol{c}_k'\right) = \sum_{k=1}^{n_2} g_k^2 \boldsymbol{c}_k \boldsymbol{c}_k' = n_2,$$

e portanto

$$n_2 = c^2 \left(\sum_{k=1}^{n_1} g_k^2 + \sum_{k=n_1+1}^{n_2} g_k^2\right) = \frac{n_1}{c^2} + c^2 \sum_{k=n_1+1}^{n_2} g_k^2 \implies n_2 \geq \frac{n_1}{c^2}\,.$$

Consequentemente,

$$c^2 \geq \frac{n_1}{n_2}\,.$$

Na expressão acima temos igualdade no caso de $g_k = 0,\ k = n_1 + 1, \ldots, n_2$. Nessas condições, $\boldsymbol{u}$ é um vector do espaço de dimensão n_1 dos vectores $\boldsymbol{c}_k$ originais e há infinitas soluções, que se obtêm rodando qualquer uma delas.

Scheffé (1943) propôs uma solução particularmente elegante:

$$\left\{\begin{array}{lcll} c_{kk} & = & \sqrt{\frac{n_1}{n_2}} - \sqrt{\frac{1}{n_1 n_2}} + \frac{1}{n_2} & k = 1, \ldots, n_1 \\ c_{kj} & = & -\sqrt{\frac{1}{n_1 n_2}} + \frac{1}{n_2} & j \neq k = 1, \ldots, n_1 \\ c_{kj} & = & \frac{1}{n_2} & j = n_1 + 1, \ldots, n_2 \end{array}\right. .$$

Substituindo na expressão definitória dos D_k, obtém-se

$$D_k = X_{1k} - \sum_{j=1}^{n_2} c_{kj} X_{2j} =$$

$$= X_{1k} - \sqrt{\frac{n_1}{n_2}} X_{2k} + \sqrt{\frac{1}{n_1 n_2}} \sum_{k=1}^{n_1} X_{2k} - \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j},$$

$k = 1, \ldots, n_1$.

Tem-se então

$$L = \frac{1}{n_1} \sum_{k=1}^{n_1} D_k = \frac{1}{n_1} \sum_{k=1}^{n_1} \left[X_{1k} - \sqrt{\frac{n_1}{n_2}} X_{2k} + \sqrt{\frac{1}{n_1 n_2}} \sum_{k=1}^{n_1} X_{2k} - \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j} \right] =$$

$$\overline{X}_1 - \frac{1}{\sqrt{n_1\, n_2}} \sum_{k=1}^{n_1} X_{2k} + \frac{1}{n_1 \sqrt{n_1\, n_2}} \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} X_{2k} - \overline{X}_2 =$$

$$= \overline{X}_1 - \frac{1}{\sqrt{n_1\, n_2}} \sum_{k=1}^{n_1} X_{2k} + \frac{1}{\sqrt{n_1\, n_2}} \sum_{k=1}^{n_1} X_{2k} - \overline{X}_2 = \overline{X}_1 - \overline{X}_2$$

e por outro lado

$$Q = \frac{1}{n_1} \sum_{k=1}^{n_1} (D_k - L)^2 =$$

$$= \frac{1}{n_1} \sum_{k=1}^{n_1} \left(X_{1k} - \sqrt{\frac{n_1}{n_2}} X_{2k} + \sqrt{\frac{1}{n_1 n_2}} \sum_{k=1}^{n_1} X_{2k} - \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j} - \overline{X}_1 + \overline{X}_2 \right)^2 =$$

$$= \frac{1}{n_1} \sum_{k=1}^{n_1} \left[\left(X_{1k} - \sqrt{\frac{n_1}{n_2}} X_{2k} \right) - \left(\overline{X}_1 - \frac{1}{n_1} \sqrt{\frac{n_1}{n_2}} \sum_{k=1}^{n_1} X_{2k} \right) \right]^2 =$$

$$\frac{1}{n_1} \sum_{k=1}^{n_1} \left[\left(X_{1k} - \sqrt{\frac{n_1}{n_2}} X_{2k} \right) - \frac{1}{n_1} \left(\sum_{k=1}^{n_1} X_{1k} - \sqrt{\frac{n_1}{n_2}} \sum_{k=1}^{n_1} X_{2k} \right) \right]^2$$

pelo que

$$\begin{cases} L = \overline{X}_1 - \overline{X}_2 \\ Q = \frac{1}{n_1} \sum_{k=1}^{n_1} \left(U_k - \overline{U}\right)^2 \end{cases},$$

onde $U_k = X_{1k} - \sqrt{\frac{n_1}{n_2}}\, X_{2k}$ e $\overline{U} = \frac{1}{n_1} \sum_{k=1}^{n_1} U_k$.

Finalmente,

$$T_{Sch} = \sqrt{n_1}\, \frac{\left(\overline{X}_1 - \overline{X}_2 - \delta\right)}{\sqrt{\frac{\sum_{k=1}^{n_1} \left(U_k - \overline{U}\right)^2}{n_1 - 1}}} \frown t_{n_1 - 1}.$$

Note-se que depois deste desenvolvimento algébrico se torna óbvio que na segunda amostra se usam, no cálculo de Q, apenas as n_1 primeiras observações — e estas são naturalmente quaisquer n_1 das n_2 da amostra com mais elementos. Podemos (devemos é porventura o termo mas correcto) escolher aleatoriamente essas n_1 observações da segunda amostra. Ou, por outro lado, usar essa aleatorização como suporte de estudos de simulação.

O trabalho de Scheffé é notável a vários títulos.

Por um lado, arreda uma daquelas "falsas evidências" — a ideia de que devemos procurar uma função invariante para permutações das componentes das amostras.

Por outro lado, obtém uma solução exacta que é válida para qualquer escolha aleatória de n_1 elementos da segunda amostra. Há, evidentemente, uma perda de eficiência neste procedimento, que não sabemos quantificar. Remetemos para Scheffé (1943)

quem tiver curiosidade em saber como comparam, em termos médios, os intervalos de confiança obtidos pelo método de Scheffé e os intervalos de confiança exactos, quando $\theta = \frac{\sigma_1^2}{\sigma_2^2}$ é conhecido.

3.2 Outros desenvolvimentos do problema de Behrens–Fisher

O problema de Behrens–Fisher teve um período de grande notoriedade, durante a fase embrionária da matematização da inferência estatística.

Durante uma fase inicial foi possível pensar que os grandes criadores de ideias estavam a falar da mesma coisa, usando termos e descrevendo procedimentos — que podemos apodar de frequencistas, bayesianos, fiducialistas — só aparentemente diferentes, chegando porém a resultados idênticos.

O problema das duas amostras independentes de populações gaussianas com variâncias diferentes mostrou, sem margem para dúvidas, que essas construções da Estatística não só eram conceptualmente diversas como podiam construir soluções diferentes para o mesmo problema. Tal originou controvérsias amargas e rancorosas — veja-se por exemplo os artigos de Fisher, de Neyman, de Bartlett e de Welch no *JRSS* **B** de 1956.

Durante muito tempo o esforço centrou-se em justificar por que razões no problema de Behrens–Fisher as soluções são diferentes, enquanto no caso de inferência sobre o valor médio de uma população, ou sobre a diferença de valores médios de duas populações em que o quociente $\frac{\sigma_1^2}{\sigma_2^2}$ é conhecido, as aborda-

gens frequencista, fiducial e bayesiana levam ao mesmo resultado. Concordamos porém com Kendall and Stuart (1961), quando comentam que o esforço seria melhor empregue em explicar por que razões no caso de uma amostra ou no caso de duas amostras de populações com iguais variâncias as soluções coincidem. (Fisher, aliás, usa também algumas frases sibilinas na sua polémica com Bartlett, que parece quererem dizer a mesma coisa; afirma ele que não se faz sentido criticar o seu trabalho por levar a intervalos de confiança que não admitem uma interpretação frequencista, pois o que há a esperar é mesmo que chegue a uma solução diferente da frequencista).

A notoriedade do problema de Behrens–Fisher advinha de ser o exemplo natural para apoiar argumentos das diversas escolas — e, claro, da notoriedade dos intervenientes na polémica. Scheffé (1943, 1944), ao mostrar que na perspectiva clássica não podia haver solução exacta para o problema (ao contrário do que Welch parece pensar nos seus trabalhos iniciais), e ao construir uma solução elegante e simples em que a aleatorização da escolha de elementos na amostra de maior dimensão permite deduzir uma variável pivotal com distribuição t exacta, ainda que com alguma perda de informação, resolveu temporariamente a questão de forma satisfatória. O problema de Behrens–Fisher passou a ser uma curiosidade, raramente merecendo mais do que algumas linhas nos manuais escolares. O próprio Fisher, no *Statistical Methods for Research Workers*, o refere como problema de escassa importância prática.

Cochran (1951) reabordou a questão no contexto mais geral de testar p relações lineares

$$\sum_{k=1}^{n} c_{kj}\sigma_k^2, \ j = 1, \ldots, p$$

entre n variâncias , com σ_k^2 estimada por s_k^2 e $\frac{(n_k-1)\,s_k^2}{\sigma_k^2} \frown \chi^2_{n_k-1}$, comentando excplicitamente (p. 19) que o clássico problema de Behrens-Fisher é equivalente a testar $\sigma_1^2 + \sigma_2^2 - \sigma_3^2 = 0$, com $\widehat{\sigma}_3^2 = \widehat{\sigma_1^2 + \sigma_2^2}$ estimado por $(\overline{x}_1 - \overline{x}_2)^2$, com 1 grau de liberdade.

Valerá a pena ressuscitar o problema? Certamente, quando mais não seja por ter sido um dos grandes desafios que se colocaram à Estatística e pelas ideias e métodos sofisticados que, devido à sua dificuldade, obrigou as diversas soluções a utilizar.

Nos últimos anos, diversos livros de Análise da Variância e de Planeamento de Experiências começaram a referir o "estimador de Satterthwaite", e também em Bioestatística a questão começa a ganhar importância prática — Samuels and Witmer (1999), por exemplo, usam constantemente os métodos de Welch e Satterthwaite. Mais relevante ainda, a questão chegou ao prestigiado *Teacher's Corner* do *American Statistician*, e Markowski and Markowski (1990) recomendam o ensino rotineiro do teste de Welch–Satterthwaite em cursos de estatística elementar, reclamando que este é mais vantajoso do que o teste t para duas amostras precedido de teste sobre a igualdade das variâncias. E o *software* R por defeito executa o teste de Welch–Satterthwaite para comparar valores médios de duas populações gaussianas.

Scheffé (1970) e Davenport e Webster (1975) trouxeram de novo este problema "esquecido" a primeiro plano, e Dunnet (1980), Fenstad (1983), Moser *et al.* (1989), Lee and Fienberg (1991) desenvolveram estudos aprofundados sobre a comparação de valores médios sob heterocedasticidade, e estimadores ponderados da variância. A aplicação pragmática em vista é proceder a testes de comparações múltiplas, emparelhadas, ou sobre contrastes, em análise da variância. Um dos resultados mais

interessantes tem que ver com a situação de, usando a metodologia de Welch–Satterthwaite, aproximar uma combinação linear de quis-quadrados por uma gama (com escala diferente de 2), no caso de alguns dos coeficientes serem negativos, com a "receita" de somar essas parcelas ao numerador e ao denominador. Uma recomendação quase herética no que respeita dedução matemática, mas que os estudos de simulação parecem de facto recomendar!

Os estudos de eficiência sobre os vários testes propostos para o problema de Behrens–Fisher parecem indicar que a solução de Welch–Aspin é a que tem melhor desempenho. Quer esta quer as soluções de Fisher e Jeffreys são preferíveis ao teste t de Student para populações homocedásticas, o qual rapidamente perde potência à medida que aumenta a diferença entre as variâncias reais.

Para efeitos práticos, na aproximação de Welch–Satterthwaite usa-se em geral uma aproximação inteira ao número fraccionário de graus de liberdade; com essa aproximação, na implementação usa-se a tabela comum dos quantis da t de Student em vez da tabela publicada por Aspin (1949), necessariamente muito parcelar, e os resultados são em geral satisfatórios. Por outro lado, a solução de Scheffé proporciona um teste exacto baseado na distribuição t. O número de graus de liberdade que se deve usar em aproximações tem sido recorrentemente debatido, recomendamos um artigo já antigo de Gaylor and Hopper (1969).

Dudewicz e os seus colaboradores (Dudewicz and Ahmed, 1998, 1999; Dudewicz *et al.*, 2007) aprofundaram o estudo de soluções exactas e aproximações assintóticas do problema de Behrens-Fisher. A panorâmica publicada por Kim and Cohen

(1998) contém pontos de vista interessantes. Sawilowsky (2002) enaltece a importância teórica do problema, e alinha pelo ponto de vista de Fisher, de que a sua importância prática é diminuta — ponto de vista de que não partilhamos. Barnard (1995), profundo mesmo no que apresenta modestamente como uma anedota, é um inestimável contributo para repensar os fundamentos da Estatística.

Babu and Dadmanaban (2002) e Neubert and Brunner (2007) trazem novas ideias para este velho problema, aproveitando progressos da estatística computacional, nomeadamente no que se refere a testes de permutação, num contexto não-paramétrico.

Fisher, no seu *Statistical Methods and Scientific Inference*, dá mais importância ao problema do que no *Statistical Methods for Research Workers*. Diz que a questão é potencialmente importante, e que só não tem sido mais considerada por em geral a recolha de dados não ser feita na perspectiva de os obter com diferentes graus de precisão. Palavras de certo modo proféticas: o desenvolvimento recente da meta-análise, resultado natural da necessidade de rentabilizar evidência estatística, tentando extrair conclusões da globalização de estudos parcelares que individualmente podem não ser conclusivos, porventura devido à escassez de informação disponível, ou procurando harmonizar resultados discrepantes de diversos estudos, confere a este problema nova relevância, desta vez prática. Nesta área, o conceito de valores de prova generalizados pode revelar-se de grande importância, veja-se Tsui and Tang (2005).

Se no passado o problema de Behrens-Fisher pode ter sido olhado como uma curiosidade matemática que passava ao lado da Estatística, o crescimento exponencial da publicaçãode dados recolhidos com protocolos e precisões diversas, e de conclusões

tantas vezes discrepantes baseadas nesses dados obrigam a comunidade estatística a não se alhear dos resultados de Fisher-Jeffreys e de Welch-Satterthwaite — ainda por cima há para todos os gostos, é só optar pela solução frequencista ou bayesiana, ou mesmo fiducialista (dizem as más-línguas que nem Fisher entendeu o que isso fosse[(1)]). O fiducialismo de Fisher foi durante muito tempo considerado o patinho feio a que esse homem genial pertinazmente se agarrou, e teve um eclipse quase total na cena estatística na segunda metade do século XX. O importante trabalho de Wilkinson (1977), e a discussão que gerou, merece uma leitura cuidada, tal como Pitman (1957). Trabalhos mais recentes tratam o fiducialismo com mais apreço, Zabell (1992), e sobretudo Efron (1998) e Hampel (2006) parecem sugerir que os desenvolvimentos recentes da Ciência estão a transformar esse patinho feio do século XX num exuberante cisne do século XXI.

E, naturalmente, procuram-se soluções para populações não gaussianas. Para além dos desenvolvimentos apresentados nos capítulos que se seguem, veja-se Akahira (2002), que estuda a comparação de médias em populações gama.

---

(1) A. W. F. Edwards, ao propor um voto de louvor ao trabalho de Wilkinson (1977, p. 145.), afirma:

> "*L. J. Savage reported Fisher as saying shortly before his death: 'I don't understand yet what fiducial probability does. We shall have to live with it for a long time before we know what it is doing for us. But it should not be ignored just because we don't have yet a clear interpretation.*"

Claro que nem Edwards nem Savage são más-línguas, apenas confirmam que haverá algum fundamento para olhar com desconfiança para essa estranha criação de Fisher.

Conta-se, também, que quando numa conferência perguntaram a Fisher se as probabilidades fiduciais obedeciam à axiomática de Kolmogorov, ele respondeu "*Kol who?*".

Capítulo 4

Normal?

Podemos facilmente indicar meia-dúzia de razões para o protagonismo do modelo gaussiano (cf. Pestana e Mendonça, 2007), mas essas mesmas razões demonstram bem a a que ponto o modelo gaussiano é "anormal". Limitamo-nos aqui — porque é o que tem interesse directo no plano desta obra — a detalhar dois pontos de importância capital no que se refere a inferência sobre a localização usando a escala, nomeadamente a caracterização da gaussiana pela independência da média e variância empíricas, e a caracterização "de Gauss" da gaussiana, mostrando que é o único modelo absolutamente contínuo em que a média empírica é o estimador de verosimilhança máxima do parâmetro de localização.

O primeiro daqueles factos indica claramente que a studentização, entendida no sentido estrito de divisão de uma variável que resulta de centrar um estimador da média por uma função de um estimador da variância para obter uma variável studentizada fulcral, fora de um contexto estritamente gaussiano é um

problema árduo (e o exemplo do resultado de Perlo (1933) de studentização no caso de uma amostra uniforme de tamanho 3 não deixa a esse respeito dúvidas).

O segundo facto apontado mostra que afinal a média não deve ser, afinal, o foco da nossa investigação quando trabalhamos com modelos diferentes do gaussiano.

4.1 Independência de $\overline{X}$ e S^2 — uma caracterização da gaussiana

A expressão da função densidade de probabilidade conjunta de vectores gaussianos permite estabelecer de forma muito simples que correlação nula arrasta independência. É por isso simples exercício mostrar, calculando

$$\operatorname{cov}\left(\overline{X}, X_k - \overline{X}\right) = \frac{1}{n}\sum_{j=1}^{n}\operatorname{cov}(X_j, X_k) - \operatorname{cov}(\overline{X}, \overline{X}) =$$

$$= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0,$$

que em populações gaussianas $\overline{X}$ e $\left(X_1 - \overline{X}, \ldots, X_n - \overline{X}\right)$ são independentes, e portanto também $\overline{X}$ e $\left((X_1 - \overline{X})^2, \ldots, (X_n - \overline{X})^2\right)$ são independentes. Consequentemente, nas populações gaussianas $\overline{X}$ e $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2$ são independentes.

A independência entre $\overline{X}$ e S^2 é no entanto característica das populações gaussianas. É uma consequência do teorema de

Darmois-Skitovich, que afirma que duas combinações lineares de réplicas indepentes de X, $\alpha_1 X_1 + \cdots + \alpha_n X_n$ e $\beta_1 X_1 + \cdots + \beta_n X_n$ são independentes, com $\alpha_k \neq 0$ e $\beta_k \neq 0$, se e só se a variável X for gaussiana.

Provamos uma versão simplificada deste resultado.

É imediato que no caso de amostras de dimensão 2 se tem $\overline{X}_2 = \frac{X_1+X_2}{2}$ e $S_2^2 = \frac{\left(X_1-X_2\right)^2}{2}$, tornando evidente o que o enunciado abaixo tem a ver com independência de $\overline{X}$ e S^2.

Teorema 4.1.1. *Seja X uma população com valor médio μ e variância σ^2, simétrica em torno de μ, X_1 e X_2 réplicas independentes de X.*

$X \frown$ Gaussiana(μ, σ) se e só se $X_1 + X_2$ e $X_1 - X_2$ forem independentes.

Demonstração:

Seja $Y = X - \mu$, simétrica em torno de 0, com valor médio 0 e variância σ^2. A sua função característica é

$$\varphi_Y(t) = \int_{-\infty}^{+\infty} \cos(ty)\ \mathrm{d}F_Y(y),$$

e $\varphi_Y(0) = 1$, $\varphi_Y'(0) = 0$ e $\varphi_Y''(0) = -\sigma^2$.

Se Y_1 e Y_2 forem réplicas independentes de Y, e $Y_1 + Y_2$ e $Y_1 - Y_2$ forem independentes, de $Y_1 = \frac{\left(Y_1+Y_2\right)+\left(Y_1-Y_2\right)}{2}$ da simetria de Y em torno de 0 conclui-se que

$$\varphi_Y(t) = \left[\varphi_Y\left(\frac{t}{2}\right)\right]^4 = \left[\varphi_Y\left(\frac{\frac{t}{2}}{2}\right)\right]^{4^2} = \cdots = \left[\varphi_Y\left(\frac{t}{2^k}\right)\right]^{4^k},\ t \in \mathbb{R},$$

e consequentemente

$$\varphi_{_Y}\left(\frac{t}{2^k}\right) = \left[\,\varphi_{_Y}(t)\right]^{\frac{1}{4^k}}.$$

Assim, $1 = \varphi_{_Y}(0) = \lim\limits_{k\to\infty} \varphi_{_Y}\left(\frac{t}{2^k}\right) = \lim\limits_{k\to\infty} \left[\,\varphi_{_Y}(t)\right]^{\frac{1}{4^k}}$, e consequentemente, para todo $t \in \mathbb{R}$, $\varphi_{_Y}(t) > 0$, e podemos definir

$$\chi(t) = \ln\, \varphi_{_Y}(t),$$

para a qual

$$\chi(0) = \ln\, \varphi_{_Y}(0) = \ln 1 = 0,$$

$$\chi'(t) = \frac{\varphi_{_Y}^{'}(t)}{\varphi_{_Y}(t)} \Longrightarrow \chi'(0) = 0,$$

e

$$\chi''(t) = \frac{\varphi_{_Y}(t)\,\varphi_{_Y}^{''}(t) - \left[\varphi_{_Y}^{'}(t)\right]^2}{\left[\varphi_{_Y}(t)\right]^2} \Longrightarrow \chi''(0) = \varphi_{_Y}^{''}(0) = -\sigma^2.$$

Aplicando duas vezes a regra de l'Hôpital,

$$\lim_{t\to 0} \frac{\chi_{_Y}(t)}{t^2} = \lim_{t\to 0} \frac{\chi_{_Y}^{'}(t)}{2t} = \lim_{t\to 0} \frac{\chi_{_Y}^{''}(t)}{2} = -\frac{\sigma^2}{2}.$$

A equação funcional $\varphi_{_Y}(t) = \left[\varphi_{_Y}\left(\frac{t}{2}\right)\right]^4$, $t \in \mathbb{R}$, mostra que $\chi(t) = 4\,\chi\left(\frac{t}{2}\right)$, $t \in \mathbb{R}$, e consequentemente,

$$\frac{\chi\,(t)}{t^2} = \frac{\chi\left(\frac{t}{2}\right)}{\left(\frac{t}{2}\right)^2},\ t \in \mathbb{R}.$$

Mas então a função $\psi(t) = \frac{\chi(t)}{t^2}$ é tal que

$$\forall\, t \neq 0,\ \psi(t) = \psi\left(\frac{t}{2}\right) = \psi\left(\frac{t}{4}\right) = \cdots = \psi\left(\frac{t}{2^n}\right) \xrightarrow[n\to\infty]{} -\frac{\sigma^2}{2}\,.$$

Consequentemente $\chi(t) = -\frac{\sigma^2}{2}\,t^2$ e daí $\varphi_{_Y}(t) = \mathrm{e}^{-\frac{\sigma^2 t^2}{2}}$.

Pelo teorema da unicidade das funções características conclui-se então que $Y \frown \mathit{Gaussiana}\,(0,\,\sigma)$.

□

4.2 A média empírica é estimador de verosimilhança máxima do parâmetro de localização?

Quase de passagem, Gauss (1809), apresenta uma dedução que mostra que a lei "normal" tem uma posição tão ímpar em Inferência quanto em Probabilidade.

Queremos estimar, por verosimilhança máxima, um parâmetro de localização λ de uma população absolutamente contínua X, cuja função densidade de probabilidade se denota f. Como é hábito $\widehat{\lambda}$ denota o estimador de verosimilhança máxima de λ.

Em que condições $\widehat{\lambda} = \overline{X}$?

Sabemos que se X for gaussiana tal é verdade. Por outro lado, no caso geral, fixada uma amostra $(x_{_1}, \ldots, x_{_n})$, $\widehat{\lambda}$ satisfaz

a equação

$$\sum_{k=1}^{n} g'(\widehat{\lambda} - x_{_k}) = 0,$$

onde denotámos $\ln[f(x_{_k} \mid \lambda)] = g(\lambda - x_{_k})$.

Analisando o caso especial $x_{_1} = \cdots = x_{_{n-1}} = 0,\ x_{_n} = nu$, a condição $\hat{\lambda} = \overline{x} = u$ implica então

$$(n-1)g'(u) + g'((1-n)u) = 0, \quad n = 2, 3, \ldots,$$

mostrando o caso especial $n = 2$ que a função g' tem que ser ímpar. Portanto,

$$g'(nu) = ng'(u),\ u \in \mathbb{R},\ n = 2, 3, \ldots$$

Segue-se que g' é uma função linear, $g'(u) = Cu \implies g(u) = \frac{1}{2}\, C\, u^2 + b$. Da condição $\mathbb{P}(\mathbb{R}) = 1$ obtém-se então

$$f(x \mid \lambda) = \sqrt{\frac{\alpha}{2\,\pi}}\ \mathrm{e}^{-\frac{1}{2}\,\alpha\,(x-\lambda)^2}\ \mathrm{I}_{\mathbb{R}}(x), \quad \alpha > 0.$$

Assim, contrariamente ao que parece ser uma convicção generalizada, mas infundada, a única família contínua para a qual o estimador de verosimilhança máxima do parâmetro de localização λ é $\overline{X}$ é a família das gaussianas, podendo o parâmetro de escala variar livremente em $\mathbb{R}^+$.

Parte II

Fugindo à Gaussiana: Studentização e Análise de Escala em Populações Não Gaussianas

Capítulo 5

Localização e Escala, Studentização Interna e Externa, Suficiência

O problema da studentização em populações não gaussianas começou a ser estudado nos anos 30 do século passado. Como já foi apontado, há imediatamente dificuldades acrescidas, por as margens do par $(\overline{X}, S^2)$ serem dependentes em qualquer população não gaussiana, como vimos no capítulo anterior.

A estrutura de dependência pode ser conhecida, e mesmo assim o cálculo para deduzir a função densidade de probabilidade de

$$T_{n-1} = \sqrt{n(n-1)}\,\frac{\overline{X}_n}{\sqrt{SS_n}}\,,$$

onde SS_n é a "*Sum of Squares*" $\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2 = (n-1)S^2$,

ser extremamente complexo, como mostra o trabalho de Perlo (1933) sobre T_2 em populações uniformes, veja-se no Apêndice A o aspecto tremendo da correspondente função densidade de probabilidade.

Correndo embora o risco de cometer injustiças, por desconhecimento, comentamos uma breve selecção de trabalhos que nos parecem pioneiros e importantes nesta área:

- Hotelling (1961) — em que é apresentado mais explicitamente a interpretação geométrica que já antes servira a Rietz (1939) para o cálculo da função densidade de probabilidade de T_1, e as dificuldades inerentes à studentização interna são torneadas deduzindo aproximações para as caudas.

- Efron (1969) — em que a representação em $\mathbb{R}^n$ anteriormente referida é ardilosamente explorada para reabordar todo o problema da studentização, e em particular se mostra que $T_n \overset{\mathrm{d}}{=} t_n$ sempre que haja simetria radial, e que a quase-simetria aumenta consideravelmente a qualidade das aproximações.

- Logan, Mallows, Rice and Shepp (1973) — em que é explicitamente questionado o recurso acrítico ao teorema limite central, propondo "estatísticas auto-normalizadas"

$$T^* = \frac{\sum_{i=1}^{n} X_i}{\left(\sum_{i=1}^{n} |X_i|^{\alpha} \right)^{\frac{1}{\alpha}}}$$

particularmente interessantes quando a distribuição parente está no domínio de atracção (no que refere o semigrupo de somas de variáveis aleatórias) de uma variável aleatória estável com índice $\alpha \in (0, 2)$.

- Margolin (1977), que no caso de uma parente exponencial relacionou a função densidade de probabilidade da estatística studentizada internamente com a inversa de uma transformada de Laplace, obtendo assim expressões não triviais.

  Já antes, Steutel (1967) tinha relacionado a divisão de um intervalo por pontos distribuidos ao acaso (isto é, com distribuição uniforme em $(0, t)$; e como se sabe $-\ln(U)$ tem distribuição exponencial) com transformadas de Laplace inversas mas sem qualquer referência à studentização.

  Gomes and Pestana (1982) abordaram a problemática geral de obtenção de funções densidade de probabilidade de estatísticas studentizadas internamente á custa de inversas de transformadas integrais (que, como Hirschmam and Widder (1955) discutiram, são em geral casos especiais da transformação de convolução, para que apresentam um "método de *kernel*" geral para inversão de transformadas integrais).

  No caso de parente uniforme, obtêm-se casos especiais da "transformada integral beta" introduzida por Pestana (1978), que para ela derivou uma fórmula de inversão.

- Tukey (1977) e Mosteller and Tukey (1977) — em que a estatística é abordada com o objectivo de "deixar os dados falar por si mesmos", sem ideias preconcebidas sobre a distribuição parente, em geral desconhecida.

> Nesta perspectiva de *análise exploratória*, aqueles autores e seus seguidores (consultar a bibliografia referida em Hoaglin, Mosteller e Tukey (1981) e nomeadamente a que corresponde ao capítulo XI desta obra) propuseram inúmeros estimadores de localização (Goodall, 1981) e de escala (Iglewicz, 1981) — estes restritos a populações simétricas — cuja "tri-eficiência" é avaliada em geral analisando o seu comportamento para pequenas amostras no caso da distribuição ser Gaussiana (caudas neutras), situação amostral perturbada (caudas moderadamente pesadas) e dividida (caudas muito pesadas).
>
> A terminologia adoptada é a de Hoaglin, Mosteller and Tukey (1981), nomeadamente nos capítulos XI e XII.

No já atrás comentado trabalho de Efron (1969), na esteira de um trabalho inovador de Hotelling (1961), mostra-se que a simetria da população parente tinha um efeito regularizador insuspeitado. A interpretação geométrica que aqueles trabalhos dão à estatística t de Student foi inspiradora, e deu novo fôlego à investigação do problema — mas com resultados de certo modo decepcionantes, sendo porventura o passo mais importante o reconhecimento de que em muitos casos valia a pena perder informação, sendo a melhor aposta dividir a amostra numa subamostra $(X_1, \ldots, X_\nu)$ para obter um estimador do valor médio, e numa outra subamostra *externa* $\left(X_{\nu+1}, \ldots, X_n\right)$ para obter um estimador da variância. Daí os conceitos de studentização externa e de studentização interna propostos por David (1981), veja-se adiante a definição 5.1.1

Neste contexto, coloca-se o problema mais geral de usar estimadores de localização cuja distribuição depende de um parâmetro perturbador de escala, e studentizar dividindo por um

estimador desse parâmetro, por forma a obter uma variável cuja distribuição já não depende do referido parâmetro perturbador.

5.1 Parâmetros de localização e de escala, studentização interna e studentização externa

Convém então recordar que genericamente $\lambda \in \mathbb{R}$ é um parâmetro de localização da variável aleatória X se a distribuição de $X - \lambda$ não depender de λ (foi "centrada", passando a ter localização 0). Assim, dada uma variável X, $\{X + \lambda\colon\ \lambda \in \mathbb{R}\}$ é uma família de localização (geralmente usa-se X centrada).

Por outro lado, $\delta > 0$ é um parâmetro de escala da variável aleatória X se a distribuição de $\frac{X}{\delta}$ não depender de δ (foi "reduzida", passando a ter escala unitária). Assim, dada uma variável X, $\{\delta X\colon\ \delta > 0\}$ é uma família de escala.

Uma família $\{\delta X + \lambda\colon\ \lambda \in \mathbb{R},\, \delta > 0\}$ é uma família de localização e escala (noutros contextos é também designada como um "tipo de Khinchine"). Em geral naquela representação X é a "variável padrão" da família, isto é a variável que está centrada em 0 e tem escala 1.

Adiante investigamos, no contexto de famílias exponenciais de Darmois–Koopman–Pitman, as situações em que uma estatística captura toda a informação contida na amostra sobre um parâmetro de localização, ou sobre um parâmetro de escala; e as situações em que um par de estatísticas consegue capturar toda a informação sobre um par de parâmetros, localização e escala.

Neste último problema, veremos que, admitindo as condições de regularidade de Koopman (1936), há apenas três famílias exponenciais com essa caracterização

1. com supporte real, a família gaussiana;
2. com suporte numa semi-recta, a família das exponenciais;
3. com suporte num segmento de $\mathbb{R}$, a família das uniformes.

Abordamos ainda a existência de um estimador suficiente da localização, quando os outros parâmetros são conhecidos, e a existência de um estimador suficiente da escala, também quando os outros parâmetros são conhecidos.

A dificuldade da studentização em populações não gaussianas, no sentido restrito que até agora usámos, advém do facto de $\overline{X}_n$ e S_n^2 serem, variáveis aleatórias dependentes, e terem uma função densidade de probabilidade conjunta difícil de determinar.

Uma forma simples de obviar este problema é considerar $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ como sendo constituída por duas subamostras disjuntas,

$$\boldsymbol{X}_\nu = (X_1, \ldots, X_\nu) \text{ e } \boldsymbol{X}_{n-\nu} = \left(X_{\nu+1}, \ldots, X_n\right)$$

e

$$T^*_{n-\nu-1} = \sqrt{\nu}\ \frac{\overline{X}_\nu - \mu}{S_{n-\nu}}$$

em que agora $\overline{X}_\nu$ e $S_{n-\nu}$, obtidas de vectores independentes, são variáveis aleatórias independentes.

Esta "studentização externa" tem a vantagem de a função densidade de probabilidade conjunta de $\left(\overline{X}_\nu, S_{n-\nu}\right)$ ser o produto

das funções densidade de probabilidade das marginais $\overline{X}_{\nu}$ e $S_{n-\nu}$, mas também tem a enorme desvantagem de usar a informação disponível de forma pouco eficiente.

Assim, apesar do atractivo acima referido, este tipo de studentização externa nunca suscitou grande interesse por parte dos estatísticos — que a ela apenas recorrem em situações particularmente complexas — e os desenvolvimentos teóricos que têm sido bem aceites referem-se a "studentização interna" em que $\overline{X}_{n}$ e S_{n} são obtidos ambos da globalidade da amostra, ainda que na generalidade das situações amostrais $\overline{X}_{n}$ e S_{n} sejam variáveis aleatórias dependentes e de função densidade de probabilidade conjunta exacta desconhecida. Todos estes desenvolvimentos levaram á definição de conceitos mais gerais de studentização, que explicitamos:

Definição 5.1.1.

Seja $\boldsymbol{X}_{n} = (X_{1}, \ldots, X_{n})$ *uma amostra de dimensão* n *extraída de uma população com função de distribuição* F_{X}*, dependente de um parâmetro de localização* λ*, e seja* $W = g(\boldsymbol{X}_{n})$ *um estimador desse parâmetro cuja função de distribuição depende de um parâmetro de escala* θ*, de que* $S = h(\boldsymbol{X}_{n})$ *é estimador. Dizemos que*

$$T^{*} = \frac{W}{S}$$

é uma estatística studentizada se a função de distribuição de T^{*} *não depender do parâmetro perturbador* θ*.*

No caso de W *e* S *serem independentes diremos que a studentização é externa, e no caso de* W *e* S *serem dependentes diremos que a studentização é interna.*

Anote-se que a studentização clássica, em populações gaussianas, é uma studentização "externa" eficiente: a independência dos estimadores de localização e de escala é uma característica da população, e não um resultado adventício de usarmos ineficientemente a informação disponível.

Adiante encontraremos outros exemplos de studentização externa — por exemplo, as populações $Exponencial(\lambda, \delta)$ caracterizam-se pela independência dos *spacings*, e consequentemente o estimador $\hat{\lambda} = X_{1:n}$ da localização e estimadores $g(X_{k:n} - X_{j:n})$ da escala que sejam funções dos *spacings* remanescentes são variáveis aleatórias independentes, permitindo assim o desenvolvimento de uma studentização "externa" e de uma análise da escala que, no que respeita à dedução dos resultados, são formalmente muito semelhantes à studentização e análise da variância em populações gaussianas.

5.2 Famílias exponenciais e estimadores suficientes dos parâmetros de localização e escala

A densidade (função densidade de probabilidade no caso absolutamente contínuo ou função massa de probabilidade no caso discreto) de uma família exponencial de variáveis aleatórias parametrizada por $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)'$, com representação

$$f_X(x|\boldsymbol{\theta}) = \exp\left[\sum_{k=1}^{s} \eta_k(\boldsymbol{\theta})\, T_k(x) - A(\boldsymbol{\theta}) + B(x)\right]$$

goza de um estatuto privilegiado em inferência estatística.

O conceito de família exponencial foi introduzido quase simultaneamente por Darmois (1935), Pitman (1936) e Koopman (1936), que mostraram que se considerarmos amostras aleatórias $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ de dimensão crescente de uma população X parametrizada por $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ e cujo suporte não depende do parâmetro $\boldsymbol{\theta}$, só no caso de famílias exponenciais existe uma estatística suficiente $T(X_1, \ldots, X_n)$ cujo número de componentes escalares não aumenta quando o tamanho da amostra aumenta.

Há conveniência em reparametrizar em termos do *parâmetro natural* $\boldsymbol{\eta} = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta}))'$.

De facto, ainda que naquela representação $T_k(x)$, $\boldsymbol{\eta}$ e $A(\boldsymbol{\eta})$ pareçam arbitrariamente definidas, elas têm de facto interpretação notável:

- $\boldsymbol{\eta}$ é o parâmetro natural, e o conjunto de valores de $\boldsymbol{\eta}$ para os quais a função densidade é finita (que é sempre um conjunto convexo) é o espaço de parâmetros natural.
- $\boldsymbol{T}$ é um vector de estatísticas suficientes — consequentemente, nas famílias exponenciais existe um vector de estatísticas suficientes cuja dimensão iguala o número de parâmetros a estimar.
- $A(\boldsymbol{\eta})$ é um factor de normalização, e a função geradora de cumulantes da estatística suficiente $T(x)$ é $K_X(\boldsymbol{t}|\boldsymbol{\eta}) = A(\boldsymbol{\eta} + \boldsymbol{t}) - A(\boldsymbol{\eta})$.

Seja X uma variável aleatória, com localização 0 e escala 1, de uma família exponencial, e

$$\left\{X_{\lambda,\delta} = \lambda + \delta X\colon\ \lambda \in \mathbb{R},\ \delta > 0\right\}$$

a família de localização-escala gerada por X. Denotamos $f = f_X$ a função densidade de probabilidade de X. Em que condições existe um par de estimadores suficientes de (λ, δ)?

Vamos responder a esta questão admitindo que se verificam as condições de regularidade de Koopman, isto é que existem as derivadas parciais requeridas em ordem aos parâmetros (o que por sua vez implica a existência das derivadas de f, ou de $\ln f$, até à ordem necessária, em cada caso).

Para desenvolver a questão, vamos necessitar de algumas extensões das equações funcionais de Cauchy.

5.2.1 Extensões das equações de Cauchy

No início o século XX foram estudadas diversas extensões de famosas equações funcionais de Cauchy. (Uma excelente fonte de informação sobre eqações funcionais é o clássico de Aczél (2006); sobre desenvolvimentos da equação funcional de Levi-Civita a que adiante recorremos, consulte-se Baker (1993).)

Entre essas extensões, interessam-nos especialmente as de Pexider (1902), que estabeleceu que as soluções regulares de $f(x+y) = \phi_1(x)\,\phi_2(y)$ e de $g(xy) = \phi_1(x)\,\phi_2(y)$ são respectivamente

$$f(x) = k_1\, \mathrm{e}^{k_2 x} \tag{5.1}$$

e

$$g(x) = k_1\, x^{k_2}, \tag{5.2}$$

e as de Levi-Civita (1913) e de Stäckel (1913), estabelecendo que as soluções regulares de $f(x+y) = \phi_1(x)\,\psi_1(y) + \phi_2(x)\,\psi_2(y)$ e

de $g(xy) = \phi_1(x)\,\psi_1(y) + \phi_2(x)\,\psi_2(y)$ são respectivamente

$$f(x) = k_1\,\mathrm{e}^{k_2 x} + k_3\,\mathrm{e}^{k_4 x} \quad \text{ou} \quad f(x) = (k_1 x + k_2)\,\mathrm{e}^{k_3 x} \tag{5.3}$$

e

$$g(x) = k_1\,x^{k_2} + k_3\,x^{k_4} \quad \text{ou} \quad g(x) = (k_1 \ln(|x|) + k_2)\,x^{k_3 x}. \tag{5.4}$$

Consequentemente, se considerarmos a equação funcional

$$h(xz + yz) = \phi_1(y, z)\psi_1(x) + \phi_2(y, z)\psi_2(x), \tag{5.5}$$

que para $z = 1$ é a equação funcional com solução (5.3) e para $y = 0$ é a equação funcional com solução (5.4), concluímos que a solução geral tem que ser linear, $h(x) = k_1 x + k_2$.

Tem-se portanto um caso especial

$$h(xz + yz) = \phi(y, z)\psi(x) \Longrightarrow h(x) = k. \tag{5.6}$$

5.2.2 Famílias de localização-escala para as quais existe um par de estimadores suficientes de (λ, δ)

Dispomos agora dos instrumentos necessários para resolver a questão, que tem que ser analisada em três contextos distintos:

1. O suporte de $X_{\lambda,\delta}$ depende de λ e de δ (suporte num segmento de recta).

2. O suporte de $X_{\lambda,\delta}$ depende apenas de λ (suporte numa semi-recta).

3. O suporte de $X_{\lambda,\delta}$ não depende de nenhum dos parâmetros (o suporte é a recta real).

1. O suporte de X é um segmento de recta

$\forall x \in \mathbb{R},\ \lambda \in \mathbb{R},\ \delta > 0, n \geq 3$

$$f_{-\lambda,\frac{1}{\delta}}(x) = \delta\, f(\delta x + \delta\lambda) = g(x)h(\lambda,\delta),$$

(observe-se que usamos $-\lambda$ e $\frac{1}{\delta}$ — isto é $\frac{X-\lambda}{\delta}$ — para na representação formal aparecer como argumento $\delta(x+\lambda)$ como na equação (5.6)) e consequentemente por (5.6) tem-se $f(x) = c$ — por outras palavras, trata-se da família de localização-escala das uniformes, extraindo o par de estatísticas $(x_{1:n}, x_{n:n})$ toda a informação da amostra relevante para estimar (λ,δ) — os estimadores de verosimilhança máxima são $\widehat{\lambda} = X_{1:n}$, $\widehat{\delta} = (X_{n:n} - X_{1:n})$.

2. O suporte de X é uma semi-recta

$\forall x \in \mathbb{R},\ \lambda \in \mathbb{R},\ \delta > 0, n \geq 3$ tem-se

$$\ln\left[f_{-\lambda,\frac{1}{\delta}}(x)\right] = u^*(\lambda,\delta)\, v(x) + A(x) + B^*(\lambda,\delta) \tag{5.7}$$

e derivando em ordem a x

$$\delta \frac{\mathrm{d}}{\mathrm{d}x} \ln[f(\delta x + \delta\lambda)] = u^*(\lambda,\delta)\, v'(x) + A'(x).$$

Então, com $u_1(\lambda,\delta) = \frac{u^*(\lambda,\delta)}{\delta}$, $u_2(\lambda,\delta) = \frac{1}{\delta}$, $v_1(x) = v'(x)$, $v_2(x) = A'(x)$,

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln[f(\delta x + \delta\lambda)] = u_1(\lambda,\delta)\, v_1(x) + u_2(\lambda,\delta)\, v_2(x),$$

cuja solução, por (5.5), é $\ln[f(x)] = c_1 x + c_2 \implies f(x) = \exp\left\{\frac{a_1}{2}x^2 + a_2 x + a_3\right\}$, onde devido a (5.7) $a_1 = 0$ e c_2, c_3 são constantes sujeitas a $\int_{\boldsymbol{S}} f(x)\,\mathrm{d}x = 1$.

Consequentemente, X é exponencial padrão na semi-recta negativa ou a exponencial padrão na semi-recta positiva. No primeiro caso, $(x_{n:n}, \sum x_k)$ é o par de estatísticas suficientes para estimar (λ, δ) (estimadores de verosimilhança máxima: $\widehat{\lambda} = X_{n:n},\ \widehat{\delta} = X_{n:n} - \overline{X}$); no segundo caso, $(x_{1:n}, \sum x_k)$ é o par de estatísticas suficientes para a localização e escala (estimadores de verosimilhança máxima: $\widehat{\lambda} = X_{1:n},\ \widehat{\delta} = \overline{X} - X_{1:n}$).

3. O suporte de X é $\mathbb{R}$

$\forall x \in \mathbb{R},\ \lambda \in \mathbb{R},\ \delta > 0, n \geq 3$ tem-se

$$\ln\left[f_{-\lambda,\frac{1}{\delta}}(x)\right] = u_1^*(\lambda,\delta)\, v_1(x) + u_2^*(\lambda,\delta)\, v_2(x) + A(x) + B^*(\lambda,\delta), \tag{5.8}$$

e derivando em ordem a x e depois em ordem a λ obtém-se

$$\frac{\partial^2}{\partial x^2} \ln\left[f(\delta x + \delta\lambda)\right] = \frac{\frac{\partial}{\partial\lambda}\, u_1^*(\lambda,\delta)}{\delta^2}\, v_1'(x) + \frac{\frac{\partial}{\partial\lambda}\, u_2^*(\lambda,\delta)}{\delta^2}\, v_2'(x),$$

que é a equação funcional (5.5), com solução

$$\frac{\partial^2}{\partial x^2} \ln\left[f(x)\right] = c_1 x + c_2 \implies f(x) = \exp\left[\frac{c_1}{6}\, x^3 + \frac{c_2}{2}\, x^2 + c_3 x + c_4\right],$$

onde as constantes c_1, c_2, c_3, c_4 estão sujeitas a $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$, o que implica que $c_1 = 0$ e $c_2 < 0$.

Trata-se portanto da família de localização-escala gaussiana, o par de estatísticas suficientes para estimar (λ, δ) — que nesta família é usualmente denotado (μ, σ) — é $(\sum x_k, \sum x_k^2)$.

5.2.3 Famílias de localização e famílias de escala; estatísticas suficientes mínimas

O problema de identificar quais as famílias de localização

$$\{X_{\lambda} = \lambda + X\colon\ \lambda \in \mathbb{R}\}$$

para as quais uma única estatística suficiente estima λ, e de identificar as famílias de escala

$$\{X_{\delta} = \delta X\colon\ \delta > 0\}$$

para as quais uma única estatística suficiente estima δ tem um tratamento análogo.

1. No caso regular de o suporte não depender dos parâmetros desconhecidos, recorrendo às equações de Pexider com soluções (5.1) e (5.2), respectivamente, tem-se:

   (a) No que se refere a famílias de localização, encontram-se como soluções a família gaussiana e as famílias Gumbel de máximos e Gumbel de mínimos — neste último caso, sendo δ uma escala conhecida, a estatística suficiente é $\sum \exp(-\frac{|x_k|}{\delta})$.

   (b) No que se refere a famílias de escala, existem como soluções as famílias de cujos representantes padrões têm densidades de probabilidade

      i. denotando c_1, $c_2 \neq 0$, c_3 parâmetros sujeitos a $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$,

$$f(x) = c_1\ \exp\left[-(x)^{c_2} + c_3\ \ln\,|x|\right],$$

que contém como caso especial a família gaussiana (com localização conhecida) e a famílias das gamas com parâmetro de forma conhecido; e

ii. sendo c_1, $c_2 < 0$, e c_3 parâmetros, sujeitos à condição $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$,

$$f(x) = c_1 \exp\left[c_2 \left(\ln |x|\right)^2 + c_3 \ln |x|\right],$$

2. No caso não regular:

   (a) No que se refere a famílias de localização, aquelas cujos elementos padrões são $X \frown Exponencial(\delta)$ (com parâmetro de escala δ conhecido), ou $-X$.

   (b) No que se refere a famílias de escala, aquelas cujos elementos padrões são

   - $X \frown Beta(p, 1)$, (com parâmetro de forma p conhecido), ou $-X$,
   - $X \frown Pareto(\alpha)$ (com parâmetro de forma α conhecido), ou $-X$,
   - variáveis do mesmo tipo de Khinchine que X, sendo a densidade de X da forma $f(x) = c\,x^{\alpha}\,\mathbb{I}_{a,b}$, $a < 0 < b$.

Deixamos ao cuidado do leitor completar esta informação, detalhando as condições de regularidade invocadas e as demonstrações, e explicitando a estatística suficiente em questão em cada caso.

5.3 Caracterização da exponencial pela independência dos espaçamentos

No Capítulo 7 usaremos a independência dos espaçamentos da exponencial (cf. Pestana e Velos, 2010, pp. 901–903) para obter diversos resultados interessantes. Infelizmente essa independência dos *spacings* não pode usar-se em outros modelos, porque é uma propriedade tão caracterizadora da exponencial quanto a independência de $\overline{X}$ e S^2 é caracterizadora da gaussiana.

A equação funcional de Pexider com solução (5.1) pode ser usada para uma demonstrar de forma elementar que os *spacings* consecutivos de um modelo absolutamente contínuo são independentes se e só se for o modelo exponencial.

Tendo invocado a referência acima, apenas temos que mostrar que se os *spacings* consecutivos forem independentes então o modelo é exponencial.

Considere-se então que X_1, X_2 é uma amostra de dimensão 2 de uma variável aleatória com função densidade de probabilidade f_X. Se $X_{1:2}$ e $X_{2:2} - X_{1:2}$ forem independentes, a sua função densidade de probabilidade conjunta factoriza, isto é

$$f_{X_{1:2},X_{2:2}-X_{1:2}}(y_1, y_2) = f_{X_{1:2}}(y_1) \times f_{X_{2:2}-X_{1:2}}(y_2).$$

Mas por outro lado

$$f_{X_{1:2},X_{2:2}-X_{1:2}}(y_1, y_2) = 2\, f_X(y_1\, f_X(y_1 + y_2).$$

Consequentemente,

$$f_X(y_1 + y_2) = \frac{f_{X_{1:2}}(y_1)}{2\, f_X(y_1)} \times f_{X_{2:2}-X_{1:2}}(y_2),$$

concluindo-se então que $f_X = k_1 \, e^{k_2 x}$; como f_X é densidade de probabilidade, $k_1 > 0$ e $k_2 = -k_1$, ou seja é a densidade de probabilidade de $X \frown Exponencial\left(\frac{1}{k_1}\right)$ no caso de suporte na semi-recta $(0, \infty)$. No caso de suporte na semi-recta negativa, temos a simetrização de uma exponencial usual.

5.4 Studentização e análise de escala em populações não gaussianas

Do atrás exposto, facilmente se percebe que os modelos não gaussianos que proporcionam resultados interessantes são o uniforme e o exponencial. Se nos contentarmos com aproximações, num contexto clássico restrito de studentização da média, assumir simetria da população parente proporciona também resultados interessantes, ainda que de interesse prático limitado. Isso justifica a nossa escolha:

No que se segue relatamos resultados de Pestana e Rocha (1993, 1995a, 1995b), Brilhante *et al.* (1996), Rocha (1995) e Brilhante (1996a, 1996b), Brilhante *et al.* (2009), que parcialmente resolvem, exacta ou aproximadamente, studentização e análise da escala em populações não gaussianas, nomeadamente:

- Explorando as ideias de Efron, de que para população parente simétrica há melhores aproximações. Rocha (1995), com uma aplicação engenhosa do teorema do valor médio, consegue expressões aproximadas para a função densidade de probabilidade de T_{n-1}.

  O princípio da abordagem é simples: exprime-se $(\overline{X}_{n+1}, SS_{n+1}, \overline{X}_n - X_{n+1})$ como função de $(\overline{X}_n, SS_n, X_{n+1})$,

conseguindo assim uma expressão recursiva para $f_{(\overline{X}_n, SS_n)}$. A utilização do teorema do valor médio leva a um resultado surpreendente: consegue-se uma expressão de aproximação em que pontos especiais do intervalo $(-1, 1)$ — os mesmos, qualquer que seja a população parente, desde que seja simétrica e tenha propriedades de regularidade muito fracas — têm um papel especial. Chamamos a atenção para a similitude com o que acontece no método de integração numérica de Gauss, mas não procuramos estabelecer qualquer nexo com famílias de polinómios ortogonais.

- Pestana e Rocha (1995a) calcularam a função densidade de variáveis studentizadas no caso de população parente exponencial ou uniforme, usando transformadas integrais inversas e cálculo fraccionário. Na base do seu trabalho está a utilização do teorema de Basu (1955) sobre estatísticas ancilares e independência.

- Brilhante *et al.* (1996) e Brilhante (1996a, 1996b) apresentaram expressões exactas para funções densidades de diversas variáveis studentizadas em populações exponenciais.

- Pestana e Rocha (1993) abordaram também uma análise de escala em populações exponenciais, usando o facto de nestas populações os espaçamentos (*spacings*) consecutivos serem independentes. Esta propriedade — uma caracterização da exponencial, tal como a independência de valor médio e variância empíricas é uma caracterização da gaussiana — permite-lhes usar o cerne das ideias de Fisher, tal como as expusemos no Capítulo 2, e obter resultados que, não sendo porventura importantes, são no entanto curiosos.

  Posteriormente, Pestana *et al.* (1999) retomaram a questão,

e por simples transformação obtiveram resultados para Paretos generalizadas. Mais recentemente, Brilhante *et al.* (2009) desenvolveram uma *ANOSp — ANalysis Of Spacings* — em populações exponenciais, capitalizando na caracterização da exponencial em termos da independência dos espaçamentos.

Capítulo 6

Studentização no Caso de População Parente Uniforme

A distribuição uniforme ocupa lugar de grande destaque na investigação estatística. O papel preponderante do modelo gaussiano, que não era disputado até aos anos sessenta, tem vindo a ser questionado nas abordagens robustas e/ou resistentes (Iglewicz, 1981, propõe um teste a aplicar rotineiramente, convicto de que as colecções de dados que se encontram na prática não são gaussianas); e o facto de $F_X^{\leftarrow}(U) \frown F_X$, onde U denota a uniforme em $(0,1)$ e $F_X^{\leftarrow}$ a inversa generalizada de uma distribuição F_X arbitrária, conjuntamente com o papel cada vez mais activo da estatística experimental, levam a que a distribuição uniforme ombreie com a gaussiana na moderna investigação es-

tatística, nomeadamente nas novas áreas da estatística computacional.

Assim, ao menos conceptualmente, é sempre possível transformar dados de qualquer distribuição arbitrária conhecida em dados uniformes, inferir nesta situação e "destransformar". Do mesmo modo, é conceptualmente possível transformar dados de qualquer distribuição F em dados com distribuição G, usando a transformação $G^{\leftarrow} \circ F$ ou algoritmos como o de Box-Muller, e assim em particular em dados gaussianos. O problema advém no entanto do facto de a distribuição dos dados ser, na prática, um ajustamento da função de distribuição empírica a um modelo usual, por métodos que podem ir dos mais ingénuos — tradição, por exemplo — aos mais sofisticados.

Tukey (1957), Box and Cox (1964) e os seus seguidores têm proposto em alternativa uma bateria de transformações de dados com o objectivo de os regularizar, para fins específicos, tais como simetrizar, homogeneizar dispersões, linearizar relações, etc. Assim, parece-nos porventura mais frutífero em muitas situações transformar os dados com vista a simetrizá-los, pois como Efron (1969) demonstrou a quase-simetria leva a que a função de distribuição de T_n seja bem aproximada pela de t_n; evidentemente, há que ter em conta que a transformação-potência que quase-simetriza foi obtida à custa de expansões em série de Taylor em torno da mediana, pelo que alcança os seus objectivos na parte central da distribuição, podendo o mesmo não acontecer nas caudas. Ora a inferência estatística recorre sobretudo às caudas extremas; e um outro factor a ter em conta na aproximação é, como já referimos, o peso das caudas da distribuição parente.

Depois desta digressão, que não podíamos omitir, retomemos então o problema da studentização interna no caso da dis-

tribuição parente uniforme.

Sejam então $X_1, \ldots, X_n$ variáveis aleatórias i.i.d., $X_i \frown Uniforme(\theta_1, \theta_1 + \theta)$. A variável

$$T_{n-1} = \frac{\sqrt{n}\,(\overline{X}_n - \mu)}{S_n}$$

foi das primeiras a ser investigada, quando a aproximação por t_{n-1} começou a ser questionada. (De facto as caudas da uniforme não são neutras, são caudas bem mais leves do que as gaussianas). A função densidade de probabilidade de T_1 foi obtida por Rider (1931), e a de T_2 por Perlo (1933); uma e outra estão registadas no Apêndice A.

Ao passar da geometria de $\mathbb{R}^2$ para a geometria de $\mathbb{R}^3$ a complexidade aumenta de tal forma, e a expressão analítica de T_2 é tal que o estudo da função densidade de probabilidade de T_n para $n \geq 3$ é pouco apelativa. Abordemos então a questão na perspectiva do conceito geral de studentização interna.

6.1 Studentização do mínimo pela amplitude

Quando $X_i \frown Uniforme(\theta_1, \theta_1 + \theta)$, segue-se que

$$\hat{\theta}_1 = \min\{X_1, \ldots, X_n\} = X_{1:n}$$

$$\hat{\theta} = R = \max\{X_1, \ldots, X_n\} - \min\{X_1, \ldots, X_n\} = X_{n:n} - X_{1:n}.$$

Consequentemente, em lugar de considerar os parâmetros de localização e escala μ e σ, estimados respectivamente por $\overline{X}_n$

e S_n, podemos antes considerar θ_1 estimado por $X_{1:n}$, como parâmetro de localização, e θ, estimado por $(X_{n:n} - X_{1:n})$, como parâmetro de escala.

Outras alternativas viáveis para a localização são μ estimado por $\frac{X_{1:n}+X_{n:n}}{2}$ ou por $X_{\frac{n+1}{2}:n}$ (com a interpretação interpolativa quando $\frac{n+1}{2}$ é da forma $k + \frac{1}{2}$), que adiante discutiremos. No que respeita a dispersão, o candidato mais promissor, quando se pretende robustez, é decerto a dispersão quartal

$$d_F = X_{\frac{2n+1-\left[\frac{n+1}{2}\right]}{2}:n} - X_{\frac{\left[\frac{n+1}{2}\right]+1}{2}:n}$$

onde $[k]$ é o maior inteiro não superior a k, e se adopta a interpretação interpolativa já atrás referida. De facto, a amplitude $R = X_{n:n} - X_{1:n}$ tem limite de ruptura 0, enquanto o limite de ruptura de d_F é 0.25 (Goodall, 1981)[(1)].

Gomes e Pestana (1982) consideraram a variável studentizada

$$T^*_{n-1} = \frac{X_{1:n} - \theta_1}{X_{n:n} - X_{1:n}}$$

e deduziram a expressão da função densidade de probabilidade de T^*_{n-1} observando que a função densidade de probabilidade de $X_{1:n} - \theta_1$ é uma transformada beta da função densidade de probabilidade de T^*_{n-1}; isto é consequência de $X_{n:n} - X_{1:n}$ ter distribuição beta e ser independente de T^*_{n-1} — de facto um célebre teorema de Basu (1955) diz que se $Z = \frac{X}{Y}$, com X e Y dependentes com distribuição envolvendo um parâmetro θ de que Y

[(1)] Convém recordar o conceito preciso do *limite de ruptura de Hampel*: Uma estatística T tem limite de ruptura $\alpha \times 100\%$ se tender para α a proporção $\frac{k}{n}$ de elementos da amostra que podem crescer (ou decrescer) sem T divergir.

é estimador suficiente e completo, e a distribuição de Z não depender de θ, então Z e Y são independentes; nestas condições, a função densidade de probabilidade conjunta de (Z, Y) é o produto das funções densidade de probabilidade marginais, do que decorre que

$$f_{ZY}(t) = \int_{-\infty}^{+\infty} f_Z\left(\frac{t}{y}\right) \frac{f_Y(y)}{|y|} \, \mathrm{d}y,$$

que pode ser interpretada como uma transformação integral de $\frac{1}{|y|} \, f_Z\left(\frac{t}{z}\right)$ com *kernel* f_Y. No caso em estudo, $f_{X_{n:n}-X_{1:n}}$ é uma $Beta(n-1, 2; 0, \theta)$ pelo que teremos uma transformada integral beta.

Assim, como a função densidade de probabilidade de T^*_{n-1} não depende de θ podemos, sem perda de generalidade, trabalhar com parentes uniformes em $(0, 1)$, a função densidade de probabilidade de $X_{1:n}$ é conhecida, o problema reduz-se à inversão da transformada beta, que foi introduzida e estudada em detalhe em Pestana (1978), onde em particular podem ser encontradas fórmulas de inversão.

As expressões gerais de Gomes e Pestana (1978) mostram que nas condições atrás, com $Y \frown Beta(p, q; 0, \theta)$, segue-se que

$$h(x) = B(p, q)\theta x^{1-p} f_X(\theta x) = \int_0^{\frac{1}{x}} (1-2y)^{q-1} \, \mathrm{d}G(y)$$

com $\mathrm{d}G(y) = y^{p-2} f_Z\left(\frac{1}{y}\right) \mathrm{d}y$, uma transformada integral cuja inversa é $G(y) = \sum_{k=0}^{q-1} \frac{(-1)^k}{k!\, y^k} \, h^{(k)}(z)\Big|_{x=\frac{1}{y}}$.

Mais geralmente, se $Y \frown Beta(\mu, \nu; 0, \theta)$, $\mu, \nu > 0$

$$f_{Z}(z) = \frac{\Gamma(\mu)}{\Gamma(\mu+\nu)}\, z^{\mu} D^{\nu} \left[x^{\mu+\nu-2} f_{X}\left(\frac{\theta}{x}\right)\right]\Big|_{x=\frac{1}{z}}$$

onde D^{ν} é a derivada de Riemann-Liouville de ordem ν.

Facilmente se calcula a função densidade de probabilidade de $T^{*}_{n-1} = \frac{X_{1:n}-\theta_1}{X_{n:n}-X_{1:n}}$:

$$f_{T^{*}_{n-1}}(t) = \frac{n-1}{(1+t)^{n}}\, \mathrm{I}_{(0,\infty)}(t),$$

ou seja $T^{*}_{n-1} \frown Pareto(n-1, 0)$.

A expressão do quantil de probabilidade $1 - \frac{p}{2}$ é

$$Q_{n-1;\ 1-\frac{p}{2}} = \left(\frac{2}{p}\right)^{\frac{1}{n-1}} - 1.$$

6.2 Studentização da mediana

Como atrás referimos, a variável studentizada

$$\tilde{T}_{n-1} = \frac{X_{\frac{n+1}{2}:n} - \theta_1}{d_{F}},$$

onde d_{F} denota a dispersão quartal, parece-nos preferível do ponto de vista de robustez. A fim de simplificar a apresentação, consideremos $n = 4k+1$, que leva a que a mediana seja estimada por $X_{2k+1:4k+1}$ e a que $d_{F} = X_{3k+1:4k+1} - X_{k+1:4k+1}$, sem haver necessidade de interpretações interpolativas. Uma vez que $\widetilde{T}_{n-1}$ é

uma variável aleatória centrada com distribuição independente do parâmetro θ, podemos sem perda de generalidade trabalhar com parentes uniformes em $(0,1)$. Uma vez que

$$f_{\left(X_{k+1:n},X_{2k+1:n},X_{3k+1:n}\right)}(x,y,z) \;=$$

$$= \frac{(4k+1)!}{\left[k!\,(k-1)!\right]^2}\,\left[x\,(1-z)\right]^k\,\left[(y-x)\,(z-y)\right]^{k-1}\;\mathrm{I}_A(x,y,z)$$

com $A = \{(x,y,z):\ 0 \le x < y < z \le 1\}$; fazendo a transformação

$$\begin{cases} \tilde{T} = \frac{X_{2k+1:n}}{X_{3k+1:n}-X_{k+1:n}} \\ U = X_{3k+1:n} - X_{k+1:n} \\ V = X_{k+1:n} \end{cases}$$

cuja inversa é

$$\begin{cases} X_{k+1:n} = V \\ X_{2k+1:n} = \tilde{T}U \\ X_{3k+1:n} = U + V \end{cases}$$

com $J = U$, segue-se que

$$f_{(\tilde{T},U,V)}(t,u,v) \;=$$

$$= \frac{(4k+1)!}{\left[k!\,(k-1)!\right]^2}\,\left[v\,(1-u-v)\right]^k\,\left[(tu-v)\,(u+v-tu)\right]^{k-1}\,u\;\mathrm{I}_A(t,u,v)$$

onde $A = \{(t,u,v): \ 0 \leq \nu < tu < u + \nu \leq 1\}$, e consequentemente

$$f_{\tilde{T}}(t) = \begin{cases} \displaystyle\int_0^1 \int_{\frac{\nu}{t}}^{\min\left\{\frac{\nu}{(t-1)}, 1-\nu\right\}} f_{(\tilde{T},U,V)}(t,u,\nu)\ \mathrm{d}u\ \mathrm{d}\nu, & 1 \leq t \leq 2 \\ \displaystyle\int_0^1 \int_{\frac{\nu}{t}}^{\frac{\nu}{(t-1)}} f_{(\tilde{T},U,V)}(t,u,\nu)\ \mathrm{d}u\ \mathrm{d}\nu, & t > 2 \\ \displaystyle\int_0^1 \int_{\frac{\nu}{(t-1)}}^{1-\nu} f_{(\tilde{T},U,V)}(t,u,\nu)\ \mathrm{d}u\ \mathrm{d}\nu, & 0 < t < 1 \end{cases}$$

expressões cuja integração numérica não oferece dúvidas de especial, mas que nos parecem menos satisfatórias por não termos conseguido uma forma explícita, tratável analiticamente.

Infelizmente é óbvio que $X_{3k+1:n} - X_{k+1:n}$ não é suficiente para θ, pelo que o caminho aberto pelo teorema de Basu nos está, neste caso, vedado.

6.3 Studentização estandardizando o mínimo

Assim, como alternativa, propomos a variável studentizada internamente

$$\tau_{n-1} = \frac{X_{1:n} - \theta_1}{\sigma_{X_{1:n}}}$$

a qual embora tenha limite de ruptura zero, tem sobre a estatística T^*_{n-1} de Gomes e Pestana (1982) a vantagem de consi-

derarmos no denominador um estimador de dispersão do estimador de localização em numerador. Esta é a abordagem proposta por Tukey e seus seguidores — veja-se por exemplo a abordagem biponderada à studentização descrita em Hoaglin *et al.* (1981).

Uma vez que os $X_i \frown Uniforme\,(\theta_1, \theta_1 + \theta)$, então

$$f_{X_{1:n}}(x) = \frac{n}{\theta^n}\,(\theta_1 + \theta - x)^{n-1}\ \mathrm{I}_{(\theta_1, \theta_1+\theta)}(x)$$

e segue-se que

$$\mathbb{E}\,[X_{1:n}] = \theta_1 + \frac{\theta}{n+1}; \quad \mathrm{var}\,(X_{1:n}) = \frac{n\,\theta^2}{(n+2)(n+1)^2}$$

e como atrás já referimos, $\hat{\theta} = R_n = X_{n:n} - X_{1:n}$.

Assim, propomos

$$\tau_{n-1} = (n+1)\sqrt{\frac{n+2}{n}}\,\frac{X_{1:n} - \theta_1}{X_{n:n} - X_{1:n}}$$

cuja densidade da probabilidade se obtém imediatamente mudando a escala da $f_{T^*_{n-1}}$ da Secção 6.1: $(X_{1:n}, X_{n:n})$. Obtém-se

$$f_{\tau_{n-1}}(t) = \frac{n-1}{n+1}\,\sqrt{\frac{n}{n+2}}\,\frac{1}{\left(1 + \sqrt{\frac{n}{n+2}}\,\frac{t}{n+1}\right)^n}\ \mathrm{I}_{(0,\infty)}(t),$$

que é a densidade de probabilidade de uma beta de segunda espécie (tipo VI de Pearson) com escala $\delta = (n+1)\sqrt{\dfrac{n+2}{n}}$.

6.4 Studentização do estimador do valor médio

Uma reparametrização em termos de (μ, θ), com $\mu = \theta_1 + \dfrac{\theta}{2}$, leva a

$$f_X(x) = \frac{1}{\theta}\, \mathrm{I}_{\left(\mu-\frac{\theta}{2}, \mu+\frac{\theta}{2}\right)}(x).$$

Consequentemente, o estimador de máxima verosimilhança do parâmetro de localização central é

$$\hat{\mu} = \frac{X_{1:n} + X_{n:n}}{2},$$

e o do parâmetro de escala é

$$\hat{\theta} = X_{n:n} - X_{1:n}.$$

A variável studentizada

$$\tau^*_{n-1} = \frac{\hat{\mu} - \mu}{\hat{\theta}}$$

tem função densidade de probabilidade

$$f_{\tau^*_{n-1}}(t) = \frac{n-1}{(1+2|t|)^n}, \ t \in \mathbb{R}$$

pelo que o quantil de probabilidade $1 - \frac{p}{2}$ é

$$Q_{\left(n-1;\ 1-\frac{p}{2}\right)} = \frac{1 - p^{\frac{1}{n-1}}}{2\, p^{\frac{1}{n-1}}}.$$

Capítulo 7

Studentização e $ANOSp$ no Caso de Populações Parentes Exponenciais

7.1 Studentização do mínimo pelo estimador da escala.

Sendo $\boldsymbol{X} = (X_1, \ldots, X_n)$ uma amostra aleatória da população $X \frown \mathit{Exponencial}(\lambda, \delta)$, $\lambda \in \mathbb{R}$, $\delta > 0$, $\widehat{\lambda} = X_{1:n}$ e $\widehat{\delta} = \overline{X} - X_{1:n}$.

Como podemos rescrever $\overline{X} - X_{1:n} = \dfrac{1}{n} \sum_{k=2}^{n} (X_{k:n} - X_{1:n})$,

torna-se evidente que os estimadores $\widehat{\lambda} \frown \mathit{Exponencial}\left(\lambda, \frac{\delta}{n}\right)$ e $\widehat{\delta} \frown \mathit{Gama}\left(n-1, \frac{\delta}{n}\right)$ são independentes.

Assim, a variável studentizada $W = \dfrac{X_{1:n} - \lambda}{\overline{X} - X_{1:n}}$ tem função densidade de probabilidade

$$f_W(x) = \int_0^{+\infty} \frac{\delta}{n}\ \mathrm{e}^{-\frac{nxy}{\delta}} \frac{y^{n-2}\ \mathrm{e}^{-\frac{ny}{\delta}}}{(\frac{\delta}{n})^{n-1}\Gamma(n-1)}\, y\, \mathrm{d}\, y\, \mathbb{I}_{\mathbb{R}}(x) =$$
$$= \frac{n-1}{(1+x)^n}\, \mathbb{I}_{\mathbb{R}}(x),$$

ou seja $W \frown \mathit{Pareto}(n-1, 0)$.

7.2 Studentização do mínimo pela amplitude.

Consideremos uma amostra aleatória ordenada $X_{1:n}, \ldots, X_{n:n}$ proveniente de uma distribuição exponencial, $X \frown \mathit{Exponencial}(\lambda, \delta)$, $\delta > 0$ e $\lambda \in \mathbb{R}$; seja a variável pivotal

$$\tau_{n-1} = \frac{X_{1:n} - \lambda}{X_{n:n} - X_{1:n}}.$$

Tendo em conta a transformação $Y = \dfrac{X - \lambda}{\delta}$, obtém-se

$$\tau_{n-1} = \frac{X_{1:n} - \lambda}{X_{n:n} - X_{1:n}} \stackrel{d}{=} \frac{Y_{1:n}}{Y_{n:n} - Y_{1:n}}$$

em que $Y_{1:n}, \ldots, Y_{n:n}$ são as estatísticas ordinais associadas a uma amostra aleatória $Y_1, \ldots, Y_n$ proveniente de uma população com distribuição exponencial padrão.

Convencionando que $X_{0:n} = \lambda$ o que corresponde $Y_{0:n} = 0$, e tendo em conta que

$$Y_{n:n} = \sum_{k=1}^{n} \left(Y_{k:n} - Y_{k-1:n} \right)$$

segue-se que

$$Y_{n:n} - Y_{1:n} = \sum_{k=2}^{n} \left(Y_{k:n} - Y_{k-1:n} \right).$$

Assim,

$$\tau_{n-1} = \frac{Y_{1:n}}{Y_{n:n} - Y_{1:n}} = \frac{Y_{1:n}}{\left(Y_{2:n} - Y_{1:n}\right) + \ldots + \left(Y_{n:n} - Y_{n-1:n}\right)}$$

e atendendo à independência dos *spacings* na população exponencial, conclui-se que $Y_{1:n}$ e $(Y_{n:n} - Y_{1:n})$ são variáveis aleatórias independentes. Simbolizando por $U = Y_{1:n}$ e $V = Y_{n:n} - Y_{1:n}$, segue-se que

$$\begin{cases} f_U(u) = n \, \mathrm{e}^{-nu} \, \mathrm{I}_{(0,\infty)}(u) \\ \\ f_V(v) = (n-1) \, \mathrm{e}^{-v} \left(1 - \mathrm{e}^{-v}\right)^{n-2} \, \mathrm{I}_{(0,\infty)}(v) \end{cases}$$

e da independência das margens de (U,V), conclui-se que

$$f_{(U,V)}(u,v) = n\ \mathrm{e}^{-nu}\ \mathrm{I}_{(0,\infty)}(u)(n-1)\ \mathrm{e}^{-v}\left(1-\mathrm{e}^{-v}\right)^{n-2}\ \mathrm{I}_{(0,\infty)}(v).$$

Considerando a transformação $(U,V) \to (S,\tau)$, com $S = U$ e $\tau = \frac{U}{V}$, cujo jacobiano tem valor absoluto $|J| = \frac{s}{t^2}$, obtém-se

$$f_{(S,\tau)}(s,t) = f_{(U,V)}\left(s, \frac{s}{t}\right)\ \frac{s}{t^2} = n(n-1)\ \mathrm{e}^{-ns}\ \mathrm{e}^{-\frac{s}{t}}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-2}\ \frac{s}{t^2}$$

com $s > 0$ e $t > 0$. Consequentemente,

$$\begin{aligned} f_{\tau_{n-1}}(t) &= \int_0^{+\infty} n(n-1)\ \mathrm{e}^{-ns}\ \mathrm{e}^{-\frac{s}{t}}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-2}\ \frac{s}{t^2}\,\mathrm{d}s = \\ &= \frac{1}{t}\int_0^{+\infty} ns\ \mathrm{e}^{-ns}\left[(n-1)\ \mathrm{e}^{-\frac{s}{t}}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-2}\frac{1}{t}\right]\mathrm{d}s, \end{aligned}$$

apropriado para proceder a integração por partes, obtendo-se

$$\begin{aligned} f_{\tau_{n-1}}(t) &= \frac{1}{t}\ ns\ \mathrm{e}^{-ns}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-1}\Bigg|_{s=0}^{s=+\infty} - \\ &\quad -\frac{1}{t}\int_0^{+\infty} n(1-ns)\ \mathrm{e}^{-ns}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-1}\mathrm{d}s \end{aligned}$$

e como $\lim\limits_{s\to\infty}\left\{ns\ \mathrm{e}^{-ns}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-1}\right\} = 0$, então

$$f_{\tau_{n-1}}(t) = -\frac{1}{t}\int_0^{+\infty} n\,\mathrm{e}^{-ns}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-1}\mathrm{d}s + \frac{1}{t}\int_0^{+\infty} n^2 s\,\mathrm{e}^{-ns}\left(1-\mathrm{e}^{-\frac{s}{t}}\right)^{n-1}\mathrm{d}s.$$

Efectuando a mudança de variável $Z = 1 - \mathrm{e}^{\left(-\frac{S}{\tau}\right)}$, vem para o valor absoluto do jacobiano da transformação $|J| = \dfrac{t}{1-z}$.

Como consequência,

$$
\begin{aligned}
&f_{\tau_{n-1}}(t) = \\
&= -n \int_0^1 \frac{\mathrm{e}^{nt\,\ln(1-z)}\, z^{n-1}}{1-z}\,\mathrm{d}z - n^2 \int_0^1 \frac{t\,\ln(1-z)\,\mathrm{e}^{nt\,\ln(1-z)}\, z^{n-1}}{1-z}\,\mathrm{d}z = \\
&= -n \int_0^1 (1-z)^{nt-1} z^{n-1}\mathrm{d}z - n^2 t \int_0^1 \ln(1-z)(1-z)^{nt-1} z^{n-1}\mathrm{d}z = \\
&= -n \left[B(n,nt) + nt\, \frac{\partial B(n,nt)}{\partial(nt)} \right].
\end{aligned}
$$

Como $\frac{\partial B(n,nt)}{\partial(nt)} = B(n,nt)\,[\psi(nt) - \psi(n+nt)]$, sendo ψ a função digama, vem então

$$
\begin{aligned}
f_{\tau_{n-1}}(t) &= -n\,B(n,nt)\,\{1 + nt\,[\psi(nt) - \psi(n+nt)]\} \\
&= n\,B(n,nt)\,\{nt\,[\psi(n+nt) - \psi(nt)] - 1\}\,,
\end{aligned}
$$

com $t > 0$.

A expressão anterior poderá apresentar outro aspecto se atendermos à seguinte fórmula de recorrência da função digama:

$$
\begin{aligned}
\psi(n+z) =& \frac{1}{(n-1)+z} + \frac{1}{(n-2)+z} + \cdots + \frac{1}{1+z} + \frac{1}{z} + \psi(z) = \\
&= \sum_{k=1}^{n} \frac{1}{(n-k)+z} + \psi(z).
\end{aligned}
$$

Considerando $z = nt$, a expressão analítica da função densidade de τ_{n-1}, assume então o seguinte aspecto:

$$f_{\tau_{n-1}}(t) = n\,B(n, nt)\left[nt\sum_{k=1}^{n}\frac{1}{(n-k)+nt} - 1\right],$$

ou seja,

$$f_{\tau_{n-1}}(t) = n\,B(n, nt)\sum_{k=1}^{n-1}\frac{nt}{(n-k)+nt},\ t > 0.$$

Brilhante (1996b) calculou a forma explícita da função densidade de probabilidade de $f_{\tau^*_{(n-1)}}$ para $n = 1, \ldots, 30, 50$, e procedeu a uma tabulação exaustiva de quantis críticos, de que apresentamos um extracto nas Tabelas 7.1 e 7.2.

Para valores mais elevados de n, recorre-se a resultados assintóticos, fáceis de estabelecer:

É óbvio que o numerador converge para 0, enquanto o denominador converge para $+\infty$. Mas como $n\,Y_{1:n} \stackrel{\mathrm{d}}{=} Y$ e por outro lado $Y_{n:n} - Y_{1:n} \stackrel{\mathrm{d}}{=} Y_{n-1:n-1}$, uma sucessão que converge para uma Gumbel se a cada termo subtrairmos $\ln(n-1)$, observa-se que

$$n\,\ln\,n\,\frac{Y_{1:n}}{Y_{n:n} - Y_{1:n}}$$

converge em distribuição para uma variável exponencial padrão.

Mais detalhadamente:

$$\mathbb{P}\left(\left|\frac{Y_{n-1:n-1}}{\ln n} - 1\right| \leq \varepsilon\right) = \mathbb{P}\left(1-\varepsilon \leq \frac{Y_{n-1:n-1}}{\ln n} \leq 1+\varepsilon\right).$$

Tabela 7.1: Quantis críticos da cauda esquerda de $\tau_{n-1} = \frac{X_{1:n} - \lambda}{X_{n:n} - X_{1:n}}$.

n	0.001	0.005	0.01	0.025	0.05	0.1
2	0.000501	0.002513	0.005051	0.012821	0.026316	0.055556
3	0.000222	0.001115	0.00224	0.005666	0.011562	0.02411
4	0.00013646	0.000684222	0.00137329	0.00347017	0.00706754	0.0146769
5	0.0000960638	0.0004816	0.000966424	0.00244062	0.00496571	0.0102906
6	0.0000730395	0.000366136	0.000734637	0.00185459	0.00377105	0.00780482
7	0.0000583455	0.000292459	0.000586757	0.0014809	0.00300995	0.00622415
8	0.0000482389	0.000241788	0.000485068	0.00122403	0.00248709	0.00513968
9	0.0000409066	0.000205028	0.000411304	0.00103775	0.0021081	0.00435433
10	0.0000353697	0.000177272	0.000355609	0.000897129	0.0018221	0.00376213
11	0.0000310563	0.000155649	0.000312225	0.000787609	0.00159942	0.00330131
12	0.0000276111	0.00013838	0.000277577	0.000700154	0.00142164	0.00293359
13	0.0000248026	0.000124303	0.000249333	0.000628874	0.00127677	0.00263405
14	0.0000224739	0.00011263	0.0022915	0.000569778	0.00115668	0.00238584
15	0.0000205148	0.00010281	0.000206216	0.000520069	0.00105568	0.00217715
16	0.0000188461	0.0000944469	0.000189438	0.000477736	0.000969682	0.00199949
17	0.0000174096	0.0000872468	0.000174994	0.000441294	0.000895657	0.00184661
18	0.0000161611	0.0000809897	0.000162442	0.000409627	0.000831338	0.0017138
19	0.0000150672	0.0000755067	0.000151443	0.00038188	0.000774985	0.00159746
20	0.0000141014	0.0000706666	0.000141734	0.000357388	0.000725247	0.00149479
21	0.0000132433	0.0000663656	0.000133107	0.000335625	0.000681054	0.00140358
22	0.0000124761	0.000062521	0.000125395	0.000316172	0.000641555	0.00132207
23	0.0000117867	0.0000590658	0.000118464	0.00029869	0.000606061	0.00124883
24	0.0000111641	0.0000559453	0.000112205	0.000282903	0.000574008	0.0011827
25	0.0000105992	0.0000531145	0.000106527	0.000268582	0.000544933	0.00112272
30	0.00000841864	0.0000421863	0.0000846071	0.000213301	0.000432716	0.000891285
50	0.00000446749	0.0000223859	0.0000448938	0.000113162	0.000229504	0.000472444

Tabela 7.2: Quantis críticos da cauda direita de $\tau_{n-1} = \frac{X_{1:n}-\lambda}{X_{n:n}-X_{1:n}}$.

n	0.9	0.95	0.975	0.99	0.995	0.999
2	4.5	9.5	19.5	49.5	99.5	499.5
3	1	1.614763	2.486079	4.216991	6.16875	14.40805
4	0.5	0.75	1.06703	1.61846	2.16449	4.04739
5	0.319126	0.462948	0.635782	0.917745	1.17975	1.99938
6	0.22899	0.325932	0.438821	0.616237	0.775077	1.24466
7	0.176015	0.247467	0.328973	0.453985	0.563222	0.874548
8	0.141563	0.197318	0.259996	0.354495	0.435671	0.661185
9	0.117566	0.162821	0.213143	0.288052	0.351592	0.524816
10	0.1	0.137805	0.179488	0.24093	0.29254	0.431222
11	0.0866474	0.11893	0.154283	0.205986	0.249079	0.363559
12	0.0761933	0.104239	0.134782	0.179165	0.215923	0.312669
13	0.0678113	0.0925182	0.1193	0.158009	0.189898	0.273188
14	0.0609576	0.0829739	0.106745	0.140945	0.168995	0.241782
15	0.0552609	0.0750686	0.0963814	0.126926	0.151881	0.21628
16	0.0504593	0.0684256	0.0876994	0.115228	0.137644	0.195213
17	0.0463632	0.0627737	0.0803323	0.105336	0.125636	0.177552
18	0.0428323	0.0579132	0.0740114	0.0968742	0.115388	0.162561
19	0.0397607	0.0536937	0.0685353	0.0895636	0.106553	0.149697
20	0.0370668	0.05	0.0637505	0.0831912	0.098865	0.13855
21	0.0346871	0.0467426	0.059538	0.0775931	0.0921224	0.128812
22	0.0325713	0.0438509	0.0558039	0.0726408	0.0861664	0.120238
23	0.0306791	0.0412684	0.0524738	0.068232	0.0808713	0.11264
24	0.0289779	0.0389496	0.0494875	0.0642849	0.0761365	0.105864
25	0.0274411	0.0368573	0.046796	0.0607329	0.0718803	0.0997897
30	0.0215675	0.0288832	0.0365665	0.047281	0.0558051	0.0769863
50	0.0111863	0.0148866	0.0187319	0.0240314	0.0282002	0.0383985

Se $\varepsilon \geq 1$,

$$\begin{aligned}
&\mathbb{P}\left(1-\varepsilon \leq \tfrac{Y_{n-1:n-1}}{\ln n} \leq 1+\varepsilon\right) = \mathbb{P}\left(\tfrac{Y_{n-1:n-1}}{\ln n} \leq 1+\varepsilon\right) \\
&= \mathbb{P}\left(Y_{n-1:n-1} \leq (1+\varepsilon)\ln n\right) \\
&= \left[1-e^{-(1+\varepsilon)\ln n}\right]^{n-1} = \left(1-\tfrac{1}{n^{1+\varepsilon}}\right)^{n-1} \\
&= \left(1-\tfrac{1}{n^{1+\varepsilon}}\right)^{n}\left(1-\tfrac{1}{n^{1+\varepsilon}}\right)^{-1} \\
&= \left[\left(1-\tfrac{1}{n^{1+\varepsilon}}\right)^{n^{1+\varepsilon}}\right]^{\frac{n}{n^{1+\varepsilon}}}\left(1-\tfrac{1}{n^{1+\varepsilon}}\right)^{-1} \xrightarrow[n\to\infty]{} 1
\end{aligned}$$

Se $\varepsilon < 1$,

$$\begin{aligned}
&\mathbb{P}\left(1-\varepsilon \leq \tfrac{Y_{n-1:n-1}}{\ln n} \leq 1+\varepsilon\right) \\
&= \mathbb{P}\left(Y_{n-1:n-1} \leq (1+\varepsilon)\ln n\right) - \mathbb{P}\left(Y_{n-1:n-1} \leq (1-\varepsilon)\ln n\right) \\
&= \left[1-e^{-(1+\varepsilon)\ln n}\right]^{n-1} - \left[1-e^{-(1-\varepsilon)\ln n}\right]^{n-1}.
\end{aligned}$$

Como

$$\begin{aligned}
&\left[1-e^{-(1-\varepsilon)\ln n}\right]^{n-1} = \left(1-\tfrac{1}{n^{1-\varepsilon}}\right)^{n-1} = \left(1-\tfrac{1}{n^{1-\varepsilon}}\right)^{n}\left(1-\tfrac{1}{n^{1-\varepsilon}}\right)^{-1} \\
&= \left[\left(1-\tfrac{1}{n^{1-\varepsilon}}\right)^{n^{1-\varepsilon}}\right]^{\frac{n}{n^{1-\varepsilon}}}\left(1-\tfrac{1}{n^{1-\varepsilon}}\right)^{-1} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

segue-se que

$$\mathbb{P}\left(1-\varepsilon \leq \tfrac{Y_{n-1:n-1}}{\ln n} \leq 1+\varepsilon\right) \xrightarrow[n\to\infty]{} 1\,.$$

Assim, $\frac{Y_{n-1:n-1}}{\ln n} \xrightarrow{\text{p}} 1$. Aplicando o teorema de Slutsky (Pestana e Velosa, 2010, p. 997–998) conclui-se que

$$n \ln n \, \frac{Y_{1:n}}{Y_{n-1:n-1}} \xrightarrow[n\to\infty]{\text{d}} Exponencial(1),$$

ou seja,

$$n \ln n \, \frac{Y_{1:n}}{Y_{n:n} - Y_{1:n}} \xrightarrow[n\to\infty]{\text{d}} Exponencial(1).$$

Um resultado interessante, mas de efeito prático limitado, uma vez que apenas propõe uma aproximação que deve ser usada com circunspecção no caso de amostras de dimensão pequena ou moderada.

7.3 Studentização usando a independência dos espaçamentos

A abordagem anterior tem decerto virtudes quando dispomos de uma amostra fiável; no entanto, em presença de erros grosseiros $\tau_{n-1} = \dfrac{X_{1:n} - \lambda}{X_{n:n} - X_{1:n}}$ tem limite de ruptura 0, uma vez que a escala é estimada pela amplitude da amostra — usando pois as estatísticas ordinais extremais.

É porventura alternativa interessante usar antes como avaliação da escala a dispersão quartal, menos eficiente do que o desvio padrão no caso de populações Gaussianas, mas bem mais robusta em outras situações, ou no caso de existência de valores periféricos ou de *outliers*. Ou ainda, mais geralmente, usando como avaliação da escala $X_{k:n} - X_{i:n},\ 1 \leq i < k \leq n$.

Vamos então considerar a variável studentizada

$$\tau^*_{n-1;i,k} = \frac{\overline{X}_n - \lambda}{X_{k:n} - X_{i:n}}$$

idêntica em distribuição a

$$\tau^*_{n-1;i,k} = \frac{\overline{Y}_n}{Y_{k:n} - Y_{i:n}}$$

com parente exponencial com valor médio 1 localizada em 0; $\overline{X}_n$ é um estimador suficiente e completo de δ e consequentemente, uma vez que a distribuição de $\tau^*_{n-1;i,k}$ não depende funcionalmente de δ, pelo teorema de Basu (1955) sabemos que $\overline{X}_n$ e $\tau^*_{n-1;i,k}$ são variáveis aleatórias independentes. Como a distribuição de $\tau^*_{n-1;i,k}$ não depende de δ, podemos sem perda de generalidade trabalhar a partir daqui como se a população parente fosse exponencial padrão.

Nestas condições a função densidade de probabilidade de $\overline{Y}_n$ é

$$f_{\overline{Y}_n}(y) = \frac{n^n y^{n-1} \mathrm{e}^{-ny}}{\Gamma(n)} \mathrm{I}_{(0,\infty)}(y).$$

Ora a função densidade de probabilidade de $Y_{k:n} - Y_{i:n}$ é

$$f_{Y_{k:n}-Y_{i:n}}(y) = \frac{\Gamma(n+1-i)}{\Gamma(k-i)\,\Gamma(n+1-k)} \mathrm{e}^{-y(n+1-k)} \left(1 - \mathrm{e}^{-y}\right)^{k-i-1} \mathrm{I}_{(0,\infty)}(y)$$

ou seja $Y_{k:n} - Y_{i:n} = -\ln(W)$, tendo W distribuição beta com parâmetros $n+1-k$ e $k-i$.

De facto, da falta de memória da exponencial é óbvio que $Y_{k:n} - Y_{i:n}$ é idêntica a $Y_{k-i:n-i}$. A relação com as betas advém

de propriedades gerais das estatística ordinais, devidas à transformação uniformizante, e ao facto de se X for uma beta com parâmetros p, q, $1-X$ ser uma beta com parâmetros q, p.

Temos assim que

$$Y_{k:n} - Y_{i:n} = \frac{\overline{Y}_n}{\tau^*_{n-1;i,k}}$$

ou, em termos de funções densidade de probabilidade

$$\int_0^{+\infty} \frac{n\,(nxy)^{n-1}\,\mathrm{e}^{-nxy}}{\Gamma(n)}\; y\, f_{\tau^*_{n-1;i,k}}(y)\;\mathrm{d}y =$$

$$= \frac{\Gamma(n+1-i)}{\Gamma(k-i)\,\Gamma(n+1-k)}\;\; \mathrm{e}^{-x(n+1-k)}\left(1-\mathrm{e}^{-x}\right)^{k-i-1}\mathrm{I}_{(0,\infty)}(x)$$

e assim, denotando $L(f;x)$ a transformada de Laplace de f no ponto x,

$$L\left(y^n f_{\tau^*_{n-1;i,k}}(y); nx\right) =$$

$$= \frac{\Gamma(n)\,\Gamma(n+1-i)}{n^n\Gamma(k-i)\,\Gamma(n+1-k)}\;\frac{\mathrm{e}^{-x(n+1-k)}\left(1-\mathrm{e}^{-x}\right)^{k-i-1}}{x^{n-1}}$$

donde

$$L\left(y^n f_{\tau^*_{n-1;i,k}}(y); x\right) =$$

$$= \frac{\Gamma(n)\,\Gamma(n+1-i)}{n\Gamma(k-i)\,\Gamma(n+1-k)}\;\frac{\mathrm{e}^{-\frac{x}{n}\;(n+1-k)}\left(1-\mathrm{e}^{-\frac{x}{n}}\right)^{(k-i-1)}}{x^{n-1}}.$$

Denotando $L^{-1}(f;y)$ a transformada de Laplace inversa de f no ponto y, vem

$$f_{\tau^*_{n-1;i,k}}(y) = \frac{\Gamma(n)\,\Gamma(n+1-i)}{n\Gamma(k-i)\,\Gamma(n+1-k)}\,\frac{1}{y^n}\times \\ \times L^{-1}\left(\frac{\mathrm{e}^{-\frac{x}{n}\ (n+1-k)}\left(1-\mathrm{e}^{-\frac{x}{n}}\right)^{(k-i-1)}}{x^{n-1}};y\right).$$

Um exemplo simples: considere-se $n=2$, $i=1$, $k=2$;

$$f_{\tau^*_{1;1,2}}(y) = \frac{1}{2y^2}\,L^{-1}\left(\frac{\mathrm{e}^{-\frac{x}{2}}}{x};y\right) = \frac{1}{2y^2}\,\mathrm{I}_{\left(\frac{1}{2},\infty\right)}(y)$$

ou seja,

$$\tau^*_{1;1,2} = \frac{X_{1:2}+X_{2:2}}{2\,(X_{2:2}-X_{1:2})}$$

é uma Pareto com índice 1 e parâmetro de localização $\frac{1}{2}$, um resultado que confere com o que obtemos, por outra via, no Apêndice A.

A situação mais interessante surge quando $n+1-k=k-i-1$. Considere-se então $\tau^*_{n-1;i,k}$ com i, k, n verificando a referida igualdade, isto é

n	i	k
$3j-1$	$j-1$	$2j$
$3j$	j	$2j+1$
$3j+1$	$j+1$	$2j+2$

Nesta situação vem

$$f_{\tau^*_{n-1;i,k}}(y) = \frac{\Gamma(n)\Gamma(2j+1)}{n\,\Gamma(j)\Gamma(j+1)}\,\frac{1}{y^n}\,L^{-1}\left(\frac{1}{x^{k-2}}\left(\frac{\mathrm{e}^{-\frac{x}{n}}\left(1-\mathrm{e}^{-\frac{x}{n}}\right)}{x}\right)^j;y\right)$$

e como

$$L^{-1}\left(\frac{\mathrm{e}^{-\frac{x}{n}}\left(1-\mathrm{e}^{-\frac{x}{n}}\right)}{x};y\right) = \mathrm{I}_{\left(\frac{1}{n},\frac{2}{n}\right)}(y)$$

e

$$L^{-1}\left(\frac{1}{x^n};y\right) = \frac{y^{n-1}}{\Gamma(n)}$$

basta recordar que a transformada de Laplace inversa de um produto é a convolução das transformadas de Laplace inversas dos factores.

Um exemplo muito simples: para $n = 3$, i=1, $k = 3$, vem

$$f_{\tau^*_{2;1,3}}(t) = \begin{cases} 0 & t < \frac{1}{3} \\ \frac{4\left(t-\frac{1}{3}\right)}{3t^3} & \frac{1}{3} \leq t < \frac{2}{3} \\ \frac{4}{9t^3} & t \geq \frac{2}{3}, \end{cases}$$

cujo gráfico se representa na Fig. 7.1.

Para para $n = 4$, $i = 2$, $k = 4$, vem

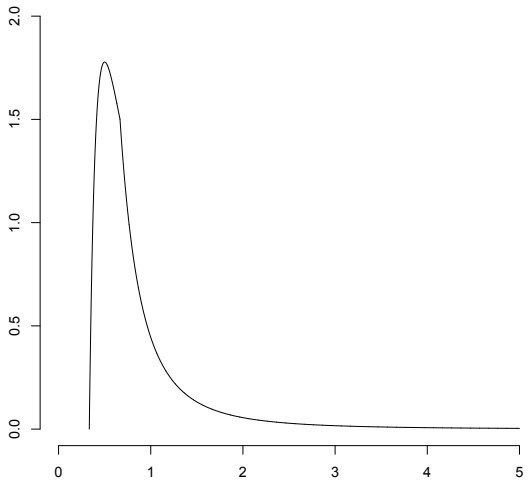

Figura 7.1: Densidade de $\tau^*_{2;1,3}$

$$f_{\tau^*_{3;2,4}}(t) = \begin{cases} 0 & t < \frac{1}{4} \\ \dfrac{3}{32t^4}(4t-1)^2 & \frac{1}{4} \leq t < \frac{1}{2} \\ \dfrac{3(8t-3)}{32t^4} & t \geq \frac{1}{2} \end{cases}$$

a que corresponde o gráfico da Fig. 7.2.

Brilhante *et al.* (1996) também procederam a uma tabulação exaustiva de quantis críticos da estatística $\tau^*_{n-1;i,k}$ para $n = 1, \ldots, 25, 30, 50$. (Tabela 7.3).

Tabela 7.3: Quantis críticos de $\tau^*_{n-1;i,k} = \frac{\overline{X}_n - \lambda}{X_{k:n} - X_{i:n}}$.

n	i	k	0.001	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	0.999
3	1	3	0.340957	0.350877	0.358697	0.375292	0.395936	0.429336	1.49071	2.10819	2.98142	4.71405	6.66667	14.9071
4	2	4	0.271553	0.289258	0.301567	0.325555	0.353308	0.395822	1.79672	2.60383	3.74135	5.99467	8.53244	19.2387
5	1	4	0.451275	0.481872	0.501286	0.536241	0.573179	0.624337	1.79163	2.31281	2.96466	4.08859	5.19617	9.00251
6	2	5	0.404304	0.439866	0.461814	0.500623	0.541023	0.597081	1.93635	2.52828	3.26727	4.54005	5.79359	10.0999
7	3	6	0.373282	0.412145	0.435759	0.477117	0.520069	0.580165	2.03545	2.67576	3.47458	4.84979	6.20392	10.855
8	2	6	0.495362	0.540184	0.566193	0.610203	0.654555	0.715025	1.93923	2.40243	2.94501	3.81446	4.6136	7.09137
9	3	7	0.469472	0.516524	0.543743	0.589912	0.636674	0.700813	2.00248	2.49299	3.0672	3.98691	4.83202	7.4518
10	4	8	0.449604	0.498404	0.526618	0.574574	0.623311	0.690388	2.05194	2.56382	3.1628	4.12196	5.00318	7.73456
11	3	8	0.540956	0.591963	0.620843	0.669181	0.717526	0.782997	1.96377	2.36465	2.81547	3.50589	4.11369	5.88409
12	4	9	0.523183	0.575748	0.605564	0.655587	0.70575	0.773818	1.99983	2.41515	2.88203	3.59686	4.22604	6.05843
13	5	10	0.508582	0.562481	0.5931	0.644558	0.696248	0.766474	2.02986	2.4572	2.93749	3.6727	4.31974	6.20397
14	4	10	0.580715	0.63532	0.665943	0.716882	0.767472	0.835382	1.95716	2.30928	2.6942	3.26553	3.75377	5.11685
15	5	11	0.567344	0.623225	0.654616	0.706903	0.758895	0.828738	1.98074	2.34184	2.73646	3.32208	3.82248	5.21933
16	6	12	0.555921	0.612927	0.644989	0.698448	0.751649	0.823149	2.00113	2.36998	2.7730	3.37101	3.88193	5.30806
17	5	12	0.61511	0.672086	0.703832	0.756351	0.808153	0.877114	1.94147	2.25603	2.59269	3.08098	3.48921	4.59421
18	6	13	0.604516	0.662577	0.694965	0.748586	0.801505	0.871973	1.95825	2.27896	2.62214	3.11981	3.53583	4.66184
19	7	14	0.595226	0.654259	0.687218	0.741813	0.795718	0.86751	1.97312	2.29927	2.64823	3.15422	3.57717	4.72183
20	6	14	0.645159	0.703758	0.736231	0.789683	0.842074	0.911308	1.92354	2.2086	2.50871	2.93625	3.28769	4.21667
21	7	15	0.636466	0.696005	0.729022	0.783394	0.836703	0.907156	1.93617	2.22574	2.53055	2.96473	3.32158	4.26483
22	8	16	0.628699	0.689089	0.722597	0.777795	0.831926	0.903471	1.94757	2.24118	2.555023	2.9904	3.35216	4.3083
23	7	16	0.671706	0.731426	0.764359	0.81832	0.870914	0.939971	1.90572	2.16709	2.43863	2.81995	3.12915	3.93126
24	8	17	0.664389	0.724931	0.758334	0.813078	0.866443	0.936514	1.91563	2.18046	2.45556	2.84183	3.15503	3.96745
25	9	18	0.657759	0.719053	0.752883	0.80834	0.862405	0.933397	1.92469	2.19266	2.47102	2.86182	3.17867	4.00052
30	10	21	0.711262	0.772902	0.806614	0.861427	0.914346	0.9831	1.87984	2.10779	2.33984	2.65864	2.91187	3.55044
50	16	34	—	0.885094	0.917836	0.970073	1.01943	1.08208	1.79297	1.95205	2.10773	2.31277	2.46933	2.84359

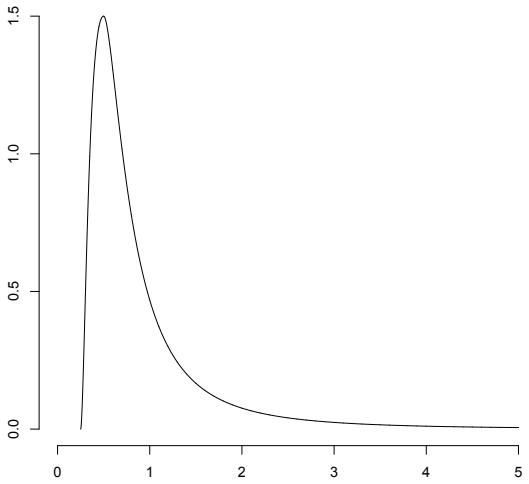

Figura 7.2: Densidade de $\tau^*_{3;2,4}$

7.4 Comparação das localizações de duas populações exponenciais, assumindo homocedasticidade

Para efectuarmos uma exposição evidenciando o essencial, vamos limitar-nos ao caso de duas amostras independentes homocedásticas (no sentido amplo de "com a mesma escala", $\delta_1 = \delta_2 = \delta$) com o mesmo tamanho $n = \frac{N}{2}$ ($N > 2$ par) por forma a obtermos o quociente de variáveis aleatórias em que o dividendo é a Laplace usual, simétrica, que como é sobejamente

sabido se obtém como diferença entre exponenciais independentes e identicamente distribuídas. O leitor pode explorar o caso geral, documentando-se em Johnson *et al.* (1995, p. 193) ou em Brilhante and Kotz (2008) sobre variáveis Laplace assimétricas.

Sejam $\boldsymbol{Y}_1$ uma amostra aleatória de tamanho $n = \frac{N}{2}$ de $Y_1 \frown \mathit{Exponencial}(\lambda_1, \delta)$, e $\boldsymbol{Y}_2$ uma amostra aleatória de tamanho $\frac{N}{2}$ de $Y_2 \frown \mathit{Exponencial}(\lambda_2, \delta)$ — por outras palavras, estamos a assumir "homocedasticidade" —, sendo as amostras $\boldsymbol{Y}_1$ e $\boldsymbol{Y}_2$ independentes. Denotamos $Y^{(1)}_{k:n}$ a k-ésima estatística ordinal ascendente de $\boldsymbol{Y}_1$, e $Y^{(2)}_{j:n}$ a j-ésima estatística ordinal ascendente de $\boldsymbol{Y}_2$, $\overline{Y}^{(1)}$ e $\overline{Y}^{(2)}$ as médias de $\boldsymbol{Y}_1$ e de $\boldsymbol{Y}_2$, respectivamente

Usando argumentos em tudo similares aos que expusemos no caso de uma população / uma amostra, facilmente se conclui que a variável studentizada W abaixo definida, resultante de dividir o resultado de centrar o estimador $Y^{(1)}_{1:n} - Y^{(2)}_{1:n}$ de $\lambda_1 - \lambda_2$ pelo resultado de reduzir o estimador $\widehat{\delta} = \frac{\left(\overline{Y}^{(1)} - Y^{(1)}_{1:n}\right) + \left(\overline{Y}^{(2)} - Y^{(2)}_{1:n}\right)}{2}$ da dispersão[(1)] comum δ,

$$W = \frac{Y^{(1)}_{1:n} - Y^{(2)}_{1:n} - (\lambda_1 - \lambda_2)}{\widehat{\delta}} = \frac{\frac{Y^{(1)}_{1:n} - \lambda_1 - (Y^{(2)}_{1:n} - \lambda_2)}{\delta}}{\frac{\widehat{\delta}}{\delta}},$$

é quociente das variáveis aleatórias independentes.

[(1)] No caso geral, com amostras de tamanhos n_1 e n_2 respectivamente, o estimador ponderado mais adequado do parâmetro de dispersão δ é $\widehat{\delta} = \frac{n_1\left(\overline{Y}^{(1)} - Y^{(1)}_{1:n_1}\right) + n_2\left(\overline{Y}^{(2)} - Y^{(2)}_{1:n_2}\right)}{n_1 + n_2}$.

Uma vez que

$$\frac{Y_{1:n}^{(1)} - Y_{1:n}^{(2)} - (\lambda_1 - \lambda_2)}{\delta} \frown \mathit{Laplace}\left(\frac{1}{n}\right)$$

e

$$\frac{\widehat{\delta}}{\delta} = \frac{\left(\overline{Y}^{(1)} - Y_{1:n}^{(1)}\right) + \left(\overline{Y}^{(2)} - Y_{1:n}^{(2)}\right)}{2\,\delta} \frown \mathit{Gama}\left(N-2, \frac{1}{N}\right),$$

deduz-se facilmente que a função densidade de probabilidade de W é

$$\begin{aligned} f_W(x) = & \int_0^{+\infty} \frac{n}{2}\, \mathrm{e}^{-\frac{n}{2}|x|\,y}\, \frac{N^{N-2} y^{N-3}\, \mathrm{e}^{-Ny}}{\Gamma(N-2)}\, y\,\mathrm{d}y\, \mathbb{I}_{\mathbb{R}}(x) = \\ & = \frac{N-2}{4\left(1+\frac{|x|}{2}\right)^{N-1}}\, \mathbb{I}_{\mathbb{R}}(x), \end{aligned}$$

e que os quantis da cauda direita são

$$w_{N,\,1-\alpha} = 2\left((2\alpha)^{-\frac{1}{N-2}} - 1\right),\ \alpha < \frac{1}{2}\,.$$

No caso geral $Y_1 \frown \mathit{Exponencial}(\lambda_k, \delta_k)$, $k = 1, 2$, como vimos na Secção 7.1,

$$Y_{1:n_k}^{(k)} - \lambda_k \frown \mathit{Exponencial}\left(\frac{\delta_k}{n_k}\right)$$

e

$$\widehat{\delta}_k = \overline{Y}^{(k)} - Y_{1:n_k}^{(k)} \frown \mathit{Gama}\left(n_k - 1, \frac{\delta_k}{n_k}\right),$$

independentes, segue-se que

$$\frac{Y^{(k)}_{1:n_k} - \lambda_k}{\overline{Y}^{(k)} - Y^{(k)}_{1:n_k}} \frown Pareto(n_k - 1, 0),$$

independentes. Consequentemente,

$$\frac{Y^{(1)}_{1:n_1} - \lambda_1}{\overline{Y}^{(1)} - Y^{(1)}_{1:n_1}} - \frac{Y^{(2)}_{1:n_2} - \lambda_2}{\overline{Y}^{(2)} - Y^{(2)}_{1:n_2}}$$

é a diferença entre duas Paretos com localização 0, independentes, intratável do ponto de vista analítico, mas de que facilmente se podem calcular aproximações dos quantis necessários.

7.5 Análise da Escala em populações exponenciais homocedásticas

A análise da variância em populações gaussianas é a análise das repercussões que a localização tem em estimadores da escala. A elegância dos resultados é uma consequência da independência entre $\overline{X}$ e S^2, que como se viu é uma propriedade característica de populações gaussianas.

Se a distribuição subjacente for exponencial, é possível usar uma propriedade característica de exponenciais — a independência entre *spacings* — para estudar a repercussão da localização λ (estimada pelo mínimo $X_{1:n}$) sobre a escala δ (estimada pela distância da média ao mínimo, $\overline{X} - X_{1:n}$). Apresentamos a situação mais elementar, de análise de escala "simples", e com

subamostras de iguais dimensões, para focar a atenção no essencial, assumindo homocedasticidadel, aqui no sentido amplo de assumirmos que as populações sob investigação têm escalas iguais. Remetemos para Pestana e Rocha (1993) e para Rocha (1995) para mais detalhes e análise de planeamentos experimentais mais sofisticados, sob hipótese de igualdade de escalas nos diversos grupos. Velosa (2001) abordou a extensão do problema de Behrens-Fisher-Welch-Satterthwaite a populações exponenciais com escalas diversas. Brilhante *et al.* (2009) retomaram a questão, desenvolvendo a *ANOSp* simples — *ANalysis Of Spacings* — para investigar a igualdade de localizações em populações exponenciais, usando a repercussão da localização na escala.

Consideremos então k amostras aleatórias independentes de dimensão n

$$\boldsymbol{X}_{i} = (X_{i1}, \ldots, X_{in}) \text{ com } i = 1, \ldots, k,$$

provenientes de populações exponenciais com a mesma escala. Assim, assumimos $X_{ij} \frown \mathit{Exponencial}\,(\lambda_i, \delta)$. É interessante, em muitas situações, testar a hipótese nula

$$H_0 : \lambda_1 = \ldots = \lambda_k \ (= \lambda).$$

Simbolizemos por $X^{(i)}_{1:n} = \min\limits_{j} X_{ij}$ o mínimo da amostra i, e mais geralmente por $X^{(i)}_{j:n}$ a j-ésima estatística ordinal da amostra i, $i = 1, \ldots, k$. Simbolizemos também por

$$X_{1:N} = \min_{i,j} X_{ij} = \min_{i} X^{(i)}_{1:n}$$

o mínimo global das k amostras consideradas em conjunto, com $N = kn$. Sob validade de H_0, as N observações, provenientes das

k amostras, poderão ser consideradas como uma única amostra, pelo que

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - X_{1:N} \right).$$

Denotando por $TSSp$ (*Total Sum of Spacings*) o somatório anterior, segue-se que

$$\begin{aligned} TSSp &= \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - X_{1:N} \right) = \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - X_{1:n}^{(i)} + X_{1:n}^{(i)} - X_{1:N} \right) = \\ &= n \sum_{i=1}^{k} \left(X_{1:n}^{(i)} - X_{1:N} \right) + \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - X_{1:n}^{(i)} \right) = \\ &= BSSp + WSSp \end{aligned}$$

onde $BSSp$ e $WSSp$ denotam os *Between Sum of Spacings* e *Within Sum of Spacings*, respectivamente, e consequentemente

$$\hat{\delta} = \frac{TSSp}{N} = \frac{BSSp}{N} + \frac{WSSp}{N} = \hat{\delta}_{B} + \hat{\delta}_{W}.$$

Verificamos assim que é possível decompor o estimador da escala $\hat{\delta}$ na soma de dois estimadores da escala $\hat{\delta}_{B}$ e $\hat{\delta}_{W}$, que refletem a variação "entre" amostras e "dentro" das amostras, respectivamente; estudemos as distribuições de $\hat{\delta}_{B}$ e $\hat{\delta}_{W}$.

Se H_0 for verdadeira, então todas as amostras têm a mesma localização λ, e como o mínimo de cada amostra tem distribuição

exponencial, ou seja $X^{(i)}_{1:n}|_{H_0} \frown Exponencial(\lambda, \frac{\delta}{n})$, segue-se que

$$\begin{aligned} BSSp &= n\sum_{i=1}^{k}\left(X^{(i)}_{1:n} - X_{1:N}\right) = \\ &= \underbrace{\underbrace{n\sum_{j=2}^{k}(k+1-j)\left(X_{j:k;1:n} - X_{j-1:k;1:n}\right)}_{Gama(k-1,\frac{\delta}{n})}}_{Gama(k-1,\delta)} \end{aligned}$$

Consequentemente $\frac{BSSp}{k-1}$ é um estimador não enviesado de δ sob a validade de H_0.

Quer H_0 seja verdadeira ou falsa, dentro da i-ésima amostra todas as variáveis X_{ij} têm a mesma localização λ_i pois $X_{ij} \frown Exp(\lambda_i, \delta)$, por hipótese, e a simples observação que

$$\begin{aligned} WSSp &= \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - X^{(i)}_{1:n}\right) = \\ &= \sum_{i=1}^{k}\sum_{j=2}^{n}\left(X^{(i)}_{j:n} - X^{(i)}_{1:n}\right) = \\ &= \underbrace{\sum_{i=1}^{k}\underbrace{\sum_{j=2}^{n}(n+1-j)\left(X^{(i)}_{j:n} - X^{(i)}_{j-1:n}\right)}_{Gama(n-1,\delta)}}_{Gama(N-k,\delta)} \end{aligned}$$

revela que $\frac{WSSp}{N-k}$ é um estimador não enviesado de δ quer H_0 seja verdadeira ou não.

A independência dos *spacings* no modelo exponencial acarreta a independência de $BSSp$ e $WSSp$, ou seja a independência dos estimadores $\hat{\delta}_{B}$ e $\hat{\delta}_{W}$.

Em resumo:

Componente	**Distribuição**	**Graus de liberdade**
$\hat{\delta}_{B}$	$Gama(k-1,\ \frac{\delta}{N})$	$k-1$
$\hat{\delta}_{W}$	$Gama(N-k, \frac{\delta}{N})$	$N-k$
$\hat{\delta}$	$Gama(N-1,\ \frac{\delta}{N})$	$N-1$

Se H_0 for verdadeira, como $\hat{\lambda}_{i} = X_{1:n}^{(i)}$, numa situação ideal será de esperar, que todos os mínimos de cada amostra sejam iguais, isto é,

$$X_{1:n}^{(1)} = X_{1:n}^{(2)} = \ldots = X_{1:n}^{(k)} = X_{1:N}$$

o que significa que

$$\sum_{i=2}^{k} \left(X_{1:n}^{(i)} - X_{1:N} \right) = 0 \implies \hat{\delta}_{B} = 0 \implies \hat{\delta} = \hat{\delta}_{W}.$$

Numa situação real, devido à flutuação amostral, parece razoável rejeitar H_0 se a contribuição de $\hat{\delta}_{B}$ para o valor da estimativa $\hat{\delta}$ for igual ou superior a um determinado valor crítico q, isto é, parece plausível rejeitar a hipótese nula se

$$\mathbb{P}\left(\frac{\hat{\delta}_{B}}{\hat{\delta}_{W}} \geq q \right) = \alpha.$$

Como

$$\frac{\hat{\delta}_{B}}{\hat{\delta}_{W}} \geq q \Longleftrightarrow \frac{BSSp}{WSSp} \geq q \Longleftrightarrow \frac{2(N-k)}{2(k-1)}\frac{BSSp}{WSSp} \geq \underbrace{\frac{2(N-k)}{2(k-1)}q}_{q^{*}}$$

segue-se que

$$\mathbb{P}\left(\frac{\hat{\delta}_{B}}{\hat{\delta}_{W}} \geq q\right) = \mathbb{P}\left(\frac{BSSp/(k-1)}{WSSp/(N-k)} \geq q^{*}\right) = \alpha$$

onde

$$\frac{BSSp/(k-1)}{WSSp/(N-k)}\Big|_{H_0} \frown F_{2(k-1);2(N-k)}.$$

Assim, para um nível de significância α, devemos rejeitar a hipótese nula $H_0 : \lambda_1 = \ldots = \lambda_k \ (= \lambda)$, se

$$\frac{(N-k)BSSp}{(k-1)WSSp} = \frac{(N-k)\hat{\delta}_{B}}{(k-1)\hat{\delta}_{W}} \geq F_{2(k-1),2(N-k);1-\alpha}$$

em que $F_{2(k-1),2(N-k);1-\alpha}$ é o quantil de probabilidade $1-\alpha$ de uma distribuição de Fisher-Snedecor com $2(k-1)$ e $2(N-k)$ graus de liberdade.

Consequentemente um quadro ANOSp (*ANalysis Of Spacings*), semelhante a um quadro ANOVA simples, pode ser utilizado, ver Tabela 7.4. Não há, obviamente, qualquer dificuldade em considerar subamostras de diferentes dimensões. No caso de um planeamento equilibrado em que as k amostras têm igual dimensão n há as simplificações decorrentes de os mínimos dos k grupos, condicionalmente a H_0, serem identicamente distribuídos, mais precisamente exponenciais com a mesma dispersão.

Tabela 7.4: Tabela da *ANOSp* (simples).

Soma de *Spacings*	Graus de liberdade	Estimador de δ	Valor de F
$BSSp = n\sum_{i=1}^{k}\left(X_{1:n}^{(i)} - X_{1:N}\right)$	$2(k-1)$	$\hat{\delta}_B^* = \frac{BSSp}{k-1}$	$F = \frac{\hat{\delta}_B^*}{\hat{\delta}_W^*}$
$WSSp = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - X_{1:n}^{(i)}\right)$	$2(N-k)$	$\hat{\delta}_W^* = \frac{WSSp}{N-k}$	

Capítulo 8

Studentização em Populações Simétricas

Seja $\{X_k\}_{k=1}^{\infty}$ uma sucessão de variáveis aleatórias i.i.d. absolutamente contínuas, com função densidade de probabilidade f. Sem perda de generalidade, admitamos que as variáveis aleatórias X_k estão centradas em zero.

8.1 Expressão recursiva para $f_{(\overline{X}_n, SS_n)}$

Denotando

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{e} \qquad SS_n = \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2,$$

a estatística de Student é

$$T_{n-1} = \sqrt{n(n-1)} \ \frac{\overline{X}_n}{\sqrt{SS_n}},$$

pelo que parece conveniente começar por deduzir uma expressão recursiva para a função densidade de probabilidade conjunta de $\left(\overline{X}_n, SS_n\right)$.

Densidade Conjunta de $\left(\overline{X}_2, SS_2\right)$

Sejam X_1 e X_2 variáveis aleatórias i.i.d., que sem perda de generalidade admitimos terem sido centradas em zero, com função densidade de probabilidade f_X. Então

$$\overline{X}_2 = \frac{X_1 + X_2}{2}; \qquad SS_2 = \frac{\left(X_1 - X_2\right)^2}{2}.$$

Consideremos os subespaços

$$D_1 = \{(X_1, X_2) : X_1 \geq X_2\}$$

$$D_2 = \{(X_1, X_2) : X_1 < X_2\}.$$

Tendo em conta o subespaço $D_1 = \{(X_1, X_2) : X_1 \geq X_2\}$, segue-se que

$$\left\{ \begin{array}{l} \overline{X}_2 = \frac{X_1+X_2}{2} \\ \\ SS_2 = \frac{\left(X_1-X_2\right)^2}{2} \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} X_1 = \overline{X}_2 + \sqrt{\frac{SS_2}{2}} \\ \\ X_2 = \overline{X}_2 - \sqrt{\frac{SS_2}{2}} \end{array} \right.$$

e assim $|J| = \dfrac{1}{\sqrt{2\,SS_{_2}}}$.

No subespaço $D_{_2} = \{(X_{_1}, X_{_2}) : X_{_1} < X_{_2}\}$ obtemos resultado semelhante,

$$\left\{ \begin{array}{l} \overline{X}_{_2} = \frac{X_{_1}+X_{_2}}{2} \\ \\ SS_{_2} = \frac{\left(X_{_1}-X_{_2}\right)^{2}}{2} \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} X_{_1} = \overline{X}_{_2} - \sqrt{\frac{SS_{_2}}{2}} \\ \\ X_{_2} = \overline{X}_{_2} + \sqrt{\frac{SS_{_2}}{2}} \end{array} \right.$$

e consequentemente,

$$f_{(\overline{X}_{_2}, SS_{_2})}(w,s) = f\Big(w+\sqrt{\tfrac{s}{2}}\Big) f\Big(w-\sqrt{\tfrac{s}{2}}\Big) \tfrac{1}{\sqrt{2s}} + f\Big(w-\sqrt{\tfrac{s}{2}}\Big) f\Big(w+\sqrt{\tfrac{s}{2}}\Big) \tfrac{1}{\sqrt{2s}},$$

ou seja

$$f_{(\overline{X}_{_2}, SS_{_2})}(w,s) = \sqrt{\frac{2}{s}}\; f\Big(w+\sqrt{\frac{s}{2}}\Big)\, f\Big(w-\sqrt{\frac{s}{2}}\Big), \qquad (8.1)$$

$w \in \mathbb{R}, s > 0$.

Densidade Conjunta de $\left(\overline{X}_{_n}, SS_{_n}\right)$

Comecemos por observar que

$$\overline{X}_{_{n+1}} = \frac{1}{n+1}\sum_{i=1}^{n+1} X_{_i} = \frac{n}{n+1}\,\frac{1}{n}\left(\sum_{i=1}^{n} X_{_i} + X_{_{n+1}}\right),$$

e assim

$$\overline{X}_{_{n+1}} = \frac{n}{n+1}\,\overline{X}_{_n} + \frac{1}{n+1}\,X_{_{n+1}}.$$

Por outro lado,

$$SS_{n+1} = \sum_{i=1}^{n+1} \left(X_i - \overline{X}_{n+1}\right)^2 = \sum_{i=1}^{n+1} \left(X_i - \frac{n\overline{X}_n + X_{n+1}}{n+1}\right)^2 =$$

$$= \sum_{i=1}^{n+1} \left(X_i - \frac{(n+1)\overline{X}_n - \overline{X}_n + X_{n+1}}{n+1}\right)^2 =$$

$$= \sum_{i=1}^{n+1} \left[X_i - \overline{X}_n + \frac{\left(\overline{X}_n - X_{n+1}\right)}{n+1}\right]^2 =$$

$$= \sum_{i=1}^{n+1} \left(X_i - \overline{X}_n\right)^2 + \frac{2\left(\overline{X}_n - X_{n+1}\right)}{n+1} \sum_{i=1}^{n+1} \left(X_i - \overline{X}_n\right) +$$

$$+ \sum_{i=1}^{n+1} \left(\frac{\overline{X}_n - X_{n+1}}{n+1}\right)^2 =$$

$$= SS_n + \left(X_{n+1} - \overline{X}_n\right)^2 - \frac{2}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2 + \frac{1}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2,$$

e consequentemente

$$SS_{n+1} = SS_n + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2.$$

Simbolizando por

$$f_n(w,s) = f_{(\overline{X}_n, SS_n)}(w,s), \quad w \in \mathbb{R}, \quad s > 0$$

e uma vez que X_{n+1} é independente de (X_n, SS_n), a função densidade de probabilidade conjunta de $\left(X_n, SS_n, X_{n+1}\right)$ é:

$$f_{\left(\overline{X}_n, SS_n, X_{n+1}\right)}\left(w, s, x_{n+1}\right) = f_n(w,s)\, f_{X_{n+1}}\left(x_{n+1}\right).$$

Definindo a variável auxiliar

$$U = \overline{X}_n - X_{n+1}$$

podemos então considerar

$$\boldsymbol{g}: \mathbb{R}^3 \longrightarrow \mathbb{R}^3$$

$$\left(\overline{X}_n, SS_n, X_{n+1}\right) \mapsto \left(\overline{X}_{n+1}, SS_{n+1}, U\right).$$

De $\overline{X}_{n+1} = \frac{1}{n+1}\left(\sum_{i=1}^{n} X_i + X_{n+1}\right)$, e tendo em conta a definição de U, obtém-se,

$$\overline{X}_{n+1} = \frac{n\overline{X}_n + X_{n+1}}{n+1} = \frac{(n+1)\overline{X}_n - \overline{X}_n + X_{n+1}}{n+1} = \overline{X}_n - \frac{U}{n+1}$$

donde

$$\overline{X}_n = \overline{X}_{n+1} + \frac{U}{n+1}.$$

De modo análogo,

$$SS_{n+1} = SS_n + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2 = SS_n + \frac{n}{n+1}U^2$$

pelo que

$$SS_n = SS_{n+1} - \frac{n}{n+1}U^2.$$

Por outro lado,

$$X_{n+1} = \overline{X}_n - U = \overline{X}_{n+1} + \frac{U}{n+1} - U,$$

e como tal

$$X_{n+1} = \overline{X}_{n+1} - \frac{n}{n+1}\, U.$$

Consequentemente, a transformação inversa de $\boldsymbol{g}$ é

$$\begin{cases} \overline{X}_n = \overline{X}_{n+1} + \dfrac{U}{n+1} \\ SS_n = SS_{n+1} - \dfrac{n}{n+1}\, U^2 \qquad (\Longrightarrow U^2 < \dfrac{n+1}{n}\, SS_{n+1}) \\ X_{n+1} = \overline{X}_{n+1} - \dfrac{n}{n+1}\, U \end{cases}$$

cujo jacobiano é

$$J = \begin{vmatrix} 1 & 0 & \frac{1}{n+1} \\ 0 & 1 & -\frac{2\,n\,U}{n+1} \\ 1 & 0 & -\frac{n}{n+1} \end{vmatrix} = -\frac{n}{n+1} - \frac{1}{n+1} = -1.$$

Então, a função densidade de probabilidade conjunta de $\left(\overline{X}_{n+1}, SS_{n+1}, U\right)$ é

$$f_{\left(\overline{X}_{n+1}, SS_{n+1}, U\right)}(w,s,u) = f_n\!\left(w + \frac{u}{n+1}, s - \frac{nu^2}{n+1}\right) f\!\left(w - \frac{nu}{n+1}\right)$$

donde, para $w \in \mathbb{R}$ e $s > 0$,

$$f_{n+1}(w,s) = \int_{-\sqrt{\frac{n+1}{n}\,s}}^{\sqrt{\frac{n+1}{n}\,s}} f_n\!\left(w + \frac{u}{n+1}, s - \frac{nu^2}{n+1}\right) f\!\left(w - \frac{nu}{n+1}\right) \mathrm{d}u.$$

Fazendo a transformação $u = \sqrt{\frac{n+1}{n}}\, s\, v$, a função densidade de probabilidade conjunta de $\left(\overline{X}_{n+1}, SS_{n+1}\right)$ assume a forma recursiva

$$f_{n+1}(w,s) = \sqrt{\frac{n+1}{n}}\, s \times$$

$$\times \int_{-1}^{+1} f_n\Big(w + \sqrt{\frac{s}{n(n+1)}}\, v, s\left(1-v^2\right)\Big) f\Big(w - \sqrt{\frac{ns}{n+1}}\, v\Big)\, \mathrm{d}v \tag{8.2}$$

em que o intervalo de integração é independente da variável s.

Como primeiro passo na utilização desta solução recursiva comecemos por calcular $f_3(w,s)$. Como em 8.1 deduzimos,

$$f_2(w,s) = \sqrt{\frac{2}{s}}\, f\Big(w + \sqrt{\frac{s}{2}}\Big) f\Big(w - \sqrt{\frac{s}{2}}\Big), \quad w \in \mathbb{R},\ s > 0.^{(1)}$$

Da expressão recursiva (8.2)

$$f_3(w,s) = \sqrt{3} \int_{-1}^{+1} \sqrt{\frac{1}{1-v^2}}\, f\left(w + \frac{v + \sqrt{3\left(1-v^2\right)}}{\sqrt{6}}\, \sqrt{s}\right) \times$$

$$\times f\left(w + \frac{v - \sqrt{3\left(1-v^2\right)}}{\sqrt{6}}\, \sqrt{s}\right) f\Big(w - \frac{2v}{\sqrt{6}}\, \sqrt{s}\Big)\, \mathrm{d}v$$

(1) No caso de f ter como suporte $\mathbb{R}^+$, então $w \in \mathbb{R}^+$ e $s \in \left(0, 2w^2\right)$; no caso de o suporte ser o intervalo (a,b), então $w \in (a,b)$ e $s \in \left(0, 2\,\min\left[(w-a)^2, (b-w)^2\right]\right)$.

ou seja,

$$f_{_3}(w,s) = \sqrt{3} \int_{-1}^{+1} \sqrt{\frac{1}{1-v^{^2}}} \prod_{i=1}^{3} f\big(w + \alpha_{_{i,3}}(v)\sqrt{s}\big)\ \mathrm{d}v \qquad (8.3)$$

com

$$\alpha_{_{1,3}}(v) = \frac{v+\sqrt{3\big(1-v^{^2}\big)}}{\sqrt{6}}\,, \quad \alpha_{_{2,3}}(v) = \frac{v-\sqrt{3\big(1-v^{^2}\big)}}{\sqrt{6}}\,, \quad \alpha_{_{3,3}}(v) = -\frac{2v}{\sqrt{6}}\,.$$

Note-se desde já que $\sum_{i=1}^{3} \alpha_{_{i,3}}(v) = 0$ e $\sum_{i=1}^{3} \alpha_{_{i,3}}^{^2}(v) = 1$. A fim de prosseguirmos, é necessário considerarmos algumas hipóteses simplificadoras:

- $H_{_1}$: f é uma função contínua em $\mathbb{R}$.

  Consequentemente, a função integranda é contínua em $(-1,+1)$, ilimitada nos extremos do intervalo. Recordando a forma generalizada do teorema do valor médio para integrais, existe $\xi_{_3} \in (-1,+1)$ tal que

  $$f_{_3}(w,s) = 2\sqrt{\frac{3}{1-\xi_{_3}^{^2}}} \prod_{i=1}^{3} f\big(w + \alpha_{_{i,3}}(\xi_{_3})\sqrt{s}\big)$$

- $H_{_2}$: $\xi_{_3} = \xi_{_3}(f)$

  De facto, no caso geral, está garantida a existência de $\xi_{_3}$ para cada par (w,s), pelo que $\xi_{_3} = \xi_{_3}(f,w,s)$. Uma vez que a função integranda de (8.2,8.3) envolve f apenas em

  $$\prod_{i=1}^{3} \left[\sqrt{s}\left(\frac{w}{\sqrt{s}} + \alpha_{_{i,3}}(v)\right)\right]$$

podemos notar que para funções densidade de probabilidade simétricas em torno de zero e suficientemente regulares, valores grandes de $\frac{w}{\sqrt{s}}$ levam a que a integração se perfaça nas caudas da densidade, e o factor multiplicativo tende a perder importância à medida que $\frac{w}{\sqrt{s}}$ aumenta. Assim, para funções densidade de probabilidade simétricas em torno de zero e regulares, $\xi_3 \cong \xi_3\Big(f, \Big|\frac{w}{\sqrt{s}}\Big|\Big)$. Se a função densidade de probabilidade depender de um factor de escala, devido à invariância de $\frac{\overline{X}_n}{\sqrt{SS_n}}$ no que respeita a escala, podemos considerar $\frac{w}{\sqrt{s}}$ grande, e assim $\xi_3 \approx \xi_3(f)$.

Note-se, no entanto, que os resultados que seguem são exactos no caso de distribuições parentes para as quais $\xi_3 \approx \xi_3(f)$ é constante (independente de w e de s), e apenas aproximado caso tal não se verifique; a aproximação é boa nomeadamente no caso de funções densidade de probabilidade simétricas em torno de zero, suficientemente regulares, e dependentes de um parâmetro de escala. A questão será discutida com mais detalhe posteriormente.

Considere-se então que, com H_1 e H_2, se toma como hipótese de indução, baseada em (8.3), que para $3, 4, \ldots, n$ se tem

$$f_n(w,s) = 2^{n-2}\sqrt{n}\left[\prod_{i=3}^{n}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right] s^{\frac{n-3}{2}} \prod_{i=1}^{n} f\big(w+\alpha_{i,n}\sqrt{s}\big), \tag{8.4}$$

$w \in \mathbb{R},\ s > 0$, com

$$\alpha_{i,n} = \frac{\xi_n}{\sqrt{n(n-1)}} + \alpha_{i,n-1}\sqrt{1-\xi_n^2},\ i = 1, \ldots, n-1$$

$$\alpha_{n,n} = -\xi_n \sqrt{\tfrac{n-1}{n}}$$

em que

$$\sum_{i=1}^{n} \alpha_{i,n} = 0 \qquad \text{e} \qquad \sum_{i=1}^{n} \alpha_{i,n}^2 = 1.$$

Note-se que (8.4), inspirada na expressão (8.3) para $f_3(w,s)$ é válida também para $f_2(w,s)$, desde que se aceite a convenção natural que os produtos vazios tomam o valor neutro do produto. Aplicando a relação de recorrência (8.4)

$$f_{n+1}(w,s) = \sqrt{\tfrac{(n+1)s}{n}}\ 2^{n-2}\sqrt{n}\left[\prod_{i=3}^{n}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right] s^{\frac{n-3}{2}} \int_{-1}^{+1} \left\{\left(1-v^2\right)^{\frac{n-3}{2}} \times \right.$$

$$\left. \times \prod_{i=1}^{n} f\left(w + \left(\tfrac{v}{\sqrt{n(n+1)}} + \alpha_{i,n}\sqrt{1-v^2}\right)\sqrt{s}\right) f\left(w - \sqrt{\tfrac{ns}{n+1}}\ v\right) \right. \mathrm{d}v$$

Denotando

$$\alpha_{i,n+1} = \alpha_{i,n+1}(v) = \tfrac{v}{\sqrt{n(n+1)}} + \alpha_{i,n}\sqrt{1-v_n^2},\ i = 1, \ldots, n$$

$$\alpha_{n+1,n+1} = \alpha_{n+1,n+1}(v) = -\sqrt{\tfrac{n}{n+1}}\ v$$

podemos escrever a expressão acima na forma

$$f_{n+1}(w,s) = 2^{n-2}\sqrt{n+1}\left[\prod_{i=3}^{n}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right] s^{\frac{n+1-3}{2}} \times$$

$$\times \int_{-1}^{+1}\left(1-v^2\right)^{\frac{n-3}{2}} \prod_{i=1}^{n+1} f\big(w+\alpha_{i,n+1}(v)\sqrt{s}\big)\,\mathrm{d}v,$$

$w \in \mathbb{R},\ s > 0$. Recorrendo à forma atrás referida do teorema do valor médio para integrais, podemos afirmar a existência de $\xi_{n+1} \in [-1,+1]$, que assumimos independente de (w,s) pelas considerações atrás feitas em H_2, tal que

$$f_{n+1}(w,s) = 2^{n-2}\sqrt{n+1}\left[\prod_{i=3}^{n}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right] s^{\frac{n+1-3}{2}} \times$$

$$\times 2\left(1-\xi_{n+1}^2\right)^{\frac{n+1-4}{2}} \prod_{i=1}^{n+1} f\big(w+\alpha_{i,n+1}(\xi_{n+1})\sqrt{s}\big) =$$

$$= 2^{n+1-2}\sqrt{n+1}\left[\prod_{i=3}^{n+1}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right] s^{\frac{n+1-3}{2}} \times$$

$$\times \prod_{i=1}^{n+1} f\big(w+\alpha_{i,n+1}(\xi_{n+1})\sqrt{s}\big)$$

com

$$\alpha_{i,n+1} = \alpha_{i,n+1}(\xi_{n+1}) = \frac{\xi_{n+1}}{\sqrt{n(n+1)}} + \alpha_{i,n}\sqrt{1-\xi_{n+1}^2},\ i = 1,\ldots,n;$$

$$\alpha_{n+1,n+1} = \alpha_{n+1,n+1}(\xi_{n+1}) = -\sqrt{\frac{n}{n+1}}\ \xi_{n+1}$$

e obviamente

$$\sum_{i=1}^{n+1} \alpha_{i,n+1}\big(\xi_{n+1}\big) = 0, \qquad \sum_{i=1}^{n+1} \alpha^{2}_{i,n+1}\left(\xi_{n+1}\right) = 1.$$

Adiante discutiremos modos mais expeditos de obter os $\alpha_{i,n}$, sob hipótese de simetria da distribuição parente. Demonstrámos assim:

Teorema 8.1.1.

Admitindo que a função densidade de probabilidade parente f é contínua, e que $\xi_n(f,w,s) = \xi_n(f)$,

$$f_n(w,s) = 2^{n-2}\sqrt{n}\left[\prod_{i=3}^{n}\left(1-\xi_i^2\right)^{\frac{i-4}{2}}\right] s^{\frac{n-3}{2}} \prod_{i=1}^{n} f\big(w+\alpha_{i,n}(\xi_n)\sqrt{s}\big),$$

$w \in \mathbb{R},\ s>0$, onde os $\alpha_{i,n}(\xi_n) = \alpha_{i,n}$ são dados por

$$\alpha_{i,n} = \frac{\xi_n}{\sqrt{n(n-1)}} + \alpha_{i,n-1}\sqrt{1-\xi_n^2},\ i = 1,2,\ldots,n-1$$

e

$$\alpha_{n,n} = -\xi_n\sqrt{\frac{n-1}{n}}\,.$$

Sob hipótese de simetria da distribuição parente, adiante usaremos o facto de $\sum_{i=1}^{n} \alpha_{i,n} = 0$ e $\sum_{i=1}^{n} \alpha^{2}_{i,n} = 1$ para demonstrar que os ξ_n não nulos são da forma $\sqrt{\frac{3}{n+1}}$ e os correspondentes $\alpha_{i,n} = (n+1-2i)\sqrt{\frac{3}{n\left(n^2-1\right)}}$.

Observe-se também que, atendendo às considerações expostas sobre $\xi_3(f,w,s) \approx \xi_3(f)$ — o que obviamente é válido para ξ_n geral — a expressão (8.4) é uma aproximação para $f_n(w,s)$ sempre que f seja aproximadamente simétrica em torno de zero e dependente de um factor de escala.

8.2 Densidade de $T_{n-1} = \sqrt{n}\ \frac{\overline{X}_n}{\sqrt{\frac{SS_n}{n-1}}}$

Considere-se a estatística de Student

$$T_{n-1} = \sqrt{n}\ \frac{\overline{X}_n}{\sqrt{\frac{SS_n}{n-1}}}$$

correspondente a uma amostra de dimensão n, $(X_1, \ldots, X_n)$, com os X_i variáveis aleatórias i.i.d. absolutamente contínuas, com densidade f.

Fazendo a transformação $\left(\overline{X}_n, SS_n\right) \to \left(T_{n-1}, U_n\right)$, com $U_n = \sqrt{SS_n}$, a transformação inversa é

$$\begin{cases} \overline{X}_n = \frac{U_n T_{n-1}}{\sqrt{n(n-1)}} \\ \\ SS_n = U_n^2 \end{cases} \Longrightarrow |J| = \frac{2U_n^2}{\sqrt{n(n-1)}}$$

Então, atendendo à expressão recursiva (8.4), a função den-

sidade de probabilidade de T_{n-1} é da forma

$$f_{T_{n-1}}(t) = \frac{2^{n-1}}{\sqrt{n-1}} \left[\prod_{i=3}^{n} \left(1 - \xi_i^2\right)^{\frac{i-4}{2}} \right] \times \\ \times \int_0^{+\infty} u^{n-1} \prod_{i=1}^{n} f\left(\left(\frac{t}{\sqrt{n(n-1)}} + \alpha_{i,n} \right) u \right) \mathrm{d}u \tag{8.5}$$

com os $\alpha_{i,n} = \alpha_{i,n}(\xi)$ definidos em (8.4), sempre que a existência dos $\xi = \xi_i = \xi_i(f_X)$ independentes de (w, s) esteja garantida.

Mesmo que este último requisito não esteja verificado senão aproximadamente — por exemplo, caso a função densidade de probabilidade f da distribuição parente seja aproximadamente simétrica em torno de zero e dependente de um parâmetro de escala, como discutimos após introduzir a hipótese H_2 acima — (8.5) é um candidato a considerar como expressão aproximada da função densidade de probabilidade de T_{n-1}, havendo eventualmente que multiplicar por uma constante normalizadora.

É fácil verificar que uma variável aleatória X é simétrica se e só se $X \overset{\mathrm{d}}{=} BX$, onde

$$B = \begin{cases} -1 & 1 \\ \\ \frac{1}{2} & \frac{1}{2} \end{cases}$$

é independente de X. De facto,

$$\varphi_B(t) = \mathbb{E}\left(\mathrm{e}^{\mathrm{i}tX} \right) = \frac{1}{2}\ \mathrm{e}^{-\mathrm{i}t} + \frac{1}{2}\ \mathrm{e}^{\mathrm{i}t} = \cos t.$$

Por outro lado, se $Z = XY$, com X e Y independentes, então

$$\begin{aligned}\varphi_{Z}(t) &= \mathbb{E}\left(\mathrm{e}^{\mathrm{i}tZ}\right) = \int_{\mathbb{R}} \mathbb{E}\left(\mathrm{e}^{\mathrm{i}tZ}|Y=y\right)\,\mathrm{d}F_{Y}(y) = \\ &= \int_{\mathbb{R}} \mathbb{E}\left(\mathrm{e}^{\mathrm{i}(ty)X}\right)\,\mathrm{d}F_{Y}(y) = \int_{\mathbb{R}} \varphi_{X}(ty)\,\mathrm{d}F_{Y}(y).\end{aligned}$$

Assim, se $X = BX$, então

$$\varphi_{X}(t) = \int_{\mathbb{R}} \cos(tx)\,\mathrm{d}F_{X}(x).$$

Ora

$$\begin{aligned}\varphi_{X}(t) &= \mathbb{E}\left(\mathrm{e}^{\mathrm{i}tX}\right) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}F_{X}(x) = \\ &= \int_{\mathbb{R}} \cos(tx)\,\mathrm{d}F_{X}(x) + \mathrm{i}\int_{\mathbb{R}} \sin(tx)\,\mathrm{d}F_{X}(x) = \\ &= \int_{\mathbb{R}} \cos(tx)\,\mathrm{d}F_{X}(x)\end{aligned}$$

se e só se F_{X} for a função de distribuição de uma variável aleatória simétrica em torno de zero, isto é, se é só se $F_{X}(x) = 1-F_{X}(-x)$ (caso a variável aleatória seja absolutamente contínua com função densidade de probabilidade f_{X}, $f_{X}(x) = f_{X}(-x)$).

Consequentemente, se $Z = XY$ com X e Y independentes, se X for simétrica, segue-se que $Z = XY = (BX)Y = B(XY)$ com B independente de XY, pelo que Z é necessariamente simétrica.

Assim, se a distribuição parente for simétrica, também T_{n-1} é simétrica. Da mesma forma, a simetria dos X_i implica a simetria de $\overline{X}_n$, e assim se a distribuição parente for simétrica o mesmo acontece com $f_n(w,s)$ no argumento w.

Na expressão (8.5),

$$f\left(\left(\frac{t}{\sqrt{n(n-1)}}+\alpha_{i,n}\right)u\right)=f\left(\left(\frac{-t}{\sqrt{n(n-1)}}+\alpha_{i,n}\right)u\right)$$

como $f(t)=f(-t)$ é par caso o conjunto dos coeficientes α_{in}, $i=1,\ldots,n$ goze da propriedade de simetria,

$$\left\{\alpha_{i,n}\right\}_{i=1}^{n}=\left\{-\alpha_{n+1-i,n}\right\}_{i=1}^{n}.$$

Consequentemente, f simétrica e $\{\alpha_{in}\}_{i=1}^{n}$ simétrica arrasta que

$$f_{T_{n-1}}(t)=f_{T_{n-1}}(-t)$$

que como atrás discutimos deve ser válido sob hipótese de a distribuição parente ser simétrica. Por razões análogas, a simetria de $\left\{\alpha_{i,n}\right\}_{i=1}^{n}=\left\{-\alpha_{n+1-i,n}\right\}_{i=1}^{n}$ arrasta a simetria de $f_n(w,s)$ como função de w, caso a distribuição parente seja simétrica.

Acabamos de mostrar que $\left\{\alpha_{i,n}\right\}_{i=1}^{n}$ simétrico implica a simetria de $f_{T_{n-1}}$, sob hipótese de simetria da distribuição parente. Vejamos agora que esta simetria, além de suficiente, é também necessária.

Comecemos por recordar que

$$\alpha_{i,n} = \frac{\xi_n}{\sqrt{n(n-1)}} + \alpha_{i,n-1}\sqrt{1-\xi_n^2},\ \ i = 1, \ldots, n-1$$

$$\alpha_{n,n} = -\xi_n \sqrt{\frac{n-1}{n}}$$

com

$$\sum_{i=1}^{n} \alpha_{i,n} = 0 \qquad \text{e} \qquad \sum_{i=1}^{n} \alpha_{i,n}^2 = 1.$$

Da primeira daquelas condições é imediato que

$$\alpha_{i,n} - \alpha_{j,n} = \left(\alpha_{i,n-1} - \alpha_{j,n-1}\right)\sqrt{1-\xi_n^2},\ \ i, j = 1, \ldots, n-1.$$

Daqui decorre que se os $\left\{\alpha_{i,n-1}\right\}_{i=1}^{n-1}$ estiverem ordenados decrescentemente[(2)], o mesmo acontece a $\left\{\alpha_{i,n}\right\}_{i=1}^{n-1}$.

Para $n = 2$, tínhamos $\alpha_{1,2} = \frac{\sqrt{2}}{2}$, $\alpha_{2,2} = -\frac{\sqrt{2}}{2}$; segue-se então que

$$\alpha_{1,3} - \alpha_{2,3} = \sqrt{2\left(1-\xi_3^2\right)}$$

Como das expressões de $\alpha_{i,3}$ (implicando a simetria que $\alpha_{2,3} = 0$ e portanto $\alpha_{1,3} - \alpha_{2,3} = -\alpha_{3,3}$) que se seguem a (8.3)

$$\alpha_{1,3} - \alpha_{2,3} = \frac{2\xi_3}{\sqrt{6}}$$

---

(2) Escolhe-se a ordenação decrescente porque $\alpha_{n,n} = -\xi_n\sqrt{\frac{n-1}{n}}$ irá ser o valor mais baixo.

segue-se que

$$2\left(1-\xi_3^2\right)=\frac{4\xi_3^2}{6}$$

donde

$$\xi_3^2=\frac{3}{4}\,.$$

Consequentemente,

$$\alpha_{1,3}=\frac{\sqrt{2}}{2},\quad \alpha_{2,3}=0,\quad \alpha_{3,3}=-\frac{\sqrt{2}}{2}$$

e portanto $\left\{\alpha_{i,3}\right\}_{i=1}^{3}$ é de facto um conjunto simétrico. Além disso $\left\{\alpha_{i,3}\right\}_{i=1}^{3}$ está em progressão aritmética de razão $r_3=-\frac{\sqrt{2}}{2}$, e consequentemente $\left\{\alpha_{i,4}\right\}_{i=1}^{3}$ está em progressão aritmética de razão $r_4=-\frac{\sqrt{2}}{2}\sqrt{1-\xi_4^2}$.

Recorrendo ao anteriormente obtido,

$$\alpha_{i,4}=\alpha_{1,4}+(i-1)\,r_4,\ i=1,2,3$$

com $r_4=-\frac{\sqrt{2}}{2}\sqrt{1-\xi_4^2}$, e $\alpha_{4,4}=-\xi_4\sqrt{\frac{3}{4}}$. Decorre então, procedendo como para $n=3$, que

$$\xi_4^2=\frac{3}{5}$$

e

$$\alpha_{1,4}=\frac{3}{\sqrt{20}}\,,\quad \alpha_{2,4}=\frac{1}{\sqrt{20}}\,,\quad \alpha_{3,4}=-\frac{1}{\sqrt{20}}\,,\quad \alpha_{4,4}=-\frac{3}{\sqrt{20}}\,.$$

Sugere isto o seguinte resultado:

Teorema 8.2.1.

Sob validade de H_0 *e de* H_1, $\left\{\alpha_{i,n}\right\}_{i=1}^{n}$ *é um conjunto simétrico, em progressão aritmética de razão*

$$r_n = -\frac{2\sqrt{3}}{\sqrt{n\left(n^2-1\right)}}.$$

Os termos da progressão são

$$\alpha_{i,n} = (n+1-2i)\sqrt{\frac{3}{n\left(n^2-1\right)}}\,,\ i=1,\ldots,n \tag{8.6}$$

e

$$\xi_n^2 = \frac{3}{n+1}\,. \tag{8.7}$$

Demonstração:

Verificámos já que as expressões acima são válidas para $k=2$. Supondo agora a validade daquelas daquelas expressões para $k=3,4,\ldots,(n-1)$, vamos verificar que se segue que é válida para $k=n$.

Note-se em primeiro lugar que (8.5) implica que $\left\{\alpha_{i,n}\right\}_{i=1}^{n-1}$ é uma progressão aritmética. Trabalhando como para a obtenção de $p^2=\xi_3^2$, deduz-se que

$$\xi_n^2 = \frac{3}{n+1}\,.$$

Consequentemente,

$$-r_n = -r_2\sqrt{\frac{1}{4}}\sqrt{\frac{2}{5}}\sqrt{\frac{3}{6}}\cdots\sqrt{\frac{n-2}{n+1}} = \sqrt{\frac{2(n-2)!\,3!}{(n+1)!}} = \frac{2\sqrt{3}}{\sqrt{n\left(n^2-1\right)}}\,.$$

Segue-se que

$$\alpha_{n,n} = \sqrt{\frac{3(n-1)}{n(n+1)}}\,.$$

Mas então

$$\alpha_{1,n} + \alpha_{1,n} + r + \alpha_{1,n} + 2r + \ldots + \alpha_{1,n} + (n-2)r + \alpha_{n,n} = 0 \iff$$

$$\iff (n-1)\alpha_{1,n} + r_n\ \tfrac{n-1}{2}\ (n-2) - \sqrt{\tfrac{3(n-1)}{n(n+1)}} = 0 \iff$$

$$\iff \alpha_{1,n} = -r_n\ \tfrac{n-2}{2} - \sqrt{\tfrac{3}{n\left(n^2-1\right)}}$$

e então

$$\begin{aligned}\alpha_{1,n} &= \tfrac{2\sqrt{3}}{\sqrt{n\left(n^2-1\right)}}\ \tfrac{n-2}{2} + \sqrt{\tfrac{3}{n\left(n^2-1\right)}} = \\ &= \tfrac{\sqrt{3}}{\sqrt{n\left(n^2-1\right)}}\ \tfrac{n-2+1}{2}\ = \sqrt{\tfrac{3(n-1)}{n(n+1)}}\ = -\alpha_{n,n}.\end{aligned}$$

Concluimos assim que $\left\{\alpha_{i,n}\right\}_{i=1}^{n}$ é um conjunto simétrico; consequentemente $\alpha_{n,n} - \alpha_{n-1,n} = \alpha_{2,n} - \alpha_{1,n} = r_n$ pelo que $\left\{\alpha_{i,n}\right\}_{i=1}^{n}$ é uma progressão aritmética.

Uma vez que $\alpha_{1,n} = \sqrt{\frac{3(n-1)}{n(n+1)}}$ e a razão é $r_n = -\frac{2\sqrt{3}}{\sqrt{n\left(n^2-1\right)}}$ segue-se que

$$\begin{aligned}\alpha_{i,n} &= \sqrt{\tfrac{3(n-1)}{n(n+1)}} - (i-1)\ \tfrac{2\sqrt{3}}{\sqrt{n\left(n^2-1\right)}} = \\ &= \tfrac{\sqrt{3}(n-1)-(i-1)2\sqrt{3}}{\sqrt{n\left(n^2-1\right)}} = \sqrt{\tfrac{3}{n\left(n^2-1\right)}}\ (n+1-2i),\end{aligned}$$

o que conclui a demonstração. □

A expressão (8.5) é, como expressão exacta, bastante restritiva, uma vez que é válida no pressuposto da existência de $\xi(f, w, s) = \xi(f)$. No entanto, como expressão aproximada para $f_{T_{n-1}}$ pode ter grande interesse, como mais adiante discutiremos.

Parece-nos assim indispensável, no entanto, mostrar que a expressão conduz à fórmula exacta para algumas distribuições parentes, mostrando consequentemente que a nossa abordagem não assume pressupostos irrealistas.

Suponha-se então que a distribuição parente é a gaussiana, isto é com função densidade de probabilidade $f(x) \propto \exp\left(-\frac{x^2}{2}\right)$. Então, de (8.5),

$$f_{T_{n-1}}(t) \propto \int_0^{+\infty} u^{n-1} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\left[\left(\frac{t}{\sqrt{n(n-1)}} + \alpha_{i,n}\right) u\right]^2\right\} \, \mathrm{d}u$$

$$\propto \int_0^{+\infty} u^{n-1} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{t^2 u^2}{n(n-1)} + \alpha_{i,n}^2 u^2 + 2\,\frac{t\,\alpha_{i,n} u^2}{\sqrt{n(n-1)}}\right]\right\} \, \mathrm{d}u$$

e como $\sum_{i=1}^{n} \alpha_{i,n} = 0$ e $\sum_{i=1}^{n} \alpha_{i,n}^2 = 1$, vem

$$f_{T_{n-1}}(t) \propto \int_0^{+\infty} u^{n-1} \exp\left\{-\frac{u^2}{2}\left(\frac{t^2}{n-1} + 1\right)\right\} \, \mathrm{d}u$$

ou seja $f_{T_{n-1}}(t) \propto \frac{1}{\left(1+\frac{t^2}{n-1}\right)^{\frac{n}{2}}}$, que é, para $(n-1)$ graus de liberdade, a expressão da função densidade de probabilidade da variável aleatória t de Student.

Como nota final, aponte-se que o trabalho de Rocha (1995) resolve satisfatoriamente (de forma aproximada, claro) o problema da studentização numa população simétrica. por outro lado, o problema de duas populações / duas amostras está longe de ser resolvido. Ainda que métodos similares aos que emprega no caso de uma amostra possam funcionar no caso de duas amostras com iguais variâncias — o que não foi estabelecido —, o problema geral não teve ainda solução.

Os cálculos envolvidos são de tal modo complexos que justificam plenamente o desenvolvimento de metodologias não paramétricas quando a população parente não é gaussiana. Apenas em situações em que grande rigor é necessário se justificaria proceder ao cálculo explícito das expressões. Como já referimos, é uma área em plena actividade, em que Mehendale and Kalluraya (2011) obtiveram recentemente êxito assinalável.

No que se refere a studentização — no sentido mais clássico, $T = \sqrt{n}\,\frac{\overline{X}-\mu}{S}$ — no caso de população parente assimétrica, o leitor interessado pode começar por Johnson (1978). Observamos que no caso de comparação de médias com base em duas amostras independentes, tem sido notado uma reversão do sentido da assimetria, e uma regularização que julgamos que se deva ao efeito simetrizador das diferenças. Martins (2005) e Rocha e Martins (2005) abordam a questão detalhadamente, e têm preciosas indicações bibliográficas para quem quiser aventurar-se neste campo.

Complementos

Em complemento ao Capítulo 8, no Apêndice A explicitamos a expressão de funções densidade de probabilidade de T_1 em alguns modelos usuais.

Não temos um apreço particular pelas transformações de dados, apesar da eloquente defesa de Tukey (1957). Pensamos que mudar de forma radical a estrutura do problema é questão que requer muita ponderação e prudência. Apresentamos no entanto, no Apêndice B, um caso de sucesso, inspirando-nos no trabalho de Box and Hill (1974); o facto de os autores usarem mínimos quadrados ponderados não é decerto irrelevante para a qualidade dos resultados que obtêm.

No Apêndice C registamos uma outra foma de estabelecer a aproximação de Welch-Satterthwaite a uma combinação linear de gamas, por poder ser de alguma utilidade a quem queira investigar questões que — *hélas* — deixámos por resolver.

Apêndice A

Exemplos de Densidades de T_1 e de T_2

O cálculo da função densidade de probabilidade da variável studentizada $T_{n-1} = \sqrt{n(n-1)}\,\frac{\overline{X}_n}{\sqrt{SS_n}}$ no caso $n = 2$ não é em geral insuportavelmente difícil (mas o exemplo de parente Cauchy mostra que também nem sempre é fácil).

Por outro lado, a partir de $n = 3$, mesmo em populações uniformes, as dificuldades são de monta, basta consultar Perlo (1933).

Pode interessar alguns leitores, para melhor avaliarem a situação, percorrer a exemplificação que segue, exibindo a função densidade de probabilidade de T_1 para algumas populações de uso corrente.

Da expressão (8.1), da função densidade conjunta de $\left(\overline{X}_2, SS_2\right)$

$$f_{(\overline{X}_2, SS_2)}(w,s) = \sqrt{\frac{2}{s}}\, f\!\left(w + \sqrt{\frac{s}{2}}\right) f\!\left(w - \sqrt{\frac{s}{2}}\right),\ w \in \mathbb{R},\ s > 0,$$

considerando

$$T_1 = \frac{\sqrt{2}\,\overline{X}_2}{\sqrt{SS_2}}$$

e a transformação $\left(\overline{X}_2, SS_2\right) \to (T,U)$ segue-se que

$$\left\{ \begin{array}{l} T = T_1 = \frac{\sqrt{2}\,\overline{X}_2}{\sqrt{SS_2}} \\ \\ U = \sqrt{SS_2} \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} \overline{X}_2 = \frac{UT}{\sqrt{2}} \\ \\ SS_2 = U^2 \end{array} \right.$$

$$\Longrightarrow |J| = \sqrt{2}\,U^2 \wedge U > 0 \wedge T \in \mathbb{R}$$

pelo que

$$\begin{aligned} f_{(T,U)}(t,u) &= f_{(\overline{X}_2, SS_2)}\!\left(\tfrac{ut}{\sqrt{2}}, u^2\right)\sqrt{2}u^2 = \\ &= \sqrt{\tfrac{2}{u^2}}\, f\!\left(\tfrac{ut}{\sqrt{2}} + \sqrt{\tfrac{u^2}{2}}\right) f\!\left(\tfrac{ut}{\sqrt{2}} - \sqrt{\tfrac{u^2}{2}}\right)\sqrt{2}\,u^2. \end{aligned}$$

Consequentemente,

$$f_{T_1}(t) = 2 \int_0^{+\infty} u\, f\!\left(\frac{u\,(t+1)}{\sqrt{2}}\right) f\!\left(\frac{u\,(t-1)}{\sqrt{2}}\right) \mathrm{d}u. \qquad \text{(A.1)}$$

(Como alternativa, podíamos ter trabalhado directamente com $f_{(X_1,X_2)}(x_1,x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ e $T_1 = \frac{X_1+X_2}{|X_1-X_2|}$.)

A aplicação de (A.1) permite calcular a expressão exacta da função densidade de probabilidade de T_1 no caso de diversas distribuições parentes correntes. Apresentamos alguns exemplos, em que se pode observar que a studentização clássica com parente não gaussiana tem resultados muito diversos, e bem diferentes do que acontece no caso de parente gaussiana — recordamos que nesse caso $T_1 \frown Cauchy(0,1)$.

Uniforme padrão

Seja X com função densidade de probabilidade $f(x) = \mathrm{I}_{(0,1)}(x)$. Assim, como $X_1,\ X_2 > 0 \Rightarrow X_1 + X_2 > |X_1 - X_2| \Rightarrow t > 1$, vem

$$u > 0 \wedge 0 < \frac{t-1}{\sqrt{2}}\, u < 1 \Rightarrow 0 \wedge u < \frac{\sqrt{2}}{t+1}$$

e consequentemente,

$$f_{T_1}(t) = 2 \int_0^{\frac{\sqrt{2}}{t+1}} u \, \mathrm{d}u = u^2 \Big]_0^{\frac{\sqrt{2}}{t+1}} \ \mathrm{I}_{(1,+\infty)}(t) = \frac{2}{(t+1)^2} \ \mathrm{I}_{(1,+\infty)}(t)$$

cujo gráfico está esboçado na Fig. A.1.

***Beta*(2, 2) padrão**

No caso de a população parente ter função densidade de probabilidade $f(x) = 6\,x\,(1-x)\ \mathrm{I}_{(0,1)}(x)$, o tratamento é análogo ao

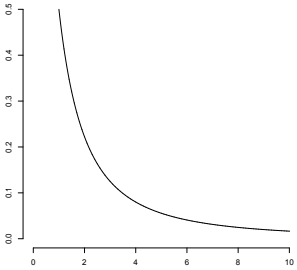

Figura A.1: Densidade de T_1, parente uniforme padrão.

acima exposto para a população $Beta(1,1)$.

$$f_{T_1}(t) = 36\,(t^2-1)\int_0^{\frac{\sqrt{2}}{t+1}} u^3(1-u\,\frac{t+1}{\sqrt{2}})\,(1-u\,\frac{t-1}{\sqrt{2}})\;\mathrm{d}u =$$

$$= 36\,(t-1)\left[\frac{1}{(1+t)^3} - \frac{8t}{5\,(1+t)^4} + \frac{2\,(t^2-1)}{3\,(1+t)^5}\right]\;\mathrm{I}_{(1,+\infty)}(t)$$

Na Fig. A.2 representa-se o gráfico da densidade de T_1 no caso de a população parente ser $Beta(2,2)$.

Uniforme em $(-1,1)$

Seja X com função densidade de probabilidade

$$f(x) = \frac{1}{2}\;\mathrm{I}_{(-1,1)}(x).$$

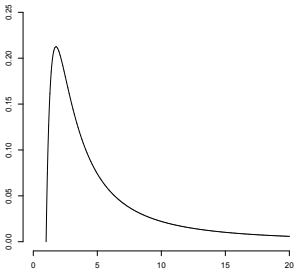

Figura A.2: Densidade de T_1, parente $Beta(2,2)$.

De igual modo,

$$u > 0 \wedge -1 < \frac{t+1}{\sqrt{2}}\, u < 1 \wedge -1 < \frac{t-1}{\sqrt{2}}\, u < 1 \Rightarrow 0 < u < \frac{\sqrt{2}}{t+1}$$

pelo que

$$f_{T_1}(t) = 2 \int_0^{\frac{\sqrt{2}}{t+1}} u\, \frac{1}{4}\, \mathrm{d}u = \frac{1}{4}\, u^2 \Big]_0^{\frac{\sqrt{2}}{t+1}} = \frac{1}{2(|t|+1)^2}\,,\ t \in \mathbb{R},$$

a que corresponde o gráfico representado na Fig. A.3.

Beta$(2,2)$ com suporte $(-1,1)$

Se $X = 2Y - 1$ com $Y \frown Beta(2,2)$ padrão, ou seja uma $Beta(2,2)$ com localização -1 e escala 2, com função densidade de probabilidade $f_X(x) = \frac{3}{4}\,(1-x^2)\ \mathrm{I}_{(-1,1)}(x)$, procedendo com

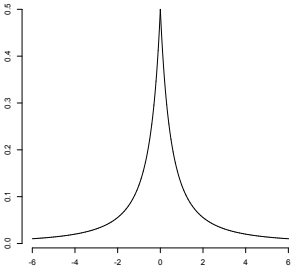

Figura A.3: Densidade de T_1, parente uniforme em $[-1, 1]$.

no caso da uniforme com aquele suporte, obtém-se

$$f_{T_1}(t) = \frac{3}{8} \frac{1 + 4|t| + t^2}{(1 + |t|)^4} \ \mathrm{I}_{\mathbb{R}}(t),$$

cujo gráfico se representa na Fig. A.4.

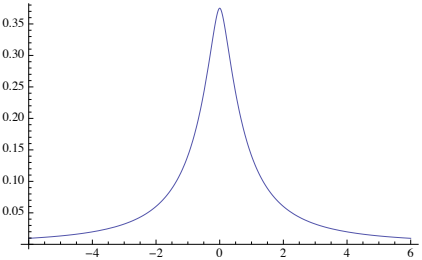

Figura A.4: Densidade de T_1, parente $Beta(2, 2; 1, 1)$.

Exponencial

Seja X com função densidade de probabilidade

$$f(x) = \frac{1}{\delta}\ \mathrm{e}^{-\frac{x}{\delta}}\ \mathrm{I}_{(0,\infty)}(x).$$

Assim, como $t > 1$

$$u > 0 \wedge \frac{u(t+1)}{\sqrt{2}} > 0 \wedge \frac{u(t-1)}{\sqrt{2}} > 0 \Rightarrow u > 0$$

e como tal,

$$\begin{aligned} f_{T_1}(t) &= 2\int_0^\infty u\,\frac{1}{\delta}\ \mathrm{e}^{-\frac{u(t+1)}{\delta\sqrt{2}}}\ \frac{1}{\delta}\ \mathrm{e}^{-\frac{u(t-1)}{\delta\sqrt{2}}}\ \mathrm{d}u \\ &= \frac{2}{\delta^2}\ \int_0^\infty u\ \mathrm{e}^{-\frac{ut\sqrt{2}}{\delta}}\ \mathrm{d}u,\ t > 1. \end{aligned}$$

Procedendo à mudança de variável

$$v = \frac{ut\sqrt{2}}{\delta}\ (\text{com } t > 1)) \Rightarrow v \in (0, +\infty)$$

e como o valor absoluto do jacobiano da transformação é $|J| = \frac{\delta}{t\sqrt{2}}$, segue-se,

$$\begin{aligned} f_{T_1}(t) &= \tfrac{2}{\delta^2}\int_0^\infty \frac{\delta v}{t\sqrt{2}}\ \mathrm{e}^{-v}\,\frac{\delta}{t\sqrt{2}}\,\mathrm{d}v = \\ &= \tfrac{1}{t^2}\int_0^\infty v\ \mathrm{e}^{-v}\,\mathrm{d}v = \frac{1}{t^2}\,\Gamma(2) = \frac{1}{t^2},\ t > 1 \end{aligned}$$

pelo que T_1 tem distribuição de Pareto com índice 1 (evidentemente, o resultado não depende de δ); o gráfico da função densidade de probabilidade está representado na Fig. A.5

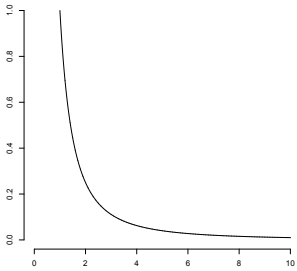

Figura A.5: Densidade de T_1, parente exponencial.

$\boldsymbol{Gama(\alpha, \delta)}$

Mais geralmente, se $X \frown Gama(\alpha)$, um tratamento idêntico ao apresentado para a $Gama(1)$ estabelece que

$$f_{T_1}(t) = \frac{\Gamma(2\,\alpha)}{2^{2(\alpha-1)}\,\Gamma^2(\alpha)}\,\frac{\left(t^2-1\right)^{\alpha-1}}{t^{2\,\alpha}}\;\mathrm{I}_{(1,+\infty)}(t).$$

Note que

$$\frac{\Gamma(2\,\alpha)}{2^{2\alpha-2}\,\Gamma^2(\alpha)}\int_1^{\infty}\frac{\left(t^2-1\right)^{\alpha-1}}{t^{2\alpha}}\,\mathrm{d}t = \frac{\Gamma(2\,\alpha)}{2^{2\alpha-1}\,\Gamma^2(\alpha)}\int_0^{1} w^{-\frac{1}{2}}(1-w)^{\alpha-1}\,\mathrm{d}w = 1$$

por aplicação directa da fórmula da duplicação do argumento da função gama. Esboça-se na Fig. A.6 o gráfico da densidade de T_1 no caso de a população parente ser $Gama(2.5, 1)$.

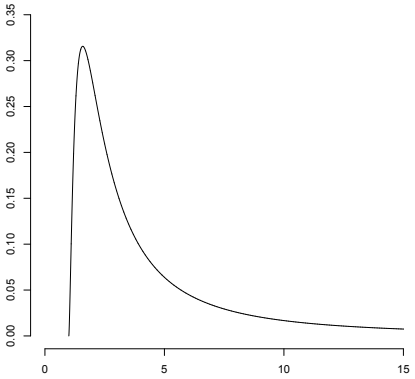

Figura A.6: Densidade de T_1, parente $Gama(2.5, 1)$.

Laplace

Seja X com função densidade de probabilidade

$$f(x) = \frac{1}{2}\ \mathrm{e}^{-|x|}\ \mathrm{I}_{\mathbb{R}}(x).$$

Tendo em conta que $t \in \mathbb{R}$, então

$$u > 0 \wedge \frac{u(t+1)}{\sqrt{2}} \in \mathbb{R} \wedge \frac{u(t-1)}{\sqrt{2}} \in \mathbb{R} \Rightarrow u > 0$$

e consequentemente

$$f_{T_1}(t) = 2 \int_0^\infty u\, \frac{1}{2}\ \exp\left(-\left|\frac{u(t+1)}{\sqrt{2}}\right|\right) \frac{1}{2}\ \exp\left(-\left|\frac{u(t-1)}{\sqrt{2}}\right|\right)\ \mathrm{d}u.$$

Como

- para $t > 1$,

$$\exp\left(-\left|\frac{u(t+1)}{\sqrt{2}}\right|\right)\exp\left(-\left|\frac{u(t-1)}{\sqrt{2}}\right|\right) = \exp\left(-\frac{2ut}{\sqrt{2}}\right) = \mathrm{e}^{-\sqrt{2}tu}$$

- para $-1 \leq t \leq 1$

$$\exp\left(-\left|\frac{u(t+1)}{\sqrt{2}}\right|\right)\exp\left(-\left|\frac{u(t-1)}{\sqrt{2}}\right|\right) = \exp\left(-\frac{2u}{\sqrt{2}}\right) = \mathrm{e}^{-\sqrt{2}u}$$

- para $t < -1$

$$\exp\left(-\left|\frac{u(t+1)}{\sqrt{2}}\right|\right)\exp\left(-\left|\frac{u(t-1)}{\sqrt{2}}\right|\right) = \exp\left(+\frac{2ut}{\sqrt{2}}\right) = \mathrm{e}^{+\sqrt{2}tu}$$

obtemos

$$f_{T_1}(t) = \begin{cases} \frac{1}{4t^2}\,, & |t| > 1 \\ \\ \frac{1}{4}\,, & |t| \leq 1 \end{cases}$$

cuja representação gráfica é esboçada na Fig. A.7.

Pareto com índice 1

Seja X com função densidade de probabilidade

$$f(x) = \frac{1}{x^2}\,\mathrm{I}_{(1,+\infty)}(x).$$

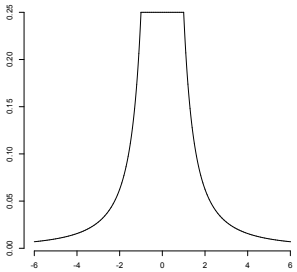

Figura A.7: Densidade de T_1, parente Laplace.

Assim, como $t > 1$,

$$u > 0 \wedge \frac{u(t+1)}{\sqrt{2}} > 1 \wedge \frac{u(t-1)}{\sqrt{2}} > 1 \Rightarrow u > \frac{\sqrt{2}}{(t-1)}$$

e consequentemente,

$$f_{T_1}(t) = 2 \int_{\frac{\sqrt{2}}{t-1}}^{\infty} u \left[\frac{u(t+1)}{\sqrt{2}}\right]^{-2} \left[\frac{u(t-1)}{\sqrt{2}}\right]^{-2} \mathrm{d}u =$$

$$= \frac{8}{(t+1)^2(t-1)^2} \int_{\frac{\sqrt{2}}{t-1}}^{\infty} \frac{1}{u^3}\, \mathrm{d}u = \frac{2}{(t+1)^2}\,,\ t \geq 1.$$

a que corresponde o gráfico da Fig. A.8.

Porventura o leitor notou que populações parentes Pareto com índice 1 e em $Uniforme(0,1)$ levam a variáveis studentizadas idênticas em distribuição.

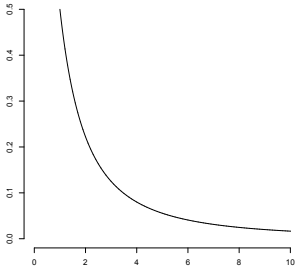

Figura A.8: Densidade de T_1, parente Pareto(1)

Tal deve-se ao facto de

$$X \frown Pareto(1) \Longrightarrow 1 - \frac{1}{X} \frown Uniforme(0,1),$$

devido à transformação uniformizante, e consequentemente

$$1 - \left(1 - \frac{1}{X}\right) \stackrel{d}{=} \frac{1}{X} \frown Uniforme(0,1),$$

caso especial da propriedade

$$X \frown Beta(p,q) \Longrightarrow 1 - X \frown Beta(q,p)$$

das distribuições beta.

Assim, considerando $X_1, X_2 \frown Pareto(1)$ independentes, tem-se que a variável studentizada para a uniforme

$$\frac{\frac{1}{X_1} + \frac{1}{X_2}}{\left|\frac{1}{X_1} - \frac{1}{X_2}\right|} = \frac{X_1 + X_2}{|X_1 - X_2|}$$

coincide com a variável studentizada no caso de população parente Pareto com índice 1.

Pareto com índice $\alpha(>0)$ qualquer.

Seja X com função densidade de probabilidade

$$f(x) = \frac{\alpha}{x^{\alpha+1}} \, \mathrm{I}_{(1,+\infty)}(x).$$

Assim, como $t \geq 1$

$$u > 0 \wedge \frac{u(t+1)}{\sqrt{2}} > 1 \wedge \frac{u(t-1)}{\sqrt{2}} > 1 \Rightarrow u > \frac{\sqrt{2}}{(t-1)}$$

e consequentemente,

$$\begin{aligned} f_{T_1}(t) &= 2 \int_{\frac{\sqrt{2}}{t-1}}^{\infty} u \left[\frac{\alpha 2^{\frac{\alpha+1}{2}}}{u^{\alpha+1}(t+1)^{\alpha+1}} \right] \left[\frac{\alpha 2^{\frac{\alpha+1}{2}}}{u^{\alpha+1}(t-1)^{\alpha+1}} \right] \mathrm{d}u = \\ &= \frac{2^{\alpha+2}\alpha^2}{(t+1)^{\alpha+1}(t-1)^{\alpha+1}} \int_{\frac{\sqrt{2}}{t-1}}^{\infty} \frac{1}{u^{2\alpha+1}} \, \mathrm{d}u = \\ &= \frac{2^{\alpha+1}\alpha}{(t+1)^{\alpha+1}(t-1)^{\alpha+1}} \, \frac{(t-1)^{2\alpha}}{2^{\alpha}} = 2\alpha \frac{(t-1)^{\alpha-1}}{(t+1)^{\alpha+1}} \, , \; t \geq 1 \end{aligned}$$

de que o resultado anterior é um caso particular — e o mesmo podemos dizer com o que obtivemos em populações exponenciais, quando consideramos a exponencial como caso limite da Pareto generalizada quando o índice tende para 0.

O aspecto do gráfico da função densidade de probabilidade, para $\alpha = 2.5$, está respresentado na Fig. A.9.

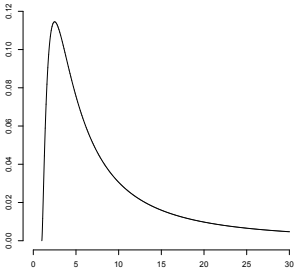

Figura A.9: Densidade de T_1, parente Pareto(1)

Cauchy.

Seja X com função densidade de probabilidade

$$f(x) = \frac{1}{\pi\ \left(1+x^2\right)},\ x \in \mathbb{R}.$$

Assim, dado que $t \in \mathbb{R}$, segue-se que

$$u > 0 \wedge \frac{u(t+1)}{\sqrt{2}} \in \mathbb{R} \wedge \frac{u(t-1)}{\sqrt{2}} \in \mathbb{R} \Rightarrow u > 0$$

e consequentemente

$$f_{T_1}(t) = 2\int_0^\infty u f\left(\frac{u(t+1)}{\sqrt{2}}\right) f\left(\frac{u(t-1)}{\sqrt{2}}\right)\,\mathrm{d}u =$$

$$= \frac{2}{\pi^2}\int_0^\infty \frac{u}{\left[1+\frac{u^2}{2}\ (t+1)^2\right]\left[1+\frac{u^2}{2}\ (t-1)^2\right]}\,\mathrm{d}u.$$

Procedendo à mudança de variáveis $\frac{u}{2}^2 = \frac{v}{(t-1)^2}$, com $t \neq 1$, o valor absoluto do jacobiano da transformação é $|J| = \frac{1}{|t-1|\sqrt{2v}}$ e $u \in (0, +\infty) \Rightarrow v \in (0, +\infty)$.

Consequentemente

$$f_{T_1}(t) = \frac{2}{\pi^2 (t-1)^2} \int_0^{+\infty} \frac{1}{\left[1 + \left(\frac{t+1}{t-1}\right)^2 v\right](1+v)} \, \mathrm{d}v,$$

cuja integração por decomposição em soma de fracções algébricas com denominador linear origina

$$f_{T_1}(t) = \frac{1}{2\pi^2 t} \ln\left(\frac{1 + \left(\frac{t+1}{t-1}\right)^2 v}{1+v}\right)\Bigg]_0^{+\infty} = \frac{1}{2\pi^2 t} \ln\left(\frac{t+1}{t-1}\right)^2, \ |t| \neq 1,$$

cuja representação gráfica se apresenta na Fig. A.10.

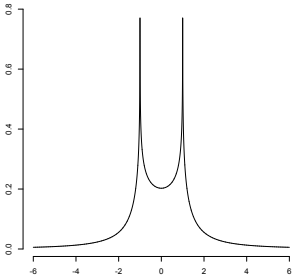

Figura A.10: Densidade de T_1, parente Cauchy

Como comentário final, saliente-se que mesmo para $n = 2$, quando a parente não é gaussiana chega-se a uma expressão analítica de T_1 bem diversa de t_1, que como observámos tem distribuição de Cauchy padrão. O peso das caudas é relevante, naturalmente, para a forma da estatística studentizada.

Mesmo com parentes simétricas — uma hipótese que Efron (1969) investigou com a sua habitual clarividência, mostrando ser muito mais forte do que aparenta — podemos, como os gráficos patentemente ilustram, chegar a densidades com uma forma bem diversa do habitual sino típico de muitas situações unimodais. Para valores mais elevados de n, basta compulsar o resultado de Perlo (1933) para parente rectangular no suporte $(-1, 1)$ e $n = 3$ (cujo gráfico se representa na Fig. A.11) para se sentir uma vertigem:

$$f_{T_2}(t) = \begin{cases} \frac{\sqrt{3}}{2(4-t^2)\sqrt{1-t^2}} \left(1 - \frac{9t^2}{4-t^2}\right) + & \\ \qquad + \frac{\sqrt{27}\,(t^2+2)}{\sqrt{(4-t^2)^5}} \operatorname{arctgh}\left(\sqrt{\frac{1-t^2}{4-t^2}}\right) & 0 \leq |t| < \frac{1}{2} \\ \frac{-9}{4\,(|t|+1)(t^2-4)} \left(\frac{1}{|t|+1} + \frac{3|t|}{t^2-4}\right) + & \\ \qquad + \frac{\sqrt{27}\,(t^2+2)}{\sqrt{(t^2-4)^5}} \operatorname{arctg}\left(\frac{\sqrt{t^2-4}}{\sqrt{3}\,(t+2)}\right) & \frac{1}{2} < |t| \end{cases}.$$

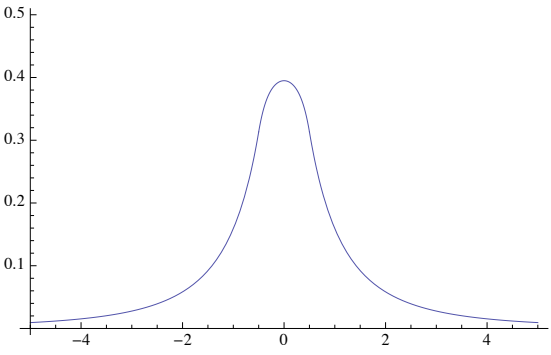

Figura A.11: Densidade de T_2, parente $Uniforme(-1, 1)$

Apêndice B

Heterocedasticidade e Transformações Estabilizadoras da Variância

No estudo do problema da localização/escala em populações gaussianas, as dificuldades surgem quando se verificam situações de heterocedasticidade. Procurámos resolver o problema da heterogeneidade de variâncias levando em conta as diferenças existentes entre os grupos. Para isso é necessário avaliar as variâncias de cada grupo, usando estimadores convenientes, que depois são combinados num estimador ponderado global.

Uma abordagem alternativa consiste em eliminar previamente as diferenças entre as dispersões dos grupos, nivelando-as por intermédio de uma transformação de dados adequada. Na-

turalmente, uma das questões que se põem relativamente a esta estratégia é a escolha da transformação apropriada. Será exposto um método para a escolha da transformação de Box-Cox apropriada para "estabilizar as variâncias".

No entanto, esta estratégia conceptualmente simples é também contestável, pois ao ser aplicada uma transformação não linear aos dados a estrutura destes é modificada, não só quanto à variância como relativamente à sua distribuição e relação com outras variáveis. Por exemplo, se considerações empíricas nos levarem a admitir uma distribuição normal para os dados originais, esta já não é adequada para os valores transformados, a menos que seja uma transformação linear, que não altera a forma.

Assim, em vez da transformação de variáveis, usa-se com frequência uma abordagem alternativa de estimação, em que são atribuídos diferentes pesos aos grupos, consoante a sua variabilidade, interpretada como inverso da informação. É exposto um método para a selecção dos pesos na estimação por mínimos quadrados ponderados proposto por Box e Hill (1974). A sua escolha dos pesos faz intervir métodos usuais de inferência bayesiana. O trabalho de Box e Hill é exemplificado num contexto de regressão, mas a sua adaptação aos modelos de análise de variância é imediata. Depois de expormos esta abordagem, terminamos com algumas considerações gerais sobre a desirabilidade de abordar o problema das médias com um teste prévio à igualdade de variâncias, e sobre transformações destinadas a conseguir o objectivo de homogeneizar variâncias.

Considerações gerais sobre testes sobre desigualdade de variâncias e sobre transformações destinadas a homogeneizar as variâncias

Alguns autores aconselham a aplicação rotineira de testes destinados a detectar desigualdade das variâncias, mas os testes paramétricos usuais são sensíveis a afastamentos da gaussianidade (Neter *et al.*, 1996, p. 763); por outro lado, o teste de Fligner-Killeen (implementado no *package* R) oferece uma alternativa não paramétrica para testar a hipótese de homogeneidade de variâncias. Em compensação, os testes t e z são relativamente robustos, tolerando afastamentos ligeiros da gaussianidade. Moser and Stevens (1992) aconselham mais investimento em ensinar o que designam por teste de Smith/Welch/Satterthwaite (SWS) e menos em testes rotineiros sobre igualdade de variâncias[(1)].

As transformações de dados são usadas com alguma regularidade em estatística. Por exemplo, num estudo de regressão entre duas variáveis pode ser conveniente transformar uma delas a fim de eliminar alguma curvatura sugerida pelo gráfico e tornar aceitável um modelo rectilíneo.

A raiz quadrada e o logaritmo são duas das funções mais usadas em transformações de variáveis. Ambas pertencem à classe mais vasta das transformações-potência de Box-Cox:

---

(1) Recomendação que gerou alguma polémica, cf. Gans (1991). Moser e Stevens terminam o seu artigo dizendo: "[...] *the SWS test is appropriate in all cases where the variance ratio is unknown. Therefore, when teaching the two-sample means test, more effort should be spent learning the qualities of the SWS test and less emphasis should be placed on the homogeneity of variance assumption*".

$$y_i^* = \begin{cases} \dfrac{y_i^{\phi} - 1}{\phi} & \phi \neq 0 \\ \\ \ln y_i & \phi = 0 \end{cases} .$$

(Mais geralmente, as transformações de Box-Cox são definidas a menos de translações e mudanças de escala; é nesse sentido que a raiz quadrada também está incluída.)

Há um padrão indiciador de variâncias desiguais: a dispersão dos resíduos tende a crescer com a localização. Uma forma simples de detectar este padrão consiste em desenhar o gráfico que tem como abcissas os valores ajustados $\widehat{y}_{jk}$ e como ordenadas os resíduos estandardizados — o padrão referido é revelado pelo aspecto cuneiforme da disposição dos pontos.

Nesta situação uma transformação-potência pode resolver a situação. Hoaglin *et al.* (1981, Cap. III e Cap. VIII) sugerem a recta de diagnóstico apropriada.

Dada a especifidade de estarmos a tentar usar uma transformação-potência que simultaneamente transforme os dados e mantenha a gaussianidade dos resíduos, vamos alongar um pouco mais os nossos comentários:

O objectivo é encontrar uma função h dos dados por tal forma que o modelo

$$h(X_{jk}) = \mu^* + \tau_j^* + \varepsilon_{jk}^*$$

seja apropriado, com os erros $\varepsilon_{jk}^* \frown Gaussiana(0, \sigma)$ mutuamente independentes, $k = 1, \ldots, n_j$, $j = 1, \ldots, k$.

No caso de σ_j^2 exibir um padrão claro de crescimento com $\mathbb{E}\left[X_{jk}\right] = \mu + \tau_j$, $j = 1, \ldots, k$, essa relação pode muitas vezes

ser expressa sob a forma

$$\sigma_j^2 = k\left(\mu + \tau_j\right)^q,$$

q e k constantes apropriadas. Neste caso, a transformação h que nos interessa é

$$h(x_{jk}) = \begin{cases} x_{jk}^{1-\frac{q}{2}} & \text{se } q \neq 2 \\ \ln\left(x_{jk}\right) & \text{se } q = 2 \text{ e todos os } x_{jk} \neq 0 \\ \ln\left(x_{jk} + 1\right) & \text{se } q = 2 \text{ e alguns dos } x_{jk} = 0 \end{cases}$$

Em geral uma boa aproximação de q pode ser obtida estimando, com mínimos quadrados, $\sigma_j^2 = \hat{k}\,\overline{x}_{j\bullet}^{\hat{q}}$. Logaritmizando obtém-se

$$\ln \sigma_j^2 = \ln \hat{k} + \hat{q} \ln \overline{x}_{j\bullet}.$$

Segue-se que o declive da recta que se ajusta aos pontos $\left(\ln \overline{x}_{j\bullet}, \ln \sigma_j^2\right)$ é uma estimativa de q.

Em muitas situações importantes considerações teóricas sugerem qual o valor apropriado de q.

Por exemplo, se o modelo gaussiano for uma aproximação de um modelo de Poisson, em que valor médio e variância são iguais, $q = 1$, e consequentemente a transformação adequada é a raiz quadrada.

Gilbert (1989, p. 66), com a sua habitual agilidade, anota que o objectivo é estabilizar a variância $\mathrm{var}[X] = \mathbb{E}\left(X - \mu\right)^2$, isto é

encontrar uma função f cuja variância seja independente de μ. Se a função f for suficientemente "regular", então expandindo em série de Taylor e supondo que termos de ordem superior a 2 são neglígiveis

$$f(x) \approx f(\mu) + (x-\mu)f'(\mu),$$

e obtém-se

$$\begin{aligned}\operatorname{var}[f(X)] &\approx \mathbb{E}\left[f(X) - f(\mu)\right]^2 = \mathbb{E}\left(X-\mu\right)^2 \left[f'(\mu)\right]^2 = \\ &= \operatorname{var}[X] \left[f'(\mu)\right]^2 = C^2,\end{aligned}$$

dado que $\mathbb{E}(X-\mu) = 0$ implica $\mathbb{E}[f(X)] \approx f(\mu)$, e consequentemente

$$f'(\mu) = \frac{C}{\sqrt{\operatorname{var}[X]}}\,.$$

No caso de $X \frown \mathit{Poisson}(\mu)$, obtém-se $f'(\mu) = \frac{C}{\sqrt{\mu}}$, e a transformação raiz quadrada, $f(x) = C\sqrt{x}$, como atrás vimos.

Mas com esta abordagem não ficamos limitados a transformações-potência. Por exemplo, se o modelo gaussiano aparecer como aproximação de um modelo binomial, nesse caso $\operatorname{var}(X) = np(1-p)$.

Segue-se que a transformação apropriada é

$$y = \arcsin\sqrt{\frac{x}{n}}\,.$$

Método de Box-Hill para estimação de parâmetros em condições de heterocedasticidade. Determinação das ponderações.

O método proposto por Box e Hill recorre à estimação por mínimos quadrados ponderados, mas para determinar as ponderações é usado o método de estabilização das variâncias.

Para levar a cabo a estimação por mínimos quadrados ponderados, é necessário calcular os pesos a atribuir às observações de cada grupo. Box e Hill consideraram o modelo linear usual para as observações,

$$Y_i = \mu_i + \epsilon_i$$

com erros gaussianos independentes, mas admitindo que possam ter variâncias distintas, parametrizadas na forma $\text{var}(\epsilon_i) = c_i \sigma^2$. Os pesos são então escolhidos como $w_i = 1/c_i$, visto que assim se tem $Var\left(\sqrt{w_i}\ Y_i\right) = \sigma^2$, constante.

Claro que numa situação realista os coeficientes c_i não serão conhecidos, e precisam de ser estimados. Para tal, é admitido que existe uma transformação de Box-Cox,

$$y_i^* = \begin{cases} \dfrac{y_i^{\phi} - 1}{\phi} & \phi \neq 0 \\ \\ \ln y_i & \phi = 0 \end{cases},$$

tal que a variância da transformada é constante, $\text{var}(Y_i^*) = \sigma^2$. Recorrendo ao método de Bartlett para estabilização das variâncias, a variância da variável transformada será

$$\mathrm{var}(Y_i^*) \simeq \mathrm{var}\,(Y_i) \left[\frac{\mathrm{d}y_i^*}{\mathrm{d}y_i}\,(\mu_i) \right]^2 ,$$

donde

$$\mathrm{var}\,(Y_i) \simeq \mathrm{var}(Y_i^*) \left[\frac{\mathrm{d}y_i^*}{\mathrm{d}y_i}(\mu_i) \right]^{-2} = \sigma^2 \mu_i^{2-2\phi}$$

e portanto $c_i \simeq \mu_i^{2-2\phi}$ e $w_i \simeq \mu_i^{2\phi-2}$. Note-se que isto significa que a variância é aproximadamente função do valor médio, uma situação habitual em regressão.

Embora μ_i seja também um parâmetro desconhecido, é o valor médio no grupo i, e portanto pode ser estimado pelo valor previsto, $\hat{y}_i$. A estimativa para os pesos é, assim, $w_i \simeq \hat{y}_i^{2-2\phi}$.

Estimação do índice da transformação

Para determinar ϕ, supõe-se que cada um dos valores médios μ_i são uma função linear, ou aproximadamente linear, de p parâmetros θ_i,

$$\mu_i = x_{i1}\theta_1 + x_{i2}\theta_2 + \cdots + x_{ip}\theta_p .$$

Num contexto de regressão linear os x_{ik} são as variáveis previsoras, numa análise de variância são zeros ou uns (variáveis indicatrizes). Estes, por sua vez, podem já ser transformações de outras variáveis cuja relação com a resposta Y_i não é linear. Os parâmetros são os coeficientes da regressão, ou do modelo de análise de variância.

Aquela igualdade pode ser escrita na forma $\mu_i = \boldsymbol{x}_i\theta$, em que $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ são os valores das variáveis independentes associados a Y_i e $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ o vector dos parâmetros.

Usando a hipótese de independência entre as observações $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$, a função de verosimilhança fixada uma amostra $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ é

$$L(\boldsymbol{y} \mid \theta,\ \phi,\ \sigma) = \frac{\left(\prod_{i=1}^{n} w_i\right)^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\ \sum_{i=1}^{n} w_i\left(y_i - \boldsymbol{x}_i\theta\right)^2\right].$$

Box e Hill admitem em seguida distribuições *a priori* não informativas independentes para os parâmetros. Consideram distribuições localmente uniformes para ϕ e θ, e para σ uma densidade proporcional a $\frac{1}{\sigma}$, como é usual para parâmetros de dispersão, obtendo uma densidade *a posteriori* $p(\theta,\ \phi,\ \sigma|\ \boldsymbol{y})$ proporcional a

$$\frac{\left(\prod_{i=1}^{n} w_i\right)^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}\ \sum_{i=1}^{n} w_i\left(y_i - \boldsymbol{x}_i\theta\right)^2\right].$$

Por integração, obtém-se a densidade *a posteriori* marginal para ϕ:

$$p(\phi \mid \boldsymbol{y}) = c\ |\boldsymbol{X'WX}|^{-\frac{1}{2}} \left(\prod_{i=1}^{n} w_i\right)^{\frac{1}{2}} \exp\left[\sum_{i=1}^{n} w_i\left(y_i - \boldsymbol{x}_i\hat{\theta}\right)^2\right],$$

onde c é uma constante, $\boldsymbol{X}$ é a matriz de $n \times p$ formada pelos elementos x_{ik} (ou seja, pelos vectores-linha $\boldsymbol{x}_i$), $\boldsymbol{W}$ uma matriz

diagonal de $n \times n$ cuja diagonal é constituída pelos pesos w_i, e $\hat{\theta}$ a estimativa de mínimos quadrados ponderados para θ,

$$\hat{\theta} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wy}.$$

Como é habitual em métodos bayesianos, ϕ será estimado pela moda da distribuição, isto é por maximização da sua densidade *a posteriori*. Dada uma estimativa para ϕ, ficam conhecidos os pesos a atribuir às observações y_i na estimação por mínimos quadrados. Draper e Smith (1998, p. 108-117), têm uma secção sobre mínimos quadrados ponderados, onde o leitor interessado pode encontrar uma fundamentação mais detalhada.

Implementação

A circularidade do processo é resolvida por iteração: com base nos valores observados y_i, são calculadas estimativas iniciais dos pesos w_i, que são então usados para determinar uma estimativa inicial para ϕ. A partir desta, obtém-se uma segunda estimativa para os pesos, que são empregues na obtenção de uma segunda estimativa para o índice ϕ, e assim sucessivamente.

O procedimento iterativo para estimar o índice ϕ é constituído pelos seguintes passos:

1. Na primeira iteração, começa-se por calcular valores iniciais de w_i com base nos valores da amostra, y_i, em vez de $\hat{y}_i$, uma vez que ainda não há estimativas dos parâmetros do modelo.

2. Selecciona-se um valor inicial razoável para o índice ϕ.

3. Determina-se a estimativa de mínimos quadrados ponderados para os parâmetros,

$$\hat{\theta} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wy},$$

e com base nela a densidade *a posteriori* $p(\phi \mid \boldsymbol{y})$.

4. Calcula-se $\hat{y}_i$ e uma nova estimativa de w_i, e volta-se ao ponto (2) até encontrar um valor de ϕ que maximize $p(\phi \mid \boldsymbol{y})$.

5. Para esse valor $\hat{\phi}$, determina-se $\hat{y}_i$ com base no modelo, e substituí-se em w_i.

6. Volta-se a (2) e começa-se nova iteração.

7. Regra de paragem: termina quando as iterações convergirem para um valor estabilizado de $\hat{\phi}$.

Box e Hill (1974) dão um exemplo de aplicação em Química, em que se partia de uma variável r, mas a fim de linearizar a relação com outras variáveis importantes se fazia a transformação r^{-1}. Neste caso, porém, o uso da transformação produzia uma heterogeneidade de variâncias bastante marcada, inexistente na variável original.

Em consequência, uma estimação pelo método usual dos mínimos quadrados levava a estimativas negativas para os parâmetros, incompatíveis com o conhecimento existente sobre a variável em estudo.

Nestas circunstâncias, mostrou-se vantajoso usar uma metodologia de mínimos quadrados ponderados. As estimativas para ϕ convergiram ao fim de apenas três iterações, e as estimativas

dos parâmetros do modelo calculadas com base no valor final de ϕ eram positivas.

Evitou-se assim que um modelo válido fosse rejeitado devido à ineficiência do método de estimação.

Note-se ainda que este método é aplicável quer a heterogeneidade de variâncias exista originariamente nos dados quer seja, como no referido exemplo, induzida por transformações efectuadas posteriormente.

Apêndice C

O estimador de Welch-Satterthwaite

Em 2.3.2 apresentámos o estimador de Welch-Satterthwaite de uma forma simples e facilmente generalizável, usada em Pestana e Velosa (2010) desde a 2ª edição desse livro. Registamos abaixo a demonstração mais "pesada" da 1ª edição, que nos parece poder vir a ser evntualmente útil para quem queira aproximar somas de gamas independentes com escalas diferentes, por exemplo para abordar, indo mais longe do que nós, o problema de análise de *spacings*, ou studentização em populações exponenciais, nomeadamente no caso de duas amostras com populações parente com diferentes escalas.

Sejam $Y_k \frown \chi^2_{\nu_k}$ independentes, $k = 1, \ldots, n$, e defina-se

$$W = \sum_{k=1}^{n} a_k Y_k.$$

O nosso objectivo é aproximar W por $\frac{Y_\nu}{\nu}$, onde $Y_\nu \frown \chi^2_\nu$. Por outras palavras, determinar o número de graus de liberdade ν por tal forma que a aproximação $W = \sum_{k=1}^{n} a_k Y_k \simeq \frac{Y_\nu}{\nu}$ seja razoável.

Faz decerto sentido exigir que as variáveis aleatórias $\sum_{k=1}^{n} a_k Y_k$ e $\frac{Y_\nu}{\nu}$ tenham os mesmo valor médio e a mesma variância, e consequentemente, como $\mathbb{E}\left(\frac{Y_\nu}{\nu}\right) = 1$, começamos por impor

$$\mathbb{E}\left[\sum_{k=1}^{n} a_k Y_k\right] = 1.$$

Recorrendo ao teorema de König vem então

$$\mathbb{E}\left[\left(\sum_{k=1}^{n} a_k Y_k\right)^2\right] = \left[\mathbb{E}\left(\sum_{k=1}^{n} a_k Y_k\right)\right]^2 + \operatorname{var}\left[\sum_{k=1}^{n} a_k Y_k\right] =$$

$$= 1 + \frac{\operatorname{var}\left[\sum_{k=1}^{n} a_k Y_k\right]}{\left[\mathbb{E}\left(\sum_{k=1}^{n} a_k Y_k\right)\right]^2}$$

e notando que queremos que

$$\mathbb{E}\left[\left(\sum_{k=1}^{n} a_k Y_k\right)^2\right] \simeq \mathbb{E}\left[\left(\frac{Y_\nu}{\nu}\right)^2\right] = \left[\mathbb{E}\left(\frac{Y_\nu}{\nu}\right)\right]^2 + \operatorname{var}\left(\frac{Y_\nu}{\nu}\right) =$$

$$= 1 + \frac{2\nu}{\nu^2} = 1 + \frac{2}{\nu},$$

vem

$$\nu = \frac{2\left[\mathbb{E}\left(\sum_{k=1}^{n} a_k Y_k\right)\right]^2}{\operatorname{var}\left[\sum_{k=1}^{n} a_k Y_k\right]}.$$

Atendendo agora a que os $Y_k \frown \chi^2_{\nu_k}$ são independentes

$$\operatorname{var}\left[\sum_{k=1}^{n} a_k Y_k\right] = \sum_{k=1}^{n} a_k^2 \operatorname{var}(Y_k) = \sum_{k=1}^{n} a_k^2 2\,\nu_k =$$

$$= 2\sum_{k=1}^{n} a_k^2 \frac{\nu_k^2}{\nu_k} = 2\sum_{k=1}^{n} \frac{a_k^2 \left[\mathbb{E}(Y_k)\right]^2}{\nu_k},$$

obtemos a *aproximação de Satterthwaite* para ν substituindo os valores médios pelas observações:

$$\widetilde{\nu} = \frac{\left(\sum_{k=1}^{n} a_k Y_k\right)^2}{\sum_{k=1}^{n} \frac{a_k^2}{\nu_k} Y_k^2}.$$

Assim, em particular, observando que

$$\frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = a_1 \frac{(n_1 - 1)\,S_1^2}{\sigma_1^2} + a_2 \frac{(n_2 - 1)\,S_2^2}{\sigma_2^2}$$

com

$$a_1 = \frac{\sigma_1^2}{n_1(n_1-1)\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]} \quad \text{e} \quad a_2 = \frac{\sigma_2^2}{n_2(n_2-1)\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]},$$

e

$$\frac{(n_1-1)\,S_1^2}{\sigma_1^2} \frown \chi^2_{n_1-1} \quad \text{e} \quad \frac{(n_2-1)\,S_2^2}{\sigma_2^2} \frown \chi^2_{n_2-1},$$

independentes, conclui-se que

$$\frac{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \simeq \frac{Y_\nu}{\nu}$$

com $Y_\nu \frown \chi^2_\nu$, podendo o número de graus de liberdade ν ser estimado por

$$\widetilde{\nu} = \frac{\left[\frac{\sigma_1^2}{n_1(n_1-1)}\,\frac{S_1^2(n_1-1)}{\sigma_1^2}+\frac{\sigma_2^2}{n_2(n_2-1)}\,\frac{S_2^2(n_2-1)}{\sigma_2^2}\right]^2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}} =$$

$$\frac{\left[\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right]^2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}}.$$

Bibliografia

(As obras precedidas do símbolo • não são directamente referidas no texto; indicam-se por serem importantes para a formação de uma perspectiva histórica sobre studentização e o problema de Behrens-Fisher, e assim, indirectamente, para os fundamentos da Estatística, ou por serem livros que explicitamente referem Welch ou Satterthwaite.)

Aczél, J. (2006). *Lectures on Functional Equations and their Applications*, Dover, New York.

Akahira, M. (2002). Confidence intervals for the difference of means: application to the Behrens-Fisher type problem, *Statistical Papers* **43**, 273–284.

Aspin, A. A. (1948). An examination and further developments of a formula arising in the problem of comparing two mean values. *Biometrika* **35**, 88–97.

Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika* **36**, 290–293.

Babu, G. J., and Padmanabhan, A. R. (2002). Resampling methods for the non-parametric Behrens–Fisher problem, *Sankhyā Ser. A* **64**, 678–692.

Baker, G. A. (1932). Distribution of the means divided by the standard deviations of samples from non-homogeneous populations. *Ann. Math. Statist.* **3**, 1–9.

Baker, J. A. (1993). On a functional equation of Aczél and Chung, *Aequat. Mathemat.* **46**, 99–111.

Barnard, G. A. (1950). Neyman and the Behrens–Fisher problem — an anedocte. *Biometrika* **15**, 67-71.

Barnard, G. A. (1995). On the Fisher-Behrens' test. *Probability and Mathematical Statistics* **37**, 203–207.

Barnett, V. (1999). *Comparative Statistical Inference*, 3rd ed., Wiley, New York.

• Bartlett, M. S. (1935). The effect of non-normality on the t distribution. *Proc. Camb. Philos. Soc.* **31**, 223–231.

Bartlett, M. S. (1936). The information available in small samples. *Proc. Camb. Philos. Soc.* **32**, 560–566.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Royal Soc.* **A 160**, 268-282.

Bartlett, M. S. (1956). Comment on Sir Ronal Fisher's paper: "On a test of significance in Pearson's *Biometrika Tables*". *J. Royal Stat. Assoc.* **B 18**, 295-296.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya* **15**, 377–380.

Behrens, W. V. (1929). Ein Beitrag zur Fehlen-Berechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbüchen* **68**,

807–837.

Belloni, A., and Didier, G. (2008). On the Behrens–Fisher problem: a globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. *Ann. Statist.* **36**, 2377–2408.

• Box, G. E. P. (1954a). Some theorems in quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variances in the one-way classification. *Ann. Math. Stat.* **25**, 290–302.

• Box, G. E. P. (1954b). Some theorems in quadratic forms applied in the study of analysis of variance problems, II: Effect of inequality of variances and of correlation of errors in the two-way classification. *Ann. Math. Stat.* **25**, 484–498.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Royal Statist. Soc.* **B26**, 211–256.

Box, G. E. P., and Hill, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics* **16**, 385–389.

Bozdogan, H., and Ramirez, D. E. (1987). An adjusted likelihood ratio algorithm for the Behrens–Fisher problem. *Statistische Hefte* **28**, 217–231.

• Bradley, R. A. (1952). The distribution of the t and F statistics for a class of non-normal populations. *Virginia J. Sci.* **3**, 1–32.

Brilhante, M. F. (1996a). Inferência sobre o parâmetro de localização de uma população exponencial I. Studentização Externa. In *A Estatística a Decifrar o Mundo*, 47–55, Salamandra, Lisboa.

Brilhante, M. F. (1996b). *Estatísticas Studentizadas com Parente Exponencial*, Dissertação de Mestrado, FCUL, Lisboa.

Brilhante, M. F., and Kotz, S. (2008). Infinite divisibility of the spacings of a Kotz–Kozubowski–Podgórski generalized Laplace model, *Statistics & Probability Letters* **78**, 2433–2436.

Brilhante, M. F., Mendonça, S., Pestana, D., and Rocha, M. L. (2009). Inference on the location parameter of exponential populations. *Discussiones Mathematicae Probability and Statistics* **29**, 115–129.

Brilhante, M. F., Pestana, D. D., e Rocha, J. (1996). Inferência sobre o parâmetro de localização de uma população exponencial II. Studentização Interna. In *A Estatística a Decifrar o Mundo*, 57–63, Salamandra, Lisboa.

Brilhante, M. F., Pestana, D. D., Rocha, J. e Velosa, S. (2001). *Inferência Estatística sobre Localização e Escala*, 166p., Sociedade Portuguesa de Estatística, Ponta Delgada.

Buot, M.-L. G., Hoşten, S., and Richards, D. St. P. (2007). Counting and locating the solutions of the polynomial systems of maximum likelihood equations, II: the Behrens–Fisher problem. *Statistica Sinica.* **17**, 1343–13541.

Casella, G., and Berger, R. L. (2002). *Statistical Inference.* Duxbury, Pacific Grove. (p. 314–315, e Ex. 8.42, p. 419).

• Chung, K. L. (1946). The approximate distribution of Student's t statistic. *Ann. Math. Statist.* **17**, 447–465.

Cobb, G. B. (1998). *Design and Analysis of Experiments.* Springer, New York. (p. 596–598).

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covari-

ance. *Math. Proc. Cambridge Philos. Soc.* **30**, 178–191.

Cochran, W. G (1951). Testing a linear relation among variances, *Biometrics* **7**, 17-32

Cochran, W. G (1964). Approximate significance levels of the Behrens-Fisher test. *Biometrics* **20**, 191–195.

Cochran, W. G., and Cox, G. M. (1957). *Experimental Designs*. Wiley, New York. (p. 100–102 e 567).

• Craig, A. T. (1932). The simultaneous distribution of the mean and standard deviation in small samples. *Ann. Math. Statist.* **3**, 126–140.

Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *C. R. Acad. Sci. Paris* **200**, 1265–1266.

Davenport, J. M., and Webster, J. T. (1975). The Behrens–Fisher problem, an old solution revisited. *Metrika* **22**, 47–54.

David, H. A. (1981). *Order Statistics*. Wiley, New York.

Dean, A., and Voss, D. (1999). *Design and Analysis of Experiments*. Springer, New York. (p. 117, 151, 165).

• DiSantostefano, R. L., and Muller, K. E. (1995). A comparison of power approximations for Satterthwaite's test, *Commun. Statist. — Simula.* **24**, 583–593.

• Dixon, W. J., and Masey Jr., F. J. (1969). *Introduction to Statistical Analysis* McGraw-Hill, New York (p. 119).

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis*. John Wiley and Sons, New York.

Drton, M. (2008). Multiple solutions to the likelihood equations in the Behrens-Fisher problem. *Statistics and Probability Letters.* **33**, 3288–32937.

Drton, M., and Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinav. J. Statistics* **33**, 247–257.

Dudewicz, E. J., and Ahmed, S. U. (1998). New exact and asymptotically optimal solution to the Behrens-Fisher problem, with tables, *Amer. J. Math. Manag. Sci.* **18**, 359–426.

Dudewicz, E. J., and Ahmed, S. U. (1999). New exact and asymptotically optimal heteroscedastic statistical procedures and tables, II, *Amer. J. Math. Manag. Sci.* **19**, 157–180.

Dudewicz, E. J., Ma, Y., Mai, S. E., and Su, H. (2007). Exact solutions to the Behrens-Fisher problem: Asymptotically optimal and finite sample efficient choice among, *J. Statist. Planning and Inference* **137**, 1584–1605.

Dunnet, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *J. Amer. Statist. Assoc.* **75**, 796–800.

Efron, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64**, 1278–1302.

Efron, B. (1998). R. A. Fisher in the 21st century, *Statistical Science* **13**, 95–122.

Fenstad, G. U. (1983). A comparison between the U and V tests in the Behrens-Fisher problem. *Biometrika* **70**, 300–302.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

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

Fisher, R. A. (1935b). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

Fisher, R. A. (1937). On a point raised by M. S. Bartlett on fiducial probability. *Ann. Eugen.* **7**, 370–375.

Fisher, R. A. (1939a). Student. *Ann. Eugen.* **9**, 1–9.

Fisher, R. A. (1939b). The comparison of samples with possibly unequal variances. *Ann. Eugen.* **9**, 174–180.

Fisher, R. A. (1941). The asymptotic approach to Behrens' integral with further tables for the d test of significance. *Ann. Eugen.* **11**, 182–187.

Fisher, R. A. (1956a). *Statistical Methods and Scientific Inference* Oliver and Boyd, Edinburgh.

Fisher, R. A. (1956b). On a test of significance in Pearson's *Biometrika Tables J. Royal Statist. Soc.* **B 18**, 56–60.

Fisher, R. A. (1990). *Statistical Methods, Experimental Design and Scientific Inference*. Oxford University Press, Oxford.

Fisher, R. A., and Healy, M. J. R. (1956). New tables of Behrens' test of significance *J. Royal Statist. Soc.* **B 18**, 212–216.

Fisher, R. A., and Yates, F. (1938). *Statistical Tables*. Oliver and Boyd, Edinburgh.

Gans, D. J. (1991). Preliminary tests on variances. *Amer. Statist.* **45**, 258.

• Gardiner, W. P., and Gettinby, G. (1998). *Experimental Design Techniques in Statistical Practice*. Horwood Publ., Chichester. (p. 25).

Gauss, F. (1809). *Theoria Motus Corporum Celestium* (trad. *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*, Dover, 1963.)

• Gayen, A. K. (1949). The distribution of 'Student's' t in random samples of any size drawn from non-normal universes. *Biometrika* **36**, 353–369.

Gaylor, D. W., and Hopper, F. N. (1969). Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. *Technometrics* **11**, 699–706.

• Geary, R. C. (1936). The distribution of 'Student's' ratio for non-normal samples. *Suppl. J. Royal Statist. Soc.* **3** 178–184.

Gilbert, N. (1989). *Biometrical Interpretation — Making Sense of Statistics in Biology*, 2nd ed., Oxford Univ. Press, Oxford.

Gomes, H., and Pestana, D. (1982). Studentized statistics. In P. Thoft-Christensen (ed.), *Reliability Theory and its Applications in Structural and Soil Mechanics*, 605–612, Nijhof, The Hague.

Gomes, M. I. and Pestana, D. (1978). The use of fractional calculus in Probability Theory. *Portugaliae Mathematica* **37**, 259–271.

Goodall, C. (1981). M-estimators of location. In D. Hoaglin, F. Mosteller and J. Tukey (eds). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.

Hampel, F. (2006). The proper fiducial argument. In *General Theory of Information Transfer and Combinatorics*, R. Ahlswede (ed.), Springer, Berlin, 512–526.

Helmert, F. R. (1875). Über der Berechnung der wahrscheinlichen Fehlers einer endlichen Anzahl wahrer Beobachtungsfehler. *Z. Mathematik Physik* **20**, 300–303.

Hirschmann, I. I., and Widder, D. V. (1955). *The Convolu-*

tion Transform, Princeton University Press, Princeton.

Hoaglin, D., Mosteller, F., and Tukey, J. (1981). *Understanding Exploratory Data Analysis and Robust Statistics*, Wiley, New York. (trad.:, *Análise Exploratória de Dados, Estatística Robusta — um Guia*, Salamandra, Lisboa.).

• Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat.* **2**, 360–378.

• Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. Math. Stat.* **11**, 271–283.

Hotelling, H. (1961). The behavior of some standard statistical tests under nonstandard conditions. *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability, I*, 319–359.

Hoyle, M. H. (1973). Transformations — an introduction and bibliography. *Internat. Statist. Rev.* **41**, 203–223.

Iglewicz, B. (1981). Robust scale estimators and confidence intervals for location. In D. Hoaglin, F. Mosteller and J. Tukey (eds). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.

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

Jeffreys, H. (1983). *Theory of Probability, 3rd ed., corrected printing.* Oxford University Press, Oxford.

Johnson, N. L. (1978). Modified t test and confidence intervals for asymmetrical populations, J. Amer. Statist. Assoc. **73**, 536–544.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2, 2nd ed., Wiley, New

York.

Kendall, M. G., and Stuart, A. (1961). *The Advanced Theory of Statistics, II: Inference and Relationship*, Griffin, London. (p. 139–154). (Veja-se Stuart *et al.* (2009) e Stuart and Ord (2009), edição actualizada deste clássico.)

Kim, S.-H., and Cohen, A. S. (1998). On the Behrens-Fisher Problem: A Review, *J. Educational and Behavioral Statistics* **23**, 356–377.

Koopman, B (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society* **39**, 399–409.

• Laderman, J. (1939). The distribution of Student's ratio for samples of two items drawn from non-normal universes. *Ann. Math. Stat.* **10**, 376–379.

Lee, A. F., and Fienberg N. S. (1991). A fitted test for the Behrens-Fisher problem. *Communic. Statist. — Theory Meth.* **20**, 653–666.

• Lee, A. F., and Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *J. Amer. Statist. Assoc.* **70**, 933–941.

Levi-Civita, T. (1913). Sulle funzioni che ammettono una formula d'addizione del tipo $f(x+y) = \sum_{i=1}^{n} X_i(x)\, Y_i(y)$. Atti Accad. Naz. Lincei, Rendic. **22**, 181–183.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem *J. Royal Statist. Soc.* **B 20**, 102–107.

Linnik, Yu. V. (1967). On the elimination of nuisance parameters in statistical problems. *Proc. Fifth Berkeley Sympos.*

Math. Statist. and Probability, Vol. I: Statistics, 267–280. Univ. California Press, Berkeley, CA.

Logan, B. F., Mallows, C. L., Rice, S. O., and Shepp, L. A. (1973). Limit distributions of self-normalized sums. *Ann. Probability* **1**, 788–809.

Margolin, B. (1977). The distribution of internally studentized statistics via Laplace transform inversion. *Biometrika* **64**, 573–582.

Markowski, C. A., and Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *Amer. Statist.* **44**, 322–326.

• Martins, J. (2005). Assimetria e Achatamento: O Comportamento da Estatística de Student em Populações não Gaussianas, Dissertação de Mestrado, Universidade de Lisboa.

• Martins, J. P. (2009). *Feira dos Momentos — Planeamento Experimental e Investigação de Localização e Escala em Populações não Gaussianas*, Universidade de Lisboa.

• Martins, J. P., e Mendonça, S. (2008). Transformações estabilizadoras da variância, *Actas do XV Congresso Anual da Sociedade Portuguesa de Estatística*, Edições SPE, Lisboa, 343-350.

McBean, E. A., Kompter, M., and Rovers, F. (1988). A Critical Examination of Approximations Implicit in Cochran's Procedure, *Ground Water Monitoring & Remediation* **8**, 83–87.

Mehendale, K., and Kalluraya, M. (2011). Behrens-Fisher Problem and Distribution Free Tests for Equality of Means, Young Statisticians Satellite Meeting, 58th Session of the ISI, Dublin.

• Mehta, C. R., and Srinivasan, R. (1970). On the Behrens-Fisher problem. *Biometrics* **57**, 649–655.

Mendonça, S. (1997). *Fugindo à Gaussiana: Modelos Lineares Generalizados e Estimação Robusta de Localização e Escala.* DEIO, FCUL, Lisboa.

• Montgomery, D. C. (1997). *Design and Analysis of Experiments.* Wiley, New York (p. 486–491).

Moser, B. K., and Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *Amer. Statist.* **46**, 19–21.

Moser, B. K., Stevens, G. R., and Watts, C. L. (1989). The two-samples t-test versus Satterthwaite's approximate F test. *Communic. Statist. — Theory and Meth.* **18**, 3963–3975.

Mosteller, F., e Rourke, R. E. K. (1993). *Estatísticas Firmes.* Salamandra, Lisboa. (trad. *Sturdy Statistics*, Addison-Wesley, 1973).

Mosteller, F., and Tukey, J. (1977). *Data Analysis and Regression: a Second Course in Statistics*, Addison–Wesley, Reading, Mass.

Murteira, B. J. F. (1988). *Estatística: Inferência e Decisão.* Imprensa Nacional/Casa da Moeda, Lisboa.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*, McGraw-Hill, New York.

Neubert, K., and Brunner, E. (2007). A studentized permutation test for the non-parametric Behrens-Fisher problem, *Computational Statistics & Data Analysis* **51**, 5192–5204.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *J. Royal Statist. Assoc.* **B 18**, 288–294.

Oehlert, G. W. (2000). *A First Course in Design and Analysis of Experiments.* Freeman, New York. (p. 132, 262, 274).

Pearson, K. (1905). On the general theory of skew correlation and nonlinear regression, *Drapers' Company Research Memoirs*, Biometric Ser. II

Perlo, V. (1933). On the distribution of 'Student's' ratio for samples of three drawn from the rectangular distribution. *Biometrika* **25**, 203–204.

Pestana, D. (1978). *Unimodality, Infinite Divisibility, and Related Topics.* Sheffield.

Pestana, D. (1981). O argumento fiducial. *Actas II Colóquio Estatística e Investigação Operacional*, SEIO, Fundão/Covilhã, 189-215.

Pestana, D., Brilhante, F., and Rocha, J. (1999). The analysis of variance revisited. In *Extreme Values and Additive Laws*, 73–77, Lisboa.

Pestana, D, e Mendonça, S. (2007). Normal?, in Estatística, Ciência Interdisciplinar, SPE, 43–58.

Pestana, D., e Rocha, J. (1992). Distribuições amostrais em situações não Gaussianas. In D. Pestana (ed.). *Estatística Robusta, Extremos e Mais Alguns Temas*, 73–80, Salamandra, Lisboa.

Pestana, D., e Rocha, J. (1993). Análise de escala — modelo exponencial. *A Estatística e o Futuro e o Futuro da Estatística*, 295–303. Salamandra, Lisboa.

Pestana, D. D., and Rocha, J. (1995a). Internally studentized

statistics with exponential parent. *Notas e Comunicações do CEAUL*

Pestana, D. D., e Rocha, J. (1995b). Studentização interna com distribuição parente simétrica. In *Bom Senso e Sensibilidade, Traves Mestras da Estatística*, 377–381, Salamandra, Lisboa.

Pestana, D., e Velosa, S. (2010). *Introdução à Probabilidade e à Estatística*. 4ª ed. Fundação Gulbenkian, Lisboa).

Pexider, J. V. (1902). Notiz über Funkcionaltheoreme. *Monatsh. Math. Phys.* **14**, 293–301.

Pitman, E. (1936). Sufficient statistics and intrinsic accuracy. *Math. Proc. Cambridge Philos. Soc.* **32**, 567–579.

Pitman, P. J. G. (1957). Statistics and Science, *J. Amer. Statist. Assoc.* **52**, 322–330.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

• Rahman, M., and Ehsanes Saleh, A. K. Md. (1974). Explicit Form of the Distribution of the Behrens–Fisher d-Statistic, *J. Royal Statist. Soc.* **B 36**, 54–60.

• Ratcliffe, J. F. (1968). The effect on the t-distribution of non-normality in the sampled population. *Applied Statistics* **17**, 71–73.

Rider, P. R. (1931). On small samples from certain non-normal universes. *Ann. Math. Statist.* **2**, 48–65.

Rietz, H. L. (1939). On the distribution of 'Student's' ratio for small samples from certain non-normal populations. *Ann. Math. Statist.* **10**, 265–274.

Rocha, J. (1995). *Localização e Escala em Situações não Clássicas.* Dissertação de Doutoramento, Universidade dos Açores, Ponta Delgada.

Rocha, J., e Martins, J. (2005). Estatística de Student com Parente não Gaussiana, *Actas do XII Congresso Anual da Sociedade Portuguesa de Estatística*, Edições SPE, Lisboa, 651–656.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*, Wiley, New York.

Sawilowsky, S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: the probable difference between two means when $\sigma_1 \neq \sigma_2$, *J. Modern Appl. Statist. Methods* **1**, 461–472.

Samuels, M. L., and Witmer, J. A. (1999). *Statistics for the Life Sciences.* Prentice Hall, Upper Saddle River. (p. 233).

Satterthwaite, F. E. (1941). Synthesis of variance, *Psychometrika* **6**, 309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110–114.

Scheffé, H. (1943). On solutions of the Behrens-Fisher problem based on the $t-$distribution. *Ann. Math. Stat.* **14**, 35–44.

Scheffé, H. (1944). A note on the Behrens-Fisher problem. *Ann. Math. Stat.* **15**, 430–432.

Scheffé, H. (1970). Practical solutions for the Behrens-Fisher problem. *J. Amer. Statist. Assoc.* **65**, 1501–1508.

• Scheffé, H. (1999). *The Analysis of Variance.* Wiley, New York. (p. 247–248).

• Siddiqui, M. M. (1964). Distribution of Student's t in sam-

ples from a rectangular universe. *Rev. Int. Statist. Inst.* **32**, 242–250.

Smith, H. (1936). The problem of comparing the results of two experiments with unequal means. *J. Council Sci. Industr. Res.* **9**, 211–212.

Sprent, P., and Smeeton, N. C. (2001). *Applied Nonparametric Statistical Methods*, 3rd ed., Chapamn and Hall/CRC, Boca Raton.

Stäckel, P. (1913). Sulla equazione funzionale $f(x+y) = \sum_{i=1}^{n} X_i(x)\, Y_i(y)$. Atti Accad. Naz. Lincei, Rendic. **22**, 392–393.

• Starkey, D. M. (1937). A test of the significance of the difference between means of samples from two normal populations without assuming equal variances. *Ann. Math. Stat.* **9**, 201–213.

Steutel, F. W. (1967). Random division of an interval. Statistics Neerland. **21**, 231–244.

Stuart, A., and Ord, K., and Arnold, S. (2009). Kendall's Advanced Theory of Statistics: Volume 1— Distribution Theory, 6th ed., Wiley, New York.

Stuart, A., Ord, K. (2009). Kendall's Advanced Theory of Statistics: Volume 2A — Classical Inference and and the Linear Model, 6th ed., Wiley, New York.

Student (1908). The probable error of a mean. *Biometrika* **6**, 1–25.

Sugiura, N., and Gupta, A. K. (1987). Maximum likelihood estimates for the Behrens-Fisher problem. *J. Japan Statist. Soc.* **17**, 55–60.

Sukhatme, P. V. (1938). On Fisher and Behrens' test of

significance for the difference in means of two normal samples. *Sankhya* **4**, 39–48.

Sundberg, R. (2010). Flat and multimodal likelihoods and model lack of fit in curved exponential families. *Scandinav. J. Statistics.* **37**, 632–643.

• Tanburn, E. (1937). Note on an approximation used by B. L. Welch. *Biometrika* **29**, 361–362.

• Trickett, W. H., Welch, B. L., and James, G. S. (1956). Further critical values for the two-means problem. *Biometrika* **43**, 203–205.

Tsui, K.-W., and Tang, S. (2005). Distributional property of the generalized p-value for the Behrens-Fisher problem with applications to multiple testing, *Statistics and Probability Letters* **77**, 1–8.

Tukey, J. W. (1957). On the comparative anatomy of transformations. *Ann. Math. Stat.* **28**, 602–632.

Tukey, J. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.

Velosa, S. (2001). Modificação do teste de Scheffé para comparação de valores médios de populações gaussianas com variâncias diferentes. *IX Congresso Anual da SPE*, Ponta Delgada.

Velosa, S. (2003). *O Problema de Behrens-Fisher*, Livraria Escolar Editora, Lisboa.

• Wald, A. (1955). Testing the difference between the means of two normal populations with unknown standard deviations. *Selected Papers in Statistics and Probability by Abraham Wald*, McGraw-Hill, New York.

• Wang, Y. Y. (1967). A comparison of several variance estimators. *Biometrika* **54** 301–305.

• Wang, Y. Y. (1971). Probabilities of Type I errors of the Welch test for the Behrens-Fisher problem. *J. Amer. Statist. Assoc.* **66** 65–76.

Welch, B. L. (1936). Specification of rules for rejecting too variable a product, with particular reference to an electrical lamp problem. *J. Roy. Statist. Soc. Suppl.* **3**, 29–48.

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

Welch, B L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *Ann. Math. Stat.* **10**, 58–69

Welch, B. L. (1947a). The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34**, 28–35.

Welch, B. L. (1947b). On the studentization of several variances. *Ann. Math. Statist.* **18**, 118–122.

Welch, B. L. (1949). Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. *Biometrika* **36**, 293–296.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika* **38**, 330–336.

Welch, B. L. (1956a). On linear combinations of several variances. *J. Amer. Statist. Assoc.* **51**, 132–148.

Welch, B. L. (1956b). Note on some criticism made by Sir Ronald Fisher. *J. Royal Stat. Assoc.* **B 18**, 297–302.

• Welch, B. L. (1958). Student' and small sample theory, *J. Amer. Statist. Assoc.* **53**, 777–788.

Wilkinson, G. N. (1977). On resolving the controversy in statistical inference (with discussin), J. Roy. Statist. Soc. **B 39**, 119–171.

Zabell, S. L (1992) R. A. Fisher and the fiducial argument. *Statistical Science* **7**, 369–387.

Zar, J. H. (1996). *Biostatistical Analysis*, Prentice Hall, Upper Saddle River. (p. 129).

Índice Remissivo

Índice de Autores

ISBN: xxx-xxx-xxxx-xx-x
Depósito Legal: xxxxxx/11