UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

META-ANÁLISE

Harmonização de Testes Usando os Valores de Prova

Fernando José Araújo Correia da Ponte Sequeira

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

(Especialidade de Probabilidades e Estatística)

2009

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

META-ANÁLISE

Harmonização de Testes Usando os Valores de Prova

Fernando José Araújo Correia da Ponte Sequeira

Tese orientada pelo Professor Doutor Dinis Duarte Pestana
e pelo Professor Doutor Sílvio Filipe Velosa

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

(Especialidade de Probabilidades e Estatística)

2009

Em memória de Orlando Oliveira e de José Carlos Rocha

Agradecimentos

◇ Os meus primeiros agradecimentos são para os meus orientadores, o Professor Doutor Dinis Pestana, a quem coube a parte de leão na invenção de temas e desafios que me colocou, e mais recentemente o Professor Sílvio Velosa, da Universidade da Madeira, que aceitou colaborar nessa orientação nos temas que investiguei nos últimos meses. Uma palavra também sobre a Professora Doutora Teresa Alpuim, que numa fase inicial da minha carreira contribuiu para despertar o meu gosto pela Matemática e pelo rigor necessário no seu tratamento.

◇ Aos meus pais, que tanto sofreram para que eu terminasse esta tese, que eu maldosamente adiava, e à minha filha, fiel depositária do meu património genético, minha Polaris desencontrada.

◇ A todos os meus colegas que colaboraram em artigos comigo: a Ivette Gomes, a Rita Vasconcelos, a Sandra Mendonça, a Sandra Aleixo, o Sílvio Velosa, a Madalena Malva, o José Rocha, a Fernanda Diamantino, e o Luís Soares de Almeida.

◇ A todos os meus colegas do Departamento (em especial à Lisete, minha ex-colega de gabinete, pela sua permanente ajuda) e todos os meus amigos que sempre me incentivaram neste trabalho e apoiaram em situações em surgia a vontade de desistir.

◇ O CEUAL — Centro de Estatística e Investigação Operacional da Faculdade de Ciências, e o DEIO, Departamento de Estatística e Investigação Operacional da FCUL, foram duas instituições que me apoiaram, e a quem expresso

os meus agradecimentos.

◇ A B, M e S, porque me acompanharam no dia a dia.

A todos o meu,

Muito Obrigado!

Sumário

Apresenta-se uma breve resenha de tópicos de Meta-Análise que mostram a importância dessa subdisciplina da Estatística na construção do conhecimento científico, viabilizando sínteses dos factos e conclusões conhecidas, e explora-se com algum detalhe o problema das sínteses usando níveis de significância descritivos.

Apresentamos uma técnica artificiosa de calcular pseudo$-p$'s, ampliando computacionalmente a amostra, e estudamos as implicações desses procedimentos na potência dos testes meta-analíticos usuais, de Fisher e de Tippett; usando os métodos de Stouffer, averiguamos o impacto do uso da amostra ampliada no número de estudos não significativos que seriam necessários para virar a decisão de rejeição global, uma das formas mais usadas de avaliar o efeito do enviesamento na publicação nas revisões sistemáticas e sínteses meta-analíticas.

Estabelecem-se ainda resultados sobre caracterização de uniformes, e estuda-se a distribuição exacta de funções de betas potencialmente úteis em eventuais extensões usando espaçamentos.

Palavras-chave: Meta-análise, combinação de $p-values$, $p-values$ aleatórios e generalizados, geração de uniformes, enviesamento na publicação.

Abstract

We present a brief overview of the main topics in Meta-Analysis, enhancing its relevance towards scientific knowledge building, via syntheses of facts and conclusions of previous work. The main original contributions are in the special field of meta-analytical syntheses based on reported $p-$values.

We develop techniques to increase the sample size with artificial *pseudo* $p-values$, and we assess the impact of those computationally augmented samples of $p-$values in the power of inferential procedures used to meta-analyse independent tests, namely the routinely used Tippett's and Fisher's methods. On the other hand, using Stouffer's method based on normal scores, we evaluate the number of non-significant results that would be needed to alter a global rejection decision, the most widely used technique of assessing the effect of publication bias.

We also present results on characterizations of uniform populations, and we compute closed form expressions to some exact distristibutions of functions of beta variables, extending our results with uniforms, which are potentially useful for further developments using order statistics and spacings.

Key-words: Meta-Analysis, combined $p-$values, random and generalized $p-$values, generation of uniform random numbers, publication bias.

Conteúdo

Capítulo 1

Introdução

A arte de combinar informação, em si mesma aparentemente inconclusiva, por forma a obter resultados significativos, ou de harmonizar resultados contraditórios, nasceu com a Estatística Matemática, nas mãos de K. Pearson[(1)], e não deixou de ser referida por Fisher. Devem-se a Cochran e a Mantzel e Haenszel os resultados pioneiros, e as recensões críticas de artigos publicados não podiam, na década de 70, continuar a ter um carácter subjectivo e impressionista. Glass cunhou em 1976 o termo Meta-Análise, formalizando uma investigação que ultrapassava as revisões sistemáticas da literatura publicada, e em 2000 Senn "arrumou" as sínteses meta-analíticas em 3 tipos:

> "*One can identify three basic types of meta-analysis depending on the nature of the data. For Type A, the outcome measure is the same in all trials being analyzed, the analyst has access to all original data and chooses to base the analysis on these data. For Type B, the outcome measure is the same in all trials being analized but the analyst uses*

(1) A discussão de sub-tabulação em Pestana e Velosa (2008) evidencia que basta **uma** parcela da estatística de qui-quadrado exceder o quantil crítico, ao nível de significância estabelecido, da distribuição de qui-quadrado com um grau de liberdade para se rejeitar a hipótese nula que está a ser avaliada; a estatítica de teste é a soma das parcelas porque por si só cada uma delas pode não apontar rejeição, mas a acumulação da evidência de todas elas conseguir esse resultado. Isto é já, na sua essência, o âmago das metodologias meta-analíticas.

summaries from each trial as the basis of analysis, either as a matter of analytic choice or because the original data are not available. For Type C, different outcomes have been measured in different trials and analysis has to proceed using 'unit-free' summaries.

Type C is what Glass originally considered in his famous [1976] *paper, being confronted with the task of summarizing the results of many different independently conducted studies in which research workers had been using different measuring instruments, despite the fact that the overall purpose of such studies* [...] *was the same.*"

Nesta tese tratamos sobretudo da harmonização do resultado de testes independentes; assim, apesar de o contexto usual ser meta-análise do Tipo B, muitos dos resultados são também aplicáveis a meta-análise do Tipo C.

Nas áreas de Medicina, de Farmacologia, e de Ciências Humanas, é quase um standard iniciar qualquer estudo de fôlego com uma recensão meta-analítica (ou referência a uma recente), e apesar de os detractores afirmarem que os adeptos da meta-análise são os crédulos que julgam que pilhas de estrume destilam um perfume de rosas, é difícil contornar as vantagens do uso sem abusos da meta-análise.

Abordamos alguns aspectos da problemática da Meta-Análise, com especial relevo para a harmonização de resultados significativos com resultados inconclusivos, usando níveis de significância descritivos ($p-values$) ou magnitudes de efeitos reportados; o incontornável problema do enviesamento na publicação devido às políticas editoriais também será investigado.

Abordamos ainda a questão relevante de a Meta-Análise servir, também, para chamar a atenção para estudos que deveriam ter sido feitos e estão em falta, e consequente relevância desta disciplina no planeamento da investigação.

Os estudos de Soares de Almeida *et al* (2008) serão usados parcialmente na exemplificação.

O Capítulo 2, de índole introdutória, baseia-se largamente na síntese preparada por Pestana, Rocha, Sequeira, e Vasconcelos (2006) a convite dos organizadores do

EQS 2005.

O Capítulo 3 apresenta resultados sobre a combinação de evidência estatística usando os níveis de significância descritiva de testes independentes. Além de resultados sobre $\min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right)$ e $X+Y-\mathcal{I}[X+Y]$, onde X e Y são variáveis aleatórias independentes com suporte em $(0,1)$, e nomeadamente quando uma delas tem distribuição uniforme (ou mais geralmente beta), interessantes em si mesmos numa perspectiva de caracterização, investiga-se a possibilidade de aumentar artificialmente a amostra de níveis de significância descritiva — que sob validade de H_0 se sabe terem distribuição uniforme — com pseudo–níveis de significância descritiva, a fim de investigar se tal permite inferência com potência melhorada. Estes resultados foram já aceites para publicação internacional indexada.

Se no Capítulo 3 nos colocamos na perspectiva de H_0 ser verdadeira, que evidentemente é pouco credível se houver uma acumulação de níveis de significância descritiva que abonem a favor da rejeição dessa hipótese, não deixamos por outro lado de aflorar o problema da *evidência* a favor da alternativa. O uso de níveis de significância descritiva como resumo da evidência estatística tem sido abundantemente criticado; por isso, abordamos a sugestão de Kulinskaya *et al.* (2008) de quantificar a evidência a favor da alternativa, usando uma função-chave apropriada para normalizar os dados, no caso de se pretender usar a normal como medida de todas as coisas.

No Capítulo 4 discutimos também os conceitos de níveis de significância descritivos aleatórios e níveis de significância descritivos generalizados ("*generalized P–values*") de Weerahandi e co-autores (Tsui and Weerahandi, 1989; Gamage and Weerahandi, 1998; Weerahandi, 2004), que se presta a sintetizar resultados dos diversos estudos que se pretende harmonizar.

A questão do viés de publicação é abordada, na perspectiva específica de estimar quantos trabalhos não publicados seriam necessários para pôr em causa os resultados de uma síntese meta-analítica em que o "efeito de gaveta" possa eventualmente ter alguma relevância. Investiga-se, naturalmente, o efeito de usar pseudo$-p$s nesse número, e consequentemente o que tal traz de benéfico aos estudos de meta-análise.

A Meta-Análise tem vindo a afirmar-se a nível internacional, e também entre nós se nota um interesse crescente nas áreas de Farmácia (académica e industrial), Medicina, Educação, Sociologia e outras. Por outro lado, tem aumentado o número de alunos pós-graduados que mostram alguma avidez por conhecer esta área. A bibliografia que referimos, estando muito longe de ser exaustiva, aborda no entanto uma grande diversidade de problemas; de facto não nos ocorre nenhum problema encontrado nas referências que consultámos que não seja referido na selecção bibliográfica que decidimos citar. As apreciações que no texto fazemos de livros, capítulos de livros, e trabalhos que consideramos reflexões críticas incontornáveis, ajudam o leitor a fazer uma primeira selecção.

Capítulo 2

Evidência Estatística e Meta-Análise

> "*Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature.*"
>
> (Glass, 1976)

2.1 A adição de Informação em Tabelas de Contingência

Numa entrevista feita a Linus Pauling, um dos raríssimos recipientes de dois prémios Nobel, perguntaram-lhe o que era necessário fazer para ser laureado com

o Nobel. Após uma breve hesitação respondeu: "Ter muitas ideias, e a coragem de deitar fora quase todas."

Admirável resposta, ilustração magnífica de duas ideias fundamentais da Probabilidade e da Estatística, respectivamente:

- Da Lei dos Grandes Números de Bernoulli decorre que se se tiver muitas ideias, mesmo que a probabilidade de cada uma delas ser profícua seja muito baixa, o número esperado de ideias com elevado potencial criativo pode ser promissor.

- A metodologia da investigação científica, de que um antepassado notável foi Guillaume d'Occam, recomenda o exercício de um espírito crítico implacável: nada deve ser admitido em Ciência sem antes ter passado pelo crivo de ser considerado irrelevante (rasoura de Occam) — e só no caso de se demonstrar que este pressuposto de irrelevância deve ser rejeitado é que a nova teoria é admitida.

Esta metodologia enformou o desenvolvimento da teoria dos testes de hipóteses, uma das áreas da Estatística mais frequentemente usadas em aplicações, nomeadamente na investigação científica. A tomada de decisões resulta da perspectiva do decisor sobre o valor de prova (*p-value*) $p = \mathbb{P}\left[T \geq T(obs.) \mid H_0 verdadeira\right]$, onde T é a estatística de teste. Por isso, é necessário conhecer a distribuição amostral de $T \mid H_0$ — ou pelo menos os seus quantis críticos —, por outras palavras H_0 tem que ser uma hipótese que permita um condicionamento fácil, em geral uma hipótese simples. Por isso, a hipótese de trabalho que motiva a investigação científica (que em geral é uma hipótese composta, porventura vaga, por exemplo que o efeito médio do tratamento experimental é *maior* do que o do tratamento de controlo) é muitas vezes relegada para hipótese alternativa H_A, optando-se por uma hipótese nula simples que é a sua negação.

Se a decisão de rejeitar a hipótese nula resultar da evidência factual $\boldsymbol{x}$ recolhida, com uma probabilidade p diminuta de rejeitar uma hipótese nula verdadeira, esta negação da negação é um progresso notável, ainda que não seja uma afirmação (uma

vez que a segunda negação é feita incorrendo um risco p de ser uma decisão errada): ficamos com uma convicção objectivamente alicerçada de que a hipótese alternativa H_A que nos motiva é promissora, e com uma quantificação do risco de se estar a rejeitar uma hipótese nula verdadeira. Por isso se diz que rejeitar a hipótese nula é tomar uma decisão forte, manter a hipótese nula é uma decisão fraca, *decorrente porventura de evidência factual insuficiente.*

Daí *a tentação de ampliar a evidência factual*, que motivou o desenvolvimento da Meta-análise.

De facto, só em situações muito excepcionais são disponibilizadas verbas para caríssimos ensaios clínicos de grande dimensão, e a alternativa é tentar sintetizar (meta-analisar) estudos diversos, mesmo que realizados com protocolos e propósitos diferentes, e havendo ainda o perigo de a comparação com um factor comum induzir dependências espúrias.

A meta-análise é o complexo de metodologias que têm como objectivo sintetizar as conclusões de um conjunto de estudos diferentes, a partir de uma integração quantitativa sistemática dos correspondentes relatórios. Ao combinar diversos resultados, é expectável que se consiga potência acrescida para detectar efeitos dos tratamentos, e maior precisão na estimação desses efeitos. A meta-análise pode incidir sobre artigos publicados analisando ensaios clínicos, ou sobre sumários disponibilizados pelos cientistas que realizaram ensaios clínicos; presentemente, favorece-se em detrimento dessas fontes o que se designa por *IPD — individual patient data*, porque permite verificação dos dados originais, actualização dos dados e *follow-up* (Oxman *et al.*, 1995).

Para substanciar a que ponto a escassez de dados pode ser responsável por os estudos serem inconclusivos, e como a globalização de vários estudos pode reverter essa situação, detemo-nos num exemplo clássico de síntese de evidência construída a partir de diversas tabelas de contingência, em que, tomadas isoladamente, não se detecta significância ao nível usual $\alpha = 0.05$.

A combinação de informação de várias tabelas de contingência foi uma das primeiras áreas da meta-análise a desenvolver-se, muito antes de a meta-análise ser

reconhecida como subdisciplina importante da Estatística, ou mesmo de ter ganho um nome próprio. No que respeita a combinação de informação de várias tabelas 2×2,

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

versando a mesma classificação duplamente dicotómica, o método da "*raiz do qui--quadrado*" é um dos mais simples:

- Para cada tabela,

$$X^2_{2,2}(obs.) = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

é (aproximadamente, é óbvio, considerando-se que a aproximação é aceitável se todos os valores esperados sob H_0 forem superiores a 5) o valor observado de um qui-quadrado[(1)] com um grau de liberdade, $Y = X^2$, onde $X \frown Gaussiana(0,1)$. Se $ad-bc > 0$, então $\frac{a}{b} > \frac{c}{d}$, ou seja a proporção de "sucessos" é maior no grupo da primeira linha do que no grupo da segunda linha; se $ad-bc < 0$, $\frac{a}{b} < \frac{c}{d}$, e dá-se portanto o inverso.

- Para cada uma das N tabelas calculamos $Y_j(obs.) = \sqrt{X^2_{2,2}(obs.)}$, $j = 1, 2, \ldots, N$, a que damos o sinal "+"se $ad-bc > 0$ e o sinal "−"se $ad-bc < 0$, a fim de regressar a observações Gaussianas.

- Somamos estes valores, e esta soma é o valor observado de uma variável aleatória $\sum_{j=1}^{N} Y_j$, com distribuição amostral $Gaussiana\left(0, \sqrt{N}\right)$, donde re-

(1) Berkson (1980) considera estatística de qui-quadrado qualquer estatística que compare frequências observadas com os seus valores esperados, e propõe à consideração do leitor cinco dessas estatísticas. Mais geralmente, as estatísticas de divergência (*power divergence statistics*) $2\,n\,\mathrm{I}^{\lambda}(\boldsymbol{O}) = \frac{2}{\lambda(\lambda+1)} \sum_{k=1}^{N} O_k \left[\left(\frac{O_k}{e_k}\right)^{\lambda} - 1\right]$ de Cressie e Read (1984), têm distribuição assintótica de qui-quadrado, com um número de graus de liberdade dado por $N - 1 -$ nº de parâmetros estimados. Note-se que para $\lambda = 1$ se obtém a usual estatística de qui-quadrado de Pearson. Um problema importante é determinar qual o valor de λ que corresponde à melhor aproximação. Veja-se Pestana and Vasconcelos (1999) ou Pestana e Velosa (2002).

sulta que $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} Y_j \frown Gaussiana(0,1)$, o que nos permite ter uma regra de decisão clara.

Na Tabela 2.1 exemplifica-se com uma situação fictícia:

Tabela 2.1: Evolução da doença no grupo experimental e no grupo de controlo.

		Caso	Controlo		X^2	$ad-bc$	Y_k
Hospital A	Não melhora	27	33	60	0.78	−	−0.88
	Melhora	41	37	78			
		68	70	138			
Hospital B	Não melhora	15	26	41	1.2	−	−1.10
	Melhora	29	32	61			
		44	58	102			
Hospital C	Não melhora	26	29	55	0.08	+	0.28
	Melhora	34	42	76			
		60	71	131			
Hospital D	Não melhora	22	22	44	0.07	−	−0.26
	Melhora	31	28	59			
		53	50	103			
Hospital E	Não melhora	22	19	41	0.02	+	0.14
	Melhora	25	23	48			
		47	42	89			
Hospital F	Não melhora	19	26	45	3.17	−	−1.78
	Melhora	28	18	46			
		47	44	91			
Hospital G	Não melhora	22	29	51	2.87	−	−1.69
	Melhora	35	24	59			
		57	53	110			
Hospital H	Não melhora	20	19	39	0.26	−	−0.51
	Melhora	20	15	35			
		40	34	74			

Suponha-se que em oito hospitais que aceitam doentes com uma determinada doença se procede a um ensaio clínico com atribuição aleatória de cada doente ao

grupo experimental (caso) ou ao grupo de controlo (controlo) a que se administra um placebo, registando-se se o doente melhora ou não melhora, obtendo-se os dados registados na Tabela 2.1. Ao nível usual $\alpha = 0.05$ nenhum dos resultados é, individualmente, significativo, não se podendo concluir heterogeneidade. Mas globalmente, será de manter a hipótese nula de que o efeito do tratamento não é distinto do da administração de um placebo?

A solução ingénua seria, naturalmente, adicionar frequências de celas correspondentes, e analisar a tabela 2×2 resultante. Mas um pouco de reflexão torna imediatamente óbvio que tal só faz sentido no caso, pouco provável, de as frequências relativas de cada classe nas diversas tabelas serem semelhantes. A metodologia atrás exposta, em contrapartida, permite sintetizar os resultados, haja ou não essa similitude de proporções.

Neste caso, observa-se $\sum_{j=1}^{8} Y_j = -5.80$ e $\frac{1}{\sqrt{8}} \sum_{j=1}^{8} Y_j = -2.05$, que é significativo, rejeitando-se então a hipótese nula de ser indiferente o tratamento experimental ou um mero placebo. Este exemplo ilustra bem as vantagens resultantes de uma síntese dos diversos estudos: porventura por escassez de dados, cada um dos estudos levou à decisão fraca que é manter H_0, mas a síntese, usando mais eficientemente toda a informação disponível nos estudos parciais, permitiu rejeitar H_0.

Este método só deve ser usado quando as diversas amostras não diferem excessivamente umas das outras, no que respeita à dimensão, e pudermos admitir independência entre elas. Em geral exige-se que a dimensão da maior seja inferior ao dobro da de menor dimensão, e que a proporção $\frac{a}{a+b}$ esteja entre 0.2 e 0.8. Em situações mais complicadas, usam-se análises mais sofisticadas, veja-se em Everitt (1992) o método de Cochran, que assume um modelo de base binomial comum. O método mais sofisticado deve-se a Mantel and Haenszel (1959), que assumem antes um modelo hipergeométrico, decerto mais "exacto", no sentido em que está mais próximo da realidade. Uma exposição detalhada dos métodos de Mantel–Haenszel encontra-se em Kuritz *et al.* (1988); note-se que diversos resultados largamente usados na análise de tabelas de contingência — teste de McNemar, teste de Cochran, teste dos multiplicadores de Lagrange, critério de Mantel–Haenszel generalizado

— podem ser reformulados como situações particulares dos métodos de Mantel–Haenszel. Deve ainda notar-se que, em sentido estrito, as conclusões são válidas para a amostra em estudo, e que a generalização para a população alvo tem que se apoiar numa análise da representatividade da amostra, veja-se Koch *et al.* (1980, 1982).

Gokhale and Kullback (1979) tratam extensivamente a problemática da globalização de informação de tabelas de contingência, que teve posteriormente grandes desenvolvimentos.

Este exemplo aborda ainda uma questão de grande simplicidade conceptual: todas as tabelas são construídas com a mesma metodologia. A situação típica em meta-análise é mais complexa. Por exemplo, Beecher (1955) estuda a eficácia terapêutica de placebos em circunstâncias diversas (dor pós-operatória, tosse, estados depressivos, angina de peito, dor de cabeça, enjoo, ansiedade, obstipação).

Os métodos estatísticos para lidar com a síntese de diversos estudos surgiram relativamente cedo (Tippett, 1931; Fisher, 1932; Cochran, 1937, 1954; Mantel and Haenszel, 1959), e pelo menos desde o trabalho pioneiro de Beecher (1955) sobre a eficácia terapêutica da administração de placebos que a meta-análise ganhou adeptos em ciências biomédicas. O termo meta-análise foi cunhado por Glass (1976), ao formalizar alguns aspectos do que até então se considerava uma panorâmica ("*review*") dos resultados publicados, veja-se a notável "*presidential address*"de Glass (1999) que Glass (1999) apresentou dos primeiros 25 anos da Meta-Análise no congresso da *American Psychology Society.* Números especiais de *Statistics in Medicine* (**6**, 1987), *Statistical Science* (**7**, 1992) e *Statistical Methods in Medical Research* (**2**, 1993), e o relatório sobre oportunidades de investigação (Draper *et al.*, 1992) do painel designado pela *National Academy of Sciences* dos EUA, catapultaram a meta-análise para o primeiro plano da investigação bioestatística, havendo presentemente uma rica literatura sobre o assunto, que cresce anualmente a ritmo exponencial. Veja-se também a notável série de artigos de Egger and Davey Smith, publicados em 1997-1998 no *British Medical Journal* a convite do editor, fazendo o ponto da situção dos padrões de publicação em Medicina e a imprescimbilidade de uma síntese meta-analítca dando uma perspectiva correcta sobre a inserção dos

resultados descritos no fluxo de investigação sobre o assunto

Neste capítulo introdutório, que em grande medida reproduz a conferência convidada de Pestana *et al.* (2006), aborda-se uma pequena fracção da problemática envolvida no campo muito vasto da síntese meta-analítica. Na secção 2 comentamos alguns dos traços mais salientes da meta-análise de ensaios clínicos e da meta-análise de testes de diagnóstico. Na secção 3 ocupamo-nos de dados resultantes de medições (em escala intervalar ou em escala de razões na classificação de Steven, 1946, que continua a ser essencial para a selecção apropriada das metodologias estatísticas adequadas para a análise dos dados), e abordamos a comparação de efeitos médios dos diversos tratamentos sem requerer homocedasticidade, o pressuposto clássico simplificador de que as dispersões nos diversos grupos são homogéneas. Aquele pressuposto faz sentido na análise de dados obtidos em experiências com planeamento tradicional, porque é de admitir que a obediência a um protocolo rígido sobre o que é medido e como se mede, e a medição dos efeitos (variável resposta) com os mesmos equipamentos, resulte em variâncias iguais nos diversos grupos, quando se condiciona à hipótese nula de homogeneidade populacional; evidentemente pressupõe-se ainda que os grupos foram constituídos com atribuição aleatória de cada unidade amostral a um dos grupos, para esbater qualquer eventual confundimento. Mas na meta-análise este discurso deixa de fazer sentido, e a heterocedasticidade — um assunto tantas vezes evitado pela aparente complexidade do teste t com um número de graus de liberdade fraccionário, usando o estimador de Smith–Welch–Satterthwaite — é, naturalmente, uma das questões angulares deste novo paradigma da Estatística.

Na secção 4 aborda-se a discussão de um exemplo de aplicação na investigação da eficiência de um fármaco, e dos possíveis efeitos secundários nocivos, porventura letais: o estudo notável sobre Ephedra e Ephredina de Shekelle *et al.* (2003), exemplar em quase todos os aspectos de Meta-análise, inclusive as suas limitações no que se refere a conclusões, quando usada com rigor.

A Meta-análise é uma área da Estatística em que, mais do que em qualquer outra, se entrelaçam questões sofisticadamente matemáticas com preocupações pragmáticas e de bom senso, e com a necessidade de ponderar cuidadosamente

como extrair conhecimento a partir da informação disponível — a qual, evidentemente, raramente reflecte de forma equilibrada a realidade, sendo evidente o viés na publicação que favorece, naturalmente, os estudos conclusivos (mesmo que as conclusões se revelem frágeis e efémeras! — veja-se Ioannidis (2005)).

Apesar de nos interessarmos mais, profissionalmente, pela perspectiva matemática, tentaremos não deixar as questões não matemáticas na sombra; e apenas na secção 3, sobre comparação de efeitos médios de tratamentos admitindo heterocedasticidade, massacramos os leitores com demonstrações pormenorizadas e técnicas. Deve-se isso à nossa convicção sobre a importância crescente deste assunto, ainda mal tratado na literatura da especialidade, convicção que é devidamente apoiada na opinião dos maiores especialistas, que não hesitam em colocar a questão da análise da escala (dispersão) no cerne do desenvolvimento actual da síntese meta-analítica:

> *In contrast,* [Draper *et al.* (1992)] *emphasizes the need to move this focus* [análise da localização] *to one of quantifying and reporting the heterogeneity between studies. This message reflects the methodological development seen in the statistics field during the last decade. Statistical methodology for meta-analysis has been moving away from approaches focused only on fitting a common estimate toward approaches that include estimating the extent and sources of heterogeneity among studies.*
>
> Stangl and Berry (2000)

Assim, apresentamos com algum detalhe o tratamento da comparação de efeitos médios no caso de aquisição de dados com variâncias distintas de grupo para grupo, um problema que não é em geral abordado mesmo em textos de nível avançado, apesar de actualmente diversos artigos publicados na prestigiosa revista *American Statistician* recomendarem o uso rotineiro do teste de Smith–Welch–Satterthwaite. A exposição detalhada e exemplificação adequada pode encontrar-se em Pestana e Velosa (2002, apresentando-se uma versão simplificada, mas menos geral, a partir da 2ª edição) ou, a nível mais abstracto, em Kendall and Stuart (1961) ou em Velosa (2003). Anote-se que o *software R* por defeito implementa o teste de Smith–Welch–Satterthwaite, sendo necessário uma instrução explicitando que se considera poder

assumir-se homocedasticidade para se obter o tradicional teste t de Student para amostras independentes, na comparação de médias de duas populações assumidas como gaussianas.

Finalmente a secção 5, de conclusões, capitaliza em toda a exposição anterior, lançando um aviso cautelar final: o excesso de informação pode tornar significativo o que é irrelevante, dando foros de conhecimento científico ao que não passa de mera observação ancorada na contingência das amostras.

Aproveitamos ainda para lançar a sugestão de publicação electrónica mais detalhada, nomeadamente remetendo para ficheiros documentais suplementares, por exemplo com os dados, e/ou descritivos do protocolo experimental, uma medida que poderia facilitar sínteses futuras. Tal era impossível há dez anos atrás, devido aos custos de publicação em suportes tradicionais; mas actualmente seria fácil, e poderia ser um passo substancial no avanço da cooperação e do esforço colectivo de organizar e construir conhecimento.

2.2 Meta-Análise, Ensaios Clínicos e Testes de Diagnóstico

Como é hoje de rigor em investigação científica, deve haver um protocolo pormenorizado do âmbito e objectivos visados na meta-análise. No que refere a testes de diagnóstico, um passo recomendável é pesquisar na *Medline*, visto ser uma fonte quase exaustiva quanto a ensaios realizados ou a decorrer em Inglês. O contacto com a *Cochrane Collaboration*[(2)] pode suplementar informação sobre estudos em outras áreas linguísticas — e o trabalho sistemático sob égide desta instituição é

[(2)] http://www.cochrane.org.

um padrão (Oxman, 1994) por que se pautam actualmente estas sínteses. A selecção dos estudos a incluir, por forma a reduzir subjectividade e enviesamento, é um passo importante (Chalmers *et al.*, 1981), veja-se a secção de considerações finais.

Há o perigo de um dos estudos dominar as conclusões, nomeadamente por ser de maior dimensão. É por isso preferível basear as conclusões em características amostrais relativas. Em dados binários, apesar de ocasionalmente os efeitos dos tratamentos serem expressos pelas diferenças de riscos, é mais usual usar riscos relativos, ou razões de vantagens (*odds ratios* — *OR*, também designados por razões de possibilidades). É também frequente uma transformação logarítmica, que ao simetrizar tem o efeito subsidiário de proporcionar uma melhor aproximação gaussiana para a distribuição amostral do estimador pretendido.

No que respeita a dados em escala intervalar ou em escala absoluta é usual recorrer a diferenças de efeitos médios divididos pelo desvio padrão global das medições, para moderar o efeito de estudos preponderantes. Quando a amostra sugere grande assimetria na população subjacente, as transformações de Box-Cox, de que a transformação logarítmica é um caso extremo, podem favorecer o recurso a metodologias mais fecundas na exploração dos dados.

Suponha-se que consideramos N ensaios clínicos, cada qual com um grupo experimental e um grupo de controlo, para a elaboração da síntese meta-analítica. Denote-se θ_k o verdadeiro efeito do tratamento no k-ésimo ensaio, $\hat{\theta}_k$ o efeito estimado do tratamento nesse k-ésimo ensaio, v_k a variância de $\hat{\theta}_k$, e usem-se como pesos $w_k = \frac{1}{v_k}$, $k = 1, \ldots, N$.

A hipótese nula natural é que o tratamento não produz efeito em qualquer um dos ensaios,

$$H_0: \quad \theta_k = 0, \ \forall\, k = 1, \ldots, N$$

Thompson (1993) justifica o uso de duas estatísticas de teste:

1. Um teste similar ao T^2 de Hotelling, $\sum_{k=1}^{N} w_k \hat{\theta}_k^2 \overset{\circ}{\frown} \chi^2_{N-1}$, quando não se tem

qualquer alternativa especificada em jogo.

2. Um teste direccional mais potente (e mais plausível e interessante do ponto de vista biomédico) $\frac{\left(\sum_{k=1}^{N} w_k \hat{\theta}_k\right)^2}{\sum_{k=1}^{N} w_k} \overset{\circ}{\frown} \chi^2_1$, quando se considera uma alternativa unilateral, H_A: $\theta_k < 0,\ \forall\, k = 1, \ldots, N$, ou H_A: $\theta_k > 0,\ \forall\, k = 1, \ldots, N$

Este segundo teste é mais potente para a alternativa H_A: $\theta_k = \theta\ (\neq 0)$ (em qualquer dos ensaios o tratamento produz o mesmo efeito médio θ), $\forall\, k = 1, \ldots, N$. Num e noutro caso, assume-se que os efeitos dos tratamentos têm distribuição gaussiana, determinante para a distribuição amostral assintótica ser qui-quadrado.

No caso de frequências em tabelas 2x2 usa-se o teste de Mantel-Haenszel (1959), que de facto pode ser deduzido como caso particular do teste direccional acima, veja-se Yusuf *et al.* (1985). O sucesso possível usando meta-análise fica bem documentado com o estudo de Collins *et al.* (1985) (cf. também os comentários feitos por Thompson and Pocock, 1991) sobre a utilização de diuréticos como prevenção de pré-eclampsia em grávidas, de que abaixo reproduzimos a Tabela 2.2.

Tabela 2.2: Prevalência de pré-eclampsia em nove ensaios clínicos sobre diuréticos.

ensaio	prev. (casos)	prev. (controlo)	*odds ratio*
Weseley	11% (14/131)	10% (14/136)	1.04
Flowers	5% (21/385)	13% (17/134)	0.40
Menzies	25% (14/57)	50% (24/48)	0.33
Fallis	16% (6/38)	45% (18/40)	0.23
Cuadros	1% (12/1011)	5% (35/750)	0.25
Landesman	10% (138/1370)	13% (175/1336)	0.74
Krans	3% (15/506)	4% (20/524)	0.77
Tervila	6% (6/108)	2% (2/103)	2.97
Campbell	42% (65/153)	39% (40/102)	1.14

$OR < 1$ significa que o tratamento é eficaz. Com os estudos acima referidos

(pelo nome do primeiro autor, como na fonte documental usada, Collins *et al.*, 1985), a estimativa-síntese de OR é 0.66, e o intervalo de confiança com coeficiente de confiança 0.95 é $(0.57, 0.79)$. Admite-se por isso, como consequência desta síntese, que o uso de um diurético durante a gravidez é uma boa prevenção contra a pré-eclampsia.

Note-se que a linha de força da investigação não reside em saber qual o diurético ou dosagem que produz melhores resultados, como nos testes *post hoc* em análise da variância. A questão que nos ocupa é: Diurético como prevenção de pré-eclampsia — Sim ou Não?

Na Tabela 2 observa-se que o estudo de Cuadros e de Landesman "afoga" os outros, por usar uma amostra de muito maior dimensão (mais do dobro de unidades amostrais do que as usadas em qualquer outro). Em particular o estudo de Tervila, cujo OR se afasta substancialmente do dos outros estudos, não tem qualquer peso na síntese final. Este é um dos problemas da meta-análise: em muitas situações, a síntese é excessivamente influenciada por apenas um dos estudos incluídos. Neste exemplo, a estimativa meta-analítica 0.66 do OR está mais próxima da estimativa 0.74 de Landesman — o que não é de admirar, pois a dimensão da amostra de Landesman, 1336, excede a de qualquer dos outros estudos, tendo um peso de 42% na amostra combinada (o estudo de Tervila tem uma amostra que não chega a ser 4% da amostra global).

Na formulação clássica de meta-análise, o modelo de efeitos fixos, que acima usámos para calcular a estimativa pontual e intervalo de confiança que sintetizam os nove estudos incluídos, gozava de algum favoritismo, porventura por ser de formulação e implementação menos trabalhosa, e fornecer intervalos de confiança mais pequenos.

Presentemente, a maior parte dos trabalhos de síntese procura explorar *IPD* — *individual patient data* e modelos hierárquicos, em que se modelam os efeitos médios: $\theta_k \frown F$ (admite-se identidade distribucional), com valor médio $\mathbb{E}(\theta_k) = \theta$ e $\text{var}[\theta_k] = \sigma^2$, enquanto na amostra recolhida os estimadores $\hat{\theta}_k$ têm valor médio $\mathbb{E}\left(\hat{\theta}_k\right) = \theta_k^*$ e variância $\text{var}\left(\hat{\theta}_k\right) = v_k$, respectivamente. O objectivo de usar

modelos com efeitos aleatórios é tornar as conclusões extensivas a uma maior generalidade de indivíduos/situações. Usam-se geralmente pesos $w_k = \frac{1}{v_k}$ para tornar preponderantes os estudos mais informativos. Para efeitos comparativos com o que concluímos quanto ao modelo de efeitos fixos: com os dados sobre diuréticos como prevenção de pré-eclampsia, o modelo de efeitos aleatórios leva a uma estimativa 0.60 de OR, sendo o intervalo de confiança de 95% $(0.40, 0.89)$, ligeiramente maior que o construído no modelo de efeitos fixos.

A meta-análise, para além da síntese de ensaios clínicos que atrás descrevemos de forma sucinta, tem um papel importante na combinação de estudos sobre testes de diagnóstico. O problema da selecção dos estudos a incluir é semelhante, mas a problemática seguinte centra-se numa questão de grande relevo: como harmonizar estudos com sensibilidades e especificidades diversas?

De facto, experimentadores diferentes podem pôr mais ou menos ênfase na sensibilidade ($s = \mathbb{P}[+ \mid D]$, a probabilidade de detectar doentes, positivos verdadeiros) ou na especificidade ($e = \mathbb{P}\left[- \mid \overline{D}\right]$, a probabilidade de não incomodar os não doentes, negativos verdadeiros) dos testes de diagnóstico, isto é escolher "pontos de corte" diversos. Como construir uma curva $SROC$ ("*Summary Receiver Operating Characteristic*")[3] a partir de pontos (1–especificidade, sensibilidade) obtidos em diversos estudos?

Comecemos por exemplificar, para clarificar ideias, o caso simples da construção da curva ROC. A porfíria aguda resulta de produção deficitária de diaminase de porfibilogénio, pelo que a determinação de níveis séricos desta enzima serve, naturalmente, como meio auxiliar de diagnóstico. Veja-se em Pestana e Velosa (2002, p. 272-276) como prevalência e outras informações correlacionadas com a a avaliação das probabilidades *a priori* de ter a doença são ponderadas pela sensibilidade $s = 93.6\%$ e pela especificidade $e = 84\%$ (quando o ponto de corte usado é 99 unidades por cm^3 de sangue, considerando-se positivos os indivíduos com teor sérico de diaminase de porfibilogénio inferior a esse valor) para se obter o valor preditivo

[3] Por simplicidade, vamos a partir de agora falar de curvas ROC, sendo claro do contexto que quando usamos pontos obtidos em diversos estudos se trata de uma síntese, e portanto de uma curva $SROC$.

positivo $\mathbb{P}\left[D|+\right]$ e o valor preditivo negativo $\mathbb{P}\left[\overline{D}|-\right]$ do teste em cada um uma das três situações analisadas.

Suponha-se então que, antes de ter sido acordado que o teste seria feito colocando a fasquia em 99, se estudava o sangue de 53 doentes e de 100 não doentes a diversos níveis, e se registava

- o número x de não doentes com resultado negativo, por terem um nível sérico superior ao patamar T,
- o número y de doentes detectados, isto é com resultado positivo, por terem um nível sérico inferior ao patamar T,

por exemplo

T	80	85	90	95	100	105	110	115	120
x	0	2	5	10	17	28	42	64	100
y	0	24	39	44	49	50	51	52	53

As correspondentes frequências relativas em cada linha dar-nos-iam então as coordenadas $(1-e,\, s)$

T	80	85	90	95	100	105	110	115	120
$1-e$	0.000	0.020	0.050	0.100	0.170	0.280	0.420	0.640	1.000
s	0.000	0.453	0.736	0.830	0.925	0.943	0.962	0.981	1.000

da curva *ROC* empírica. Adiante se indica como se constrói uma curva alisada a partir de algumas coordenadas.

Evidentemente as decisões da Organização Mundial de Saúde, ou de entidades nacionais como o Infarmed, não devem basear-se apenas na evidência de um estudo, feito em geral com recursos limitados.

Mas a simples aglomeração de dados de diversos estudos pode dar resultados absurdos, devido a confundimento. Por exemplo, suponha-se que num dos estudos se tem

diagnóstico \estado	D – doente	$\overline{D}$ – não doente
positivo	200	10
negativo	300	60

Tem-se sensibilidade $= \frac{200}{200+300} = 0.4$, especificidade $= \frac{60}{10+60} = 0.86$, e o $OR = \frac{200/10}{300/60} = 4$ indica que esta análise tem um poder discriminativo razoável (OR=1 é equivalente a $s = 1 - e$, pelo que consideramos $OR \approx 1$ um indicador de escasso poder discriminativo).

Noutro estudo, os dados são

diagnóstico \estado	D – doente	$\overline{D}$ – não doente
positivo	20	200
negativo	5	200

Tem-se sensibilidade $= \frac{20}{20+5} = 0.8$, especificidade $= \frac{200}{200+200} = 0.5$, e o $OR = \frac{20/200}{5/200} = 4$, indicando que também esta análise tem um poder discriminativo razoável.

A simples aglutinação só dá asneira:

diagnóstico \estado	D – doente	$\overline{D}$ – não doente
positivo	220	210
negativo	305	260

com sensibilidade $= \frac{220}{220+305} = 0.42$, especificidade $= \frac{260}{210+260} = 0.55$, e o $OR = \frac{220/210}{305/260} = 0.89$ indicando que esta análise tem um poder discriminativo péssimo. Este efeito perverso da adição dos valores observados em tabelas de contingência é conhecido por paradoxo de Simpson[(4)].

A forma de meta-analisar tem que ser muito mais sofisticada. Para um bom sumário da sensibilidade/especificidade correspondentes a testes com pontos de

[(4)] Deveria de facto ser denominado paradoxo de Yule (1903), e não de Simpson (1951).

corte diversos, faz-se um gráfico de pontos cujas coordenadas são: abcissa — proporção de falsos positivos $= \mathbb{P}\left[+ \mid \overline{D}\right] = 1 - e$; ordenada — proporção de positivos verdadeiros$= \mathbb{P}\left[+ \mid D\right] = s$. É de esperar um crescimento rápido da sensibilidade de 0 a valores próximos de 0.90 quando a abcissa $x = 1-$ especificidade varia de 0 a 0.10, seguindo-se um crescimento muito lento até 1, como se observa na Figura 2.1. Por outras palavras, esperamos que a curva *ROC* esteja consideravelmente acima da bissectriz $s = 1 - e$.

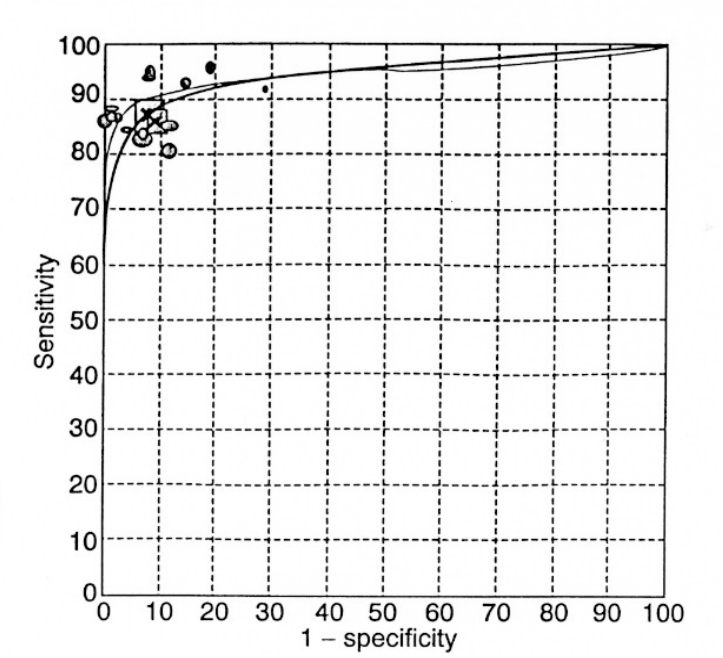

Figura 2.1: Curva ROC em estudos de ultrassons para estenose da artéria carótida.

Se a curva *ROC* passasse pelo ponto (0,1), isso significaria que era possível ter a situação ideal de 100% de sensibilidade e de especificidade. Mas não é possível ter o melhor de dois mundos, é evidente que para aumentar a sensibilidade alguma coisa há que sacrificar na especificidade, e vice-versa. Mas por outro lado também esperamos que a curva *ROC* se afaste visivelmente da bissectriz $s = 1-e \Leftrightarrow OR = 1$, que corresponderia a um teste de diagnóstico muito pouco recomendável, no sentido em que o ganho em sensibilidade de ponto de corte para ponto de corte corresponde exactamente à perda de especificidade.

O *odds ratio* OR pode ser escrito $OR = \frac{LR^+}{LR^-}$, onde $LR^+ = \frac{s}{1-e}$ e $LR^- = \frac{1-s}{e}$ são as "razões de verosimilhanças" que nos permitem, usando a fórmula de Bayes, transformar as vantagens pré-teste em vantagens pós-teste. Se por outro lado considerarmos como avaliação da assimetria nos ganhos de sensibilidade *versus* perdas de especificidade $S = LR^+ \times LR^-$, admitindo que há assimetria entre o que se ganha numa das características e se perde na outra, ao alterar o ponto de corte (isto é, o teste favorece, para cada ponto de corte, o diagnóstico correcto ou de doentes ou de não doentes), parece razoável procurar uma curva exprimindo OR como função de S. Espera-se então que uma transformação logarítmica linearize essa relação:

$$\ln OR = \ln LR^+ - \ln LR^- \approx \alpha \ln S + \beta = \alpha \left(\ln LR^+ + \ln LR^- \right) + \beta$$

equivalente a

$$\ln \frac{a}{b} + \ln \frac{b+d}{a+c} - \ln \frac{c}{d} - \ln \frac{b+d}{a+c} \approx \alpha \left(\ln \frac{a}{b} + \ln \frac{b+d}{a+c} + \ln \frac{c}{d} + \ln \frac{b+d}{a+c} \right) + \beta$$

ou ainda

$$\ln \frac{a}{b} - \ln \frac{c}{d} \approx \alpha \left(\ln \frac{a}{b} + \ln \frac{c}{d} \right) + \beta^*,$$

onde $\beta^* = \beta + \ln \left(\frac{b+d}{a+c} \right)^{2\alpha}$.

Uma forma simples de estimar essa curva *ROC* (Littenberg *et al.*, 1990) é então, para cada uma das tabelas 2×2

	D	$\overline{D}$
$+$	a_k	b_k
$-$	c_k	d_k

- Calcular as expressões $\hat{U}_k$ e $\hat{V}_k$, ajustados por causa de eventuais zeros,

$$\hat{U}_k = \ln \frac{c_k + 0.5}{d_k + 0.5} \qquad \text{e} \qquad \hat{V}_k = \ln \frac{a_k + 0.5}{b_k + 0.5}\,.$$

- Ajustar uma recta (ajustamento resistente, cf. Pestana e Velosa, 2002, p. 169–178, ou usando mínimos quadrados ponderados) aos pontos $\left(\hat{V}_k + \hat{U}_k,\, \hat{V}_k - \hat{U}_k\right)$.

- Deduzir daquela recta de regressão a expressão analítica da curva *ROC*, como exemplificado na Figura 2.1. Note-se que isto corresponde a admitir uma forma muito genérica para as curvas *ROC*,

$$\left(\frac{s}{1-e}\right)^{1-\alpha}\left(\frac{e}{1-s}\right)^{1+\alpha} = \frac{e^{\alpha+1}\,(1-e)^{\alpha-1}}{s^{\alpha-1}\,(1-s)^{\alpha+1}} = \exp(\beta) = C.$$

Se não houver condicionantes extra, parece natural definir o ponto de corte óptimo, no sentido em que dá um equilíbrio apelativo no que refere sensibilidade e especificidade, com base no ponto mais próximo do óptimo ideal — inatingível — (0,1). Por outras palavras, trata-se de minimizar o quadrado da distância $d^2 = (1-s)^2 + (1-e)^2$ de um ponto genérico $(1-e, s)$ ao canto superior esquerdo do quadrado $\{(x,y):\ 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$.

Mas, evidentemente, é em geral mais sensato considerar penalizações definindo funções de perda L_1 e L_2 associadas aos diagnósticos errados, diagnosticar negativo um doente ou diagnosticar positivo um não doente, respectivamente, e o problema efectivo é procurar o ponto da curva *ROC* ajustada que minimiza a soma $L_1 + L_2$ das perdas.

Veja-se também van Belle (2008, secção 4.10), que aborda com detalhe e pragmatismo a construção de uma curva *ROC*.

Um foco de preocupação é se a curva *ROC* assim estimada denuncia heterogeneidade no que refere a patamares de decisão (*thresholds*), ou na capacidade discriminativa, ou em ambos. Deve procurar explicar-se essa heterogeneidade, indicar ao menos se deriva de características específicas diversas dos testes de diagnóstico que estamos a tentar harmonizar, de diferenças populacionais, ou de planeamentos experimentais diversos.

2.3 Uma Questão de Escala — Como Lidar com Heterogeneidade em Populações Gaussianas

A investigação científica obriga a inúmeras comparações de efeitos médios que se admite poderem ser modelados por gaussianas, quer na situação em que se tem medições repetidas, ou antes e depois de um tratamento (caso em que se usa o teste t de Student para dados correlacionados), quer na situação em que se deseja comparar o efeito de dois tratamentos, um dos quais é, muitas vezes, um placebo, usando amostras independentes (usa-se o teste t para amostras independentes); a comparação simultânea de diversos tratamentos, com base, também, em amostras independentes, faz-se recorrendo a análise da variância. A studentização elimina o parâmetro perturbador de escala para viabilizar a inferência sobre a localização média populacional; a análise da variância, mais sofisticada, usa o facto de a variância ser o menor de todos os momentos de segunda ordem para verificar a influência que "perturbações amostrais" da média (médias dentro dos grupos) induzem nas somas de quadrados, e assim testar uma hipótese geral de igualdade de todas as médias.

Na teoria clássica assume-se igualdade de variâncias nos diversos grupos que estão a ser comparados. Por isso, nessa abordagem clássica, o teste de igualdade de médias em populações gaussianas com a mesma variância é, afinal, um teste de homogeneidade populacional.

Será razoável assumir que as variâncias das diversas subpopulações são iguais, quando afinal o que recolhemos são as respostas a tratamentos diferentes, que estão a induzir, eventualmente, a diferenciação das subpopulações de uma mesma população de base?

Sim, sem dúvida, se o protocolo experimental for bom e devidamente implementado. Na fase de planeamento, a regra de ouro é: controlo tudo o que puder, o que não puder controlar aleatorize. Assim, se quisermos comparar quatro tratamentos e usarmos ainda um placebo que estabeleça um nível 0 de eficácia, ao

admitir uma nova unidade experimental deve usar-se uma metodologia de aleatorização. Neste caso, por exemplo, usar um programa como o Excel em que a simples instrução "= rand( )" tem como retorno um pseudo-aleatório em $(0, 1)$, e depois alocar essa unidade experimental ao grupo 1, 2, 3, 4 ou 5 consoante esse número aleatório esteja em $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, ou $(0.8, 1]$, respectivamente. Cada unidade é alocada a qualquer um dos grupos com probabilidade 0.2, independentemente do que acontece a qualquer outra, pelo que podemos considerar que as amostras obtidas são realizações de amostras aleatórias independentes. Os grupos vão ter dimensões eventualmente diferentes[(5)], mas provavelmente pouco diferentes, e é razoável esperar que essa aleatorização tenha disciplinado os factores de confundimento que não tínhamos sabido controlar. Por isso, e porque fazer médias é proceder a um alisamento bastante radical, que reduz substancialmente a dispersão (o erro padrão da média é $\frac{\sigma}{\sqrt{n}}$), a dispersão dos dados reflecte alguma

(5) Se quisermos grupos da mesma dimensão, decidimos à partida qual é essa dimensão, e vai-se procedendo como descrito, até qualquer um dos grupos ficar completo. A partir daí, para garantir características análogas às do caso em que se permite grupos com diferentes dimensões, quando o pseudo-aleatório corresponde a um grupo já completo essa unidade amostral é rejeitada.

O procedimento ingénuo que seria ir distribuindo apenas pelos grupos ainda incompletos traz o risco de confundimento. Suponha-se por exemplo que se está a fazer um estudo sobre tratamento de problemas respiratórios, e que se procura ter 12 indivíduos no grupo experimental e 12 no grupo de controlo. No caso de este ficar completo quando ainda só temos 5 indivíduos no grupo experimental e por isso decidirmos enfiar os próximos 7 doentes no grupo experimental, e tal ocorrer na primavera, incluiria porventura no grupo experimental um número desequilibrado de doentes com febre dos fenos.

Evidentemente em muitas situações a escassez de unidades amostrais torna aquele modo de proceder escandalosamente perdulário. Uma alternativa mais inteligente é gerar logo à cabeça, na situação exemplificada, 12 números da distribuição uniforme discreta com suporte $\{1, 2, \ldots 24\}$, sem repetição, do tipo $8, 12, 1, 23, 14, 11, 19, 5, 2, 7, 16, 22$. Os 12 indivíduos a incluir no grupo experimental seriam então os 1º, 2º, 5º, 7º, 8º, 11º, 12º, 14º, 16º, 19º, 22º e 23º que aceitassem integrar o estudo, enquanto os 3º, 4º, 6º, 9º, 10º, 13º, 15º, 17º, 18º, 20º, 21º e 24º seriam incluídos no grupo de controlo. São questões elementares, mas muitas vezes ultrapassadas com alguma ligeireza na fase de recolha de dados, invalidando ou pelo menos comprometendo o valor de trabalhos que mereceriam melhor destino.

Grupos equilibrados proporcionam ortogonalidade, e a interpretação dos resultados é mais evidente, veja-se Gilbert (1989).

variabilidade residual, e espera-se que seja dominada pelo erro de medição, decorrente do protocolo experimental, da preparação técnica de quem a faz, e da acribia dos instrumentos de medição usados. Ora, espera-se que a este respeito qualquer dos grupos tenha processamento análogo, e por isso é razoável, na perspectiva do planeamento de experiências clássico, admitir homocedasticidade.

Tal não é o caso, evidentemente, em sínteses meta-analíticas. Muitas vezes, o que há de comum nos diversos estudos considerados é apenas o facto de usarem um tratamento semelhante (por exemplo, administração de um diurético, que nem sequer tem que ser o mesmo) para o grupo de controlo.

Apesar de haver agora a convicção de que as questões de heterogeneidade merecem um tratamento protagonista, raros são os planeamentos experimentais em que a questão de dispersões eventualmente diferentes é abordada. Muitas vezes procede-se a transformações de dados para se obter uma melhor aproximação gaussiana, mas sem qualquer atenção à heterocedasticidade de raiz, ou induzida pela transformação usada.

Mosteller and Chalmers (1992), por exemplo, referem de passagem que "*In a paper in the dissertation, Larholt, Tsiatis and Gelber (1989) developed methods that considerably improve the coverage* [...] *by using a t-distribution with degrees of freedom based on an approximation suggested by Satterthwaite*". Mas em seguida escrevem laconicamente: "*The details are too extensive to give here.*"

A importância da aproximação de Smith-Welch-Satterthwaite, nomeadamente em áreas biomédicas, é patente da atenção que este tema tem nas edições mais recentes de famosos livros de Bioestatística como Samuels and Witmer (1999) e Zar (1999), ou de Planeamento de Experiências, como Oehlert (2000), e recomendações várias publicadas na *American Statistician* (Mose and Stevens, 1992, por exemplo). Mas as bases estatístico-matemáticas apenas se encontram em livros avançados, como Kendall and Stuart (1961) ou Casella and Berger (2002), com grandes saltos "ao cuidado do leitor" entre os passos. A demonstração que seguidamente apresentamos é consideravelmente mais simples do que a publicada em Pestana e Velosa (2002).

Comecemos por notar que se extrairmos amostras aleatórias independentes $\boldsymbol{X} = \left(X_1, \ldots, X_{n_1}\right)$ e $\boldsymbol{Y} = \left(Y_1, \ldots, Y_{n_2}\right)$ de populações $X \frown Gaussiana\left(\mu_X, \sigma_X\right)$ e $Y \frown Gaussiana\left(\mu_Y, \sigma_Y\right)$, respectivamente, então

$$\overline{X} \frown Gaussiana\left(\mu_X, \frac{\sigma_X}{\sqrt{n_1}}\right) \text{ e } \overline{Y} \frown Gaussiana\left(\mu_Y, \frac{\sigma_Y}{\sqrt{n_2}}\right).$$

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

$$\overline{X} - \overline{Y} \frown Gaussiana\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}\right).$$

Estandardizando temos

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \frown Gaussiana\left(0, 1\right),$$

um resultado de pouco préstimos se o objectivo for inferência sobre a diferença das médias $\mu_X - \mu_Y$, pois nem $\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$, sob $H_0 : \mu_X - \mu_Y = \Delta$ será uma boa estatística de teste, devido aos parâmetros perturbadores σ_X^2 e σ_Y^2, nem $\frac{\overline{X} - \overline{Y} - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$ será uma variável frulcral interessante para $\mu_X - \mu_Y$, pela mesma razão, a menos que estejamos na situação invulgar de σ_X^2 e σ_Y^2 serem conhecidos quando $\mu_X - \mu_Y$ é desconhecido!

Como σ_X^2 e σ_Y^2 são desconhecidos, é natural recorrer aos estimadores centrados S_X^2 e S_Y^2, e usar a "studentização"

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}.$$

Uma vez que a dedução da distribuição exacta daquela variável não parece factível, vamos aproximar o quadrado $\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}$ do denominador por uma variável $\sigma^2 \frac{Y_\nu}{\nu}$, onde $Y_\nu \frown \chi_\nu^2$, e ν e σ^2 são escolhidos de tal forma que

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

e

$$\text{var}\left(\frac{S_{X}^{2}}{n_{1}}+\frac{S_{Y}^{2}}{n_{2}}\right)=\text{var}\left(\sigma^{2}\,\frac{Y_{\nu}}{\nu}\right)=\frac{2\,\sigma^{4}}{\nu}\,.$$

Ora $\mathbb{E}\left(\frac{S_{X}^{2}}{n_{1}}+\frac{S_{Y}^{2}}{n_{2}}\right)=\frac{\sigma_{X}^{2}}{n_{1}}+\frac{\sigma_{Y}^{2}}{n_{2}}$, e de $\frac{(n_{1}-1)\,S_{X}^{2}}{\sigma_{X}^{2}}\frown\chi_{n_{1}-1}^{2}$ e $\frac{(n_{2}-1)\,S_{Y}^{2}}{\sigma_{Y}^{2}}\frown\chi_{n_{2}-1}^{2}$ segue-se que $\text{var}\left(\frac{S_{X}^{2}}{n_{1}}+\frac{S_{Y}^{2}}{n_{2}}\right)=\frac{2\sigma_{X}^{4}}{n_{1}^{2}(n_{1}-1)}+\frac{2\sigma_{Y}^{4}}{n_{2}^{2}(n_{2}-1)}$. Assim, a aproximação que estamos a fazer usa uma variável com o mesmo valor esperado e a mesma variância da que vai substituir sse

$$\frac{\sigma_{X}^{2}}{n_{1}}+\frac{\sigma_{Y}^{2}}{n_{2}}=\sigma^{2}$$

e

$$\nu=\frac{\left(\frac{\sigma_{X}^{2}}{n_{1}}+\frac{\sigma_{Y}^{2}}{n_{2}}\right)^{2}}{\frac{\sigma_{X}^{4}}{n_{1}^{2}(n_{1}-1)}+\frac{\sigma_{Y}^{4}}{n_{2}^{2}(n_{2}-1)}}\,.$$

Consequentemente, procedemos à aproximação

$$\frac{\overline{X}-\overline{Y}-(\mu_{X}-\mu_{Y})}{\sqrt{\frac{S_{X}^{2}}{n_{1}}+\frac{S_{Y}^{2}}{n_{2}}}}\approx\frac{\overline{X}-\overline{Y}-(\mu_{X}-\mu_{Y})}{\sqrt{\sigma^{2}\,\frac{Y_{\nu}}{\nu}}}=\frac{\overline{X}-\overline{Y}-(\mu_{X}-\mu_{Y})}{\sigma}\Big/\sqrt{\frac{Y_{\nu}}{\nu}}\,.$$

Mas como $\sigma^{2}=\frac{\sigma_{X}^{2}}{n_{1}}+\frac{\sigma_{Y}^{2}}{n_{2}}$, 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, em populações gaussianas, mutuamente independentes. Quer isto dizer que

$$\frac{\overline{X}-\overline{Y}-(\mu_{X}-\mu_{Y})}{\sqrt{\frac{S_{X}^{2}}{n_{1}}+\frac{S_{Y}^{2}}{n_{2}}}}\approx t_{\nu}\,,$$

onde ν pode ser estimado por

$$\tilde{\nu}=\frac{\left(\frac{S_{X}^{2}}{n_{1}}+\frac{S_{Y}^{2}}{n_{2}}\right)^{2}}{\frac{S_{X}^{4}}{n_{1}^{2}(n_{1}-1)}+\frac{S_{Y}^{4}}{n_{2}^{2}(n_{2}-1)}}\,.$$

Provámos assim, para a variável $\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}}$, fulcral para $\mu_X-\mu_Y$,

Teorema.

Se $\boldsymbol{X}=\left(X_1,\ldots,X_{n_1}\right)$ *com os* $X_k \frown$ *Gaussiana* (μ_X,σ_X) *independentes, for independente de* $\boldsymbol{Y}=\left(Y_1,\ldots,Y_{n_2}\right)$, *com os* $Y_j \frown$ *Gaussiana* (μ_Y,σ_Y) *independentes, então*

$$T^* = \frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} \simeq t_{\widetilde{\nu}}$$

onde o número de graus de liberdade $\widetilde{\nu}$ *é estimado por*

$$\frac{\left(\frac{s_X^2}{n_1}+\frac{s_Y^2}{n_2}\right)^2}{\frac{s_X^4}{n_1^2(n_1-1)}+\frac{s_Y^4}{n_2^2(n_2-1)}}.$$

Na prática, aproximamos por um número de graus de liberdade natural.

Mais geralmente, para contrastes $\sum_{k=1}^{g} w_k\mu_k = 0$ usamos

$$t = \frac{\sum_{k=1}^{g} w_k\overline{x}_{k\bullet}}{\sqrt{\sum_{k=1}^{g}\frac{w_k^2 s_k^2}{n_k}}}$$

que tem distribuição aproximada pela de uma t com ν graus de liberdade, onde o número de graus de liberdade é estimado 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) ou Pestana e Velosa (2001, p. 506–510).

Não deixa de ser interessante apontar que os poucos livros que incluem este resultado referem, em geral, apenas o trabalho de Satterthwaite (1946), o qual atribui a ideia original a Smith (1936).

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

Kendall and Stuart (1961), pelo contrário, atribuem todo o crédito a Welch (1936, 1938), anterior e com sofisticação matemática muito superior à do trabalho de Satterthwaite. Welch (1951) voltou ao assunto mais tarde, mostrando como a sua abordagem podia ser reformulada por forma a comparar $k \geq 2$ médias.

Samuels and Witmer (1999), que consistentemente aborda a comparação de médias nesta perspectiva mais alargada, credita a Welch e a Satterthwaite os métodos que emprega, sem no entanto dar qualquer referência bibliográfica precisa.

O excelente *package* estatístico R, de utilização livre (pode obter-se em http://cran.r–project.org/), ao testar a diferença entre duas médias usa por defeito a estatística de Welch–Satterthwaite (para se conseguir realizar o teste clássico é necessário declarar explicitamente que consideramos que as duas populações têm variâncias iguais, `> t.test(A, B, var.equal=TRUE)`; neste ponto, diverge da versão comercial S, que por defeito executa o teste t clássico).

Anotamos ainda que também se pode fazer o *download* do *package* adicional

meta v. 0.5, da autoria de Guido Schwarzer, que permite realizar facilmente algumas tarefas de síntese meta-analítica (perdoe-se o pleonasmo), em particular no que se refere a modelos de efeitos fixos e de efeitos aleatórios, e funções para testar enviesamento. Alternativamente, *MetaStat*, de Rudner, Glass, Evartt, and Emery, também pode ser obtido livremente em http://www.edres.org/meta/metastat.htm.

2.4 Meta-Análise na Avaliação Retrospectiva de Tratamentos

Quando se administra a diferentes indivíduos a mesma dose de um medicamento, é de esperar grande variabilidade das concentrações plásmicas. Tal deve-se a grande número de covariáveis — idade, sexo, peso, medicamentos que tomou ou toma, doenças, características genéticas e outros marcadores biológicos. Mas por debaixo deste aparente caos superficial existe coerência, de que se ocupa a farmacocinética. O conhecimento detalhado de como decorre a absorção, a distribuição e a eliminação de medicamentos e dos metabolitos derivados com o decorrrer do tempo é essencial para a determinação de dosagens e intervalos entre a sua administração, nomeadamente na tentativa de controlar efeitos secundários adversos.

Não admira portanto que foruns legislativos internacionais procurem impor regras estritas de avaliação no que se refere a fármacos usados em "medicina natural", ou a produtos vendidos como aditivos alimentares, publicitados como benéficos ou favoráveis a efeitos pretendidos ansiosamente por faixas expressivas de consumidores.

De facto, a miragem de conseguir, com pouco esforço e sem sacrifícios, realizar ambições — como controlar o peso sem dietas penosas, prescindir de fumar sem *stress*, conseguir melhor desempenho sexual ou desportivo do que aquilo para que a

natureza nos dotou — é a base de uma próspera cadeia económica que movimenta uma fatia importante do PIB de qualquer país. Por vezes, o êxito desses produtos é a raiz da sua desgraça — a lei dos grandes números traz para as luzes da ribalta efeitos secundários adversos, que podem mesmo levar a incapacidade permanente ou mesmo morte.

O alarme público decorrente da publicitação desses factos na comunicação social determina então a acção de agências oficiais.

Num caso exemplar — suspeita acidentes vasculares cerebrais, e eventualmente de mortes decorrentes do consumo de Ephedra e Ephedrina, substâncias muito populares em programas de perda de peso e/ou de ganho muscular e correspondente progresso em actividades desportivas — a *Agency for Healthcare Research and Quality*, dependente do *Department of Health and Human Services* do governo dos EUA, encomendou um estudo retrospectivo (Shekelle *et al.*, 2003), que recorreu intensivamente a Meta-análise para investigar questões de eficiência/segurança:

- Aditivos alimentares contendo Ephedra/Ephedrina, eventualmente em conjugação com cafeína, provocam efectivamente redução de peso?
- Aditivos alimentares contendo Ephedra/Ephedrina, eventualmente em conjugação com cafeína, melhoram de facto o desempenho físico?
- Pode estabelecer-se uma relação causal entre o risco de efeitos secundários nocivos (náusea, vómitos, ansiedade, depressão, palpitações, hiperactividade involuntária, enfarte de miocárdio, acidente vascular cerebral, convulsões, morte) e o uso aditivos alimentares contendo Ephedra/Ephedrina?

O relatório de Shekelle *et al.*, (2003), a muitos títulos exemplar, não deixa por isso de ser uma montanha a parir um rato.

As duzentas e cinquenta páginas de relatório, baseado em 56 artigos aceites para revisão[6] e apoiado num aparato documental de 139 artigos sobre as áreas de

[6] E tendo sido excluídos outros 460, por falharem em algum dos padrões — ensaios clínicos durando pelo menos oito semanas, por exemplo — exigidos.

reflexão relevantes, incluindo 12 especificamente sobre Meta-análise — tema que é objecto de exposição das páginas 18 a 24, e está explicitamente presente em todo o estudo —, e algumas dezenas sobre ensaios clínicos, produzem apenas as seguintes conclusões:

> "**Conclusions.** *Ephredine, ephredine plus caffeine, and ephedra-containing dietary supplements with or without herbs containing caffeine all promote modest amounts of weight loss over the short term. There are no data regarding long-term effects on weight loss. Single-dose ephedrine plus caffeine had a modest effect on athletic performance. The available trials do not provide any evidence about ephedrine or ephedra-containing dietary supplements, as they are used by the general population, to enhance athletic performance. Use of ephedra or ephedrine plus caffeine is associated with an increased risk of gastrointestinal, psychiatric, and autonomic symptoms. The adverse event reports contain a sufficient number of cases of death, myocardial infarction, cerebrovascular accident, seizure, or serious psychiatric illness in young adults to warrant a hypothesis-testing study, such as a case-control study, to support or refute the hypothesis that consumption of ephedra or ephedrine may be causally related to these adverse events.*"

2.5 Considerações Finais

Se não nos falha a memória, já Descartes afirmava que o bom senso foi a coisa que Deus melhor distribuiu: cada qual está contente com o quinhão que lhe coube em sorte.

Muito ou pouco nos tenha calhado em sorte, há que usá-lo porfiadamente, e como é natural também na análise estatística e interpretação dos resultados. Gilbert (1989), é um texto admirável, sempre a recordar que as ideias são mais importantes do que as fórmulas.

As ideias da meta-análise são estimulantes, e correspondem a uma evolução natural da Estatística. No passado, a recolecção acrítica de dados levou a muito disparate, sendo a correcção deles a invenção das metodologias de Amostragem: nem tudo o que vem à rede é peixe, só devemos analisar os dados representativos da população, obtidos com uma metodologia de recolha "ao acaso", antítese do "por acaso"ou "à toa".

Fisher, na primeira edição do *Statistical Methods for Research Workers*, de 1925, mudou o paradigma da Estatística, mostrando que o Planeamento de Experiências é uma sub-disciplina mais fértil do que a Amostragem: em lugar de observar passivamente, devemos *produzir* os dados que vão ser analisados, não só por forma a obter uma melhor relação custo/benefício mas também a poder estabelecer relações de causalidade, mais interessantes do que mera associação estatística que os estudos observacionais possibilitam. Um dos problemas dessa fase da Estatística era a escassez de dados, nomeadamente em áreas de Biomedicina, e foi necessário combinar inteligentemente resultados assintóticos com resultados exactos para pequenas amostras. O advento de recursos computacionais acessíveis permitiu o desenvolvimento da Estatística Computacional e de métodos de computação intensiva, em que do pouco se faz muito: poucos dados são usados para, construída a função de distribuição empírica, gerar muitos mais, e fazer estudos de simulação.

Mas o advento desses recursos computacionais trouxe novos problemas, porque permitiu também a aquisição de dados, aos milhões. Mais uma vez a Estatística se vê forçada a evoluir, agora na mineração do que nessa ganga enorme tem importância e pode servir para alicerçar decisões, investindo na nova disciplina de *Data Mining*.

Fisher desenvolveu os seus métodos na convicção de que o protocolo experimental nos defendia de metodologias sofisticadas e adversas que, na sua opinião,

eram mero ludismo de matemáticos. Em particular, apesar de ter apresentado a primeira solução credível da comparação de efeitos médios em situação de heterocedasticidade em populações gaussianas — o célebre problema de Behrens-Fisher —, na última edição por ele revista do *Statistical Methods for Research Workers*, p. 125, ainda escreve: "*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*".

Esta opinião drástica sobre a inutilidade dos métodos de estatística matemática em condições de heterocedasticidade é posteriormente moderada no *Statistical Methods and Scientific Inference*, p. 97. "*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.*" A importância actual de processos *ARCH* (*Auto-Regressive Conditional Heteroskedasticity*), e da meta-análise de ensaios clínicos e de testes de diagnóstico mostram a que ponto o futuro da Estatística se foi afastando da sua visão. A atribuição do Nobel de Economia de 2003 a Eagle e a Granger, em parte justificada pelo desenvolvimento dos processos ARCH, só vem sublinhar a necessidade de, cada vez mais, dar atenção à escala dos fenómenos, nomeadamente quando se comparam tratamentos que podem modificar a escala intrínseca do fenómeno no grupo de "casos", o que naturalmente não se espera que aconteça com os "controlos".

Com todas as potencialidades deste novo método meta-analítico de sintetizar conhecimento, usando agora fontes diversificadas em que a informação foi recolhida e relatada com protocolos diversos, há erros, limitações e cautelas que não devem ser escamoteados.

Logo à cabeça, a constatação de que as fontes documentais têm um viés positivo: na literatura científica não têm em geral acolhimento trabalhos inconclusivos, pelo que a evidência estatística, eventualmente pouco relevante caso a caso, aponta toda no mesmo sentido, e acaba por fabricar uma convicção forte que, porventura, é

apenas fruto de aparências.

Naturalmente os estudos inconclusivos raramente atingem o estatuto de publicáveis, pelo que um dos maiores problemas que os cultores da meta-análise enfrentam é o enviesamento decorrente de só encontrarem resultados que apontam na mesma direcção.

A situação pode ser mesmo muito preocupante, porque com os vultuosos investimentos em investigação o conceito de replicabilidade tem que ser repensado. De facto, um critério usado em muitas revistas científicas, particularmente na área biomédica, é a existência de valores de prova ($p-values$) que confiram um estatuto de grande improbabilidade à hipótese nula; ora a definição de valor de prova contém em si mesmo o reconhecimento de que valores piores do que o observado podem ser obtidos — com a probabilidade p, apenas — sendo H_0 verdadeira.

Concretizemos. Suponha-se, porque o ridículo ajuda, que alguns cientistas muito sérios decidiam reviver a teoria de que os espermatozóides produzidos pelo testículo direito originam rapazes, e os produzidos pelo testículo esquerdo originam raparigas, exposta — numa perspectiva de história das ideias, entenda-se — no notável livro de McManus (2002). Algumas centenas de equipas estavam a investigar "cientificamente"o assunto, na linha do que McManus relata que se passava efectivamente: Arranjavam, cada uma delas, a sua colecção de *voluntários*, e com os usuais cuidados de aleatorização se constituiam dois grupos; os indivíduos de um dos grupos iam com a instrução de que quando lhes apetecesse ... deveriam amarrar com toda a força o dito direito, enquanto os do outro grupo tinham a instrução de amarrar o dito esquerdo; ao fim de dois anos contabilizava-se o número de descendentes masculinos e femininos em cada um dos grupos[(7)].

[(7)] Os fundamentalistas desta teoria sobre a determinação do sexo clamariam evidentemente que os desvios factuais se deviam a alguns dos voluntários serem uns maricas que não se tinham amarrado com a determinação que a causa da Ciência merece — era este o consolo que restava, no relato de McManus, a algum indivíduo a quem saísse uma desilusão depois de tanto sacrifício, se quisesse continuar a acreditar na virtude da respectiva *com sorte* — escrevemos assim, porque se não tivesse sorte o marido optava pela outra explicação, a menos que fosse dos que acreditam que a alta percentagem de filhos extra-conjugais advém de trocas nas maternidades ...

Possivelmente alguns desses estudos originariam tabelas de contingência com forte evidência de que não se deveria considerar equiprovável gerar rapazes e raparigas consoante se fizesse um esforço para usar o testículo direito ou o testículo esquerdo. Seriam esses os estudos com alguma chance de serem publicados!

Relatos que são tão curiosos como esta investigação fictícia, que usamos como caricatura, ocorrem com frequência em secções (ou deveremos escrever sexões?) como o "Antes do Tempo", do Expresso Revista, ou na ampla divulgação dos prémios Ig Nobel (http://www.improbable.com/ig/ig-top.html; Abrahams, 2002, 2005)[8], ou mesmo em programas de televisão[9].

Voltando à seriedade e circunspecção de um artigo científico:

Ioannidis (2005) não se coíbe de, logo do título do seu artigo, denunciar que a maior parte dos trabalhos publicados têm conclusões falsas, e no corpo do artigo sugere que a metodologia da investigação científica está a ter um efeito perverso, confundindo enviesamento com significância.

---

[8] Se acharam a nossa ficção sobre esquerdo e direito forçada, leiam as páginas 198–201 de Abrahams (2005), que justificam a atribuição do prémio Ig Nobel de 1995 a Buebel, Shannahoff-Khalsa e Boyle (1991) por provarem, depois de um árduo trabalho de treino de 23 voluntários em inspirarem ar apenas através de uma das narinas, que o uso diferenciado das narinas na respiração altera as capacidades cognitivas: como os premiados relatam, "*Our results indicate that uninostril predominance is associated with varying ratios of cognitive performance and also that altering the phases of the nasal cycle by forced breathing in the nondominant nostril can influence cognitive performance*"..

[9] Não há muito tempo tivemos a ditosa sorte de observar três cientistas, que tiveram que actuar como seus próprios voluntários, mergulhando uma parte nobre da sua anatomia em água quente, a 45^{0}, no limiar da dor (? — os esgares iniciais levavam a crer que o limiar pode ter sido ultrapassado, mas quem investiga por gosto ...), para estabelecerem que ao fim de alguns dias desse tratamento o número de espermatozóides "vivos"declinava, até atingir níveis que poderiam garantir a esterilidade, sem incómodo para as parceiras, e sem ser um método irreversível como a vasectomia. Não podíamos, uma vez que vem a talhe de foice, deixar de registar esta recomendação aos muitos casais jovens ainda sem condições económicas para terem filhos, pois apesar de o preço da energia estar a subir acreditamos que aquecer água continua a ser o método mais barato de adiar a prole. Se não se pretender apenas adiar, o talhe da foice acima referido pode ser uma alternativa radicalmente barata à vasectomia.

Utts (1991) merece uma leitura cuidada, pelo esmero com que discute replicação em Ciência. No fim da leitura fica-se com a desconfortável sensação que que a objectividade científica é um mito. A evidência estatística a favor da existência de comparável percepção extrasensorial, cuidadosamente analisada por Utts, não demove os cépticos; no entanto — e é pena Utts não se ter lembrado de fazer essa comparação — é da mesma ordem de grandeza da evidência de que a administração de um placebo pode ajudar a curar, afirmação feita periodicamente em grandes revistas da especialidade, e que nunca vimos contestada.

O efeito de em Meta-análise não se estar a contabilizar equilibradamente a evidência inconclusiva ("*file-draw effect*") pode ser avaliado, por exemplo estimando o número de casos não significativos necessários para contrabalançar a evidência fornecida pelos valores de prova dos casos significativos publicados, como faz Utts (1991).

Simes (1986), advoga o registo de todas as investigações autorizadas pelas comissões éticas que se pronunciam sobre ensaios clínicos, por forma a estarem disponíveis para meta-análises fiáveis todos os estudos, quer apontem num quer noutro sentido. A discussão sobre a metodologia de selecção de estudos a incluir numa análise de síntese é, evidentemente, um tema recorrente, ainda que as recomendações contidas em Chalmers *et al.* (1981) continuem a ser largamente consensuais: partir de uma inclusão liberal de estudos, evitar o viés devido ao respeito por autores ou correntes de opinião (o que se faz ocultando fontes de informação e conclusões, e deixando apenas as descrições metodológicas e protocolos, que é o que nesta fase importa), e solicitar uma avaliação cega, no que respeita aqueles pontos, de pelo menos dois avaliadores independentes, que apenas se pronunciam sobre metodologias de investigação da área específica em estudo, protocolos experimentais e metodologias de análise estatística. Os estudos seleccionados nesta primeira fase são então disponibilizados na íntegra (fontes e conclusões), e pontuados no que refere validade e qualidade, com possível repescagem de estudos anteriormente sacrificados, seguindo-se uma reunião final dos avaliadores, para decisão definitiva sobre inclusões e exclusões.

O espaço de publicação em revista é escasso, e o estilo usual dos artigos científicos

despojado. Raramente há detalhes suficientes sobre o protocolo experimental, que poderiam ter alguma utilidade na harmonização que se busca nos estudos de síntese meta-analíticos. Seria porventura útil que as grandes revistas científicas passassem a disponibilizar em edição electrónica versões com "extras", como se usa nos filmes de sucesso em DVD, em que esses pormenores (que em estudos de síntese podem ser pormaiores) tivessem o seu lugar.

O paradigma bayesiano é, no contexto de meta-análise, natural ("[...] *Suppose that the meta-analyst is interested in some effect θ, common to all studies. After updating with observations from the first study, x_1, the posterior distribution is $f(\theta \mid x_1) \propto f(x_1 \mid \theta)\, f(\theta)$, where $f(\theta)$ is the prior distribution of the effect θ, and $f(x_1 \mid \theta)$ is the likelihood of the data given the effect. Now the meta-analyst can use $f(\theta \mid x_1)$ as a prior for the next analysis, producing $f(\theta \mid x_1, x_2) \propto f(x_2 \mid \theta) f(\theta \mid x_1)$ and so on. This updating scheme is common in all applications of Bayesian methods and demonstrates how all Bayesian analysis can be seen as meta-analysis* — escrevem Stangl and Berry, 2000, p. 7). As razões aduzidas para as escolhas de distribuições *a priori*, no entanto, nem sempre ultrapassam, afinal, a conveniência matemática: porquê modelar com uma t a variabilidade dos valores médios populacionais entre estudos no modelo completo apresentado por Wakefield and Rahman (2000), quando há modelos de caudas pesadas mais convincentes (leis estáveis)? Decerto apenas por melhor tratabilidade matemática!

A controvérsia meta-análise de ensaios clínicos *versus* estudos experimentais de grande dimensão é até certo ponto um falso problema, uma vez que em última instância estudos de meta-análise virão a integrar qualquer estudo relevante anterior e assim, naturalmente, as descobertas feitas com ensaios clínicos de grande dimensão. De qualquer modo, Mosteller and Chalmers (1992) referem que nas quatro situações em que houve estudos de síntese e ensaios clínicos de grande dimensão não integrados nessas sínteses, as conclusões são idênticas.

Mais grave é a suspeita de que um único dos estudos integrados na meta-análise domine as conclusões (Thompson and Pocock, 1991).

A meta-análise foi inventada para tornear o problema de muitos ensaios clínicos

levarem a conclusões baseadas em evidência fraca. Criou-se a esperança de, com uma síntese apropriada, evidenciar efeitos que cada um deles isoladamente não tinha conseguido alicerçar. No entanto, a nosso ver, é neste ponto que reside toda a fragilidade desta nova disciplina estatística, nascida para corrigir a fraca evidência fornecida por estudos de dimensão modesta. Por exemplo, o valor observado da estatística do qui-quadrado na tabela de contingência

15	19	34
23	55	78
38	74	112

é $X^2_1(obs.) = 2.261$, não significativo ao nível $\alpha = 0.05$. Duplicando as frequências

30	38	68
46	110	156
76	148	224

a estatística de teste tem um valor observado duplo, $\tilde{X}^2_1(obs.) = 4.522$, que é significativo àquele nível, apesar de a razão de vantagens (*odds ratio*) ser a mesma! Mais geralmente, se os valores observados a, b, c, d de uma tabela de contingência 2×2 forem multiplicados por um escalar λ o valor da estatística de teste

$$\tilde{X}^2_{2,2}(obs.) = \frac{\lambda n(\lambda a\, \lambda d - \lambda b\, \lambda c)^2}{(\lambda a + \lambda b)(\lambda a + \lambda c)(\lambda b + \lambda d)(\lambda c + \lambda d)} = \lambda X^2_{2,2}(obs.)$$

também vem multiplicado por λ.

O mesmo se passa com testes paramétricos, nomeadamente os testes clássicos para comparação de efeitos médios de dois tratamentos, na situação clássica de populações gaussianas. Há regras para determinar a dimensão amostral necessária para estimar parâmetros com a precisão desejada; a atitude displicente de recolher amostras de grande dimensão, porque assim "decerto proporcionam a precisão desejada"pode ser desastrosa: o erro padrão $\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}$ pode tornar-se tão pequeno que mesmo diferenças irrelevantes entre as médias podem aparecer como significativas.

Por outras palavras, o excesso de informação é uma verdadeira vertigem das pequenas diferenças: mesmo diferenças mínimas se tornam irremediavelmente sig-

nificativas, sem que isso tenha qualquer tipo de relevância. A Amostragem há muito tempo que desenvolveu a teoria de qual a dimensão amostral que se deve usar para obter uma precisão (*error bound*) pré-fixada. Enquanto em ensaios clínicos se embarcar em estudos de grande dimensão ou meta-análise de estudos, sem um controlo prévio do que é razoável observar para que a detecção de efeitos seja considerada relevante, parece-nos mais um retrocesso do que um progresso.

Capítulo 3

Harmonização de testes independentes

Suponha-se a investigação de uma hipótese nula H_0 contra uma alternativa H_A usando a estatística de teste T, de que se conhece a distribuição de $T|_{H_0\ verd.}$, há um historial de publicação de resultados reportando níveis de significância descritivos — valores de prova, *p–values* — $p_k = \mathbb{P}\left[T|_{H_0\,verd.} > T_k(obs.)\right]$, $k = 1, \ldots, n$. Como harmonizar esses diversos resultados?

Pelo teorema da transformação uniformizante, e de $X \frown Beta(p,q) \Longrightarrow 1 - X \frown Beta(q,p)$ — donde, em particular, $X \frown Uniforme(0,1) \Longrightarrow 1 - X \frown Uniforme(0,1)$ — conclui-se que $\{p_1, \ldots, p_n\}$ é uma amostra de dimensão n de uma população uniforme padrão.

É também sobejamente conhecido que $X \frown Uniforme(0,1) \Longrightarrow -\ln X \frown Exponencial(1) \Longrightarrow -2\ln X \frown \chi^2_{(2)}$.

Por outro lado, ainda pelo teorema das transformação uniformizante, denotando Φ a função de distribuição da variável aleatória gaussiana padrão, o conjunto dos $z - scores$ $\{\Phi^{-1}(1-p_1), \ldots, \Phi^{-1}(1-p_n)\}$ é uma amostra de dimensão n de uma população gaussiana padrão (e, obviamente, usando as variáveis antitéticas, o mesmo se pode dizer de $\{\Phi^{-1}(p_1), \ldots, \Phi^{-1}(p_n)\}$, o que contudo não trará grandes vantagens, dada a estrita dependência entre as duas amostras aleatórias de que são

realizações). Esta metodologia de usar os *scores* gaussianos $z_k := \Phi^{-1}(p_k)$, em lugar dos níveis de significância descritivos p obtidos para construir um juízo global sobre a hipótese nula H_0, foi proposta por Stouffer *et al.* (1949). Esta forma de proceder á enfaticamente recomendada por Mosteller and Bush (1954).

Temos assim um contexto que nos permite usar um alargado leque de resultados sobre caracterizações de uniformes, de gaussianas e de quis-quadrado (e mais especialmente de exponenciais), suas estatísticas ordinais, espaçamentos, e somas.

Os trabalhos pioneiros são de Tippett (1931), que advoga rejeição da hipótese nula, ao nível de significância α, no teste combinado se $p_{1:n} < 1 - (1-\alpha)^{\frac{1}{k}}$, e de Fisher (1932), que na quarta edição do seu famoso tratado *Statistical Methods for Research Workers* propõe o uso do teste combinado

$$T = -2\sum_{k=1}^{n} \ln P_k \underset{|_{H_0\ verd.}}{\frown} \chi^2_{(2n)}.$$

Note-se que na proposta de Tippett a região de rejeição é a união das regiões de rejeição de cada um dos testes. Note-se que $P_{(1:n)} \underset{|_{H_0\ verd.}}{\frown} Beta(1,n)$. Mais geralmente, $P_{(k:n)} \underset{|_{H_0\ verd.}}{\frown} Beta(k, n+1-k)$, observação que foi usada por Wilkinson (1951): rejeitar H_0 ao nível de significância $\alpha = \mathbb{P}[\text{rejeitar } H_0 | H_0 \text{verdadeira}]$ se $p_{k:n} < c$, onde

$$\int_0^c \frac{t^{k-1}(1-t)^{n-k}}{B(k, n+1-k)}\,\mathrm{d}t = \alpha.$$

Assim o facto de os níveis de significância descritiva p de testes independentes usando a estatística T para decidir sobre uma hipótese nula H_0 constituírem, quando a hipótese nula é verdadeira, uma amostra da população uniforme padrão, tem uma relevância indiscutível, e é o ponto de partida que nos permite recorrer à teoria das estatísticas ordinais e espaçamentos da uniforme, da exponencial e da normal.

A utilidade de sínteses meta-analíticas em Medicina, nomeadamente quando os estudos disponíveis usaram um número reduzido de casos, está bem estabelecida, basta consultar o conjunto de artigos de Davey Smith and Egger (1998), Egger and

Davey Smith (1997, 1997a, 1997b, 1998) e Egger *et al.* (1998) "seleccionados" pelo editor do *British Medical Journal*.

A construção da Ciência é um empreendimento colectivo. Em ciências da saúde, em particular, estudos de dimensão modesta podem ser em si mesmos um quase inaudível murmúrio, mas quando muitos desses estudos baseados em informação escassa são combinados podem gritar bem alto resultados importantes. A publicação atempada de revisões sistemáticas — de resultados publicados mas também, se possível, de resultados que não passaram de relatórios não publicados — é a base de trabalho necessária para avaliação de parâmetros importantes, tais como a prevalência ou a taxa de contágio de doenças, sensibilidade e especificidade de análises clínicas e os consequentes valores preditivos positivo e negativo, ou o ponto de corte adequado para optimizar, em termos de uma função de perda adequada, os valores preditivos de um teste clínico.

As monografias recentes de Hartung *et al.* (2008), de Kulinskaya *et al.* (2008), e de Borenstein *et al.* (2009) apresentam panorâmicas interessantes (e bem diversas!) de questões de meta-análise. Outros textos introdutórios, mais breves e simples, que nos foram úteis a desbravar este tema foram o capítulo 12 de Woodward (2005), o capítulo 16 de Senn (2007), e o capítulo 12 de Longford (2008). Glass(1999), que cunhou o termo meta-análise, é uma apresentação cheia de sageza, não só sobre meta-análise como sobre os caminhos da Ciência em geral.

Na secção 1 apresentamos uma caracterização da uniforme, alternativa à de Paul (2003), recorrendo a propriedades das estatísticas uniformes de populações uniformes. Na secção 2 desenvolvemos as propriedades que serão o ponto de partida de geração de amostras uniformes, com a finalidade de aumentar computacionalmente uma amostra de níveis de significância descritivos $\{p_{_k}\}$ por forma a dispor de amostras de maior dimensão de "pseudo-níveis de significância descritivos que nos permitam testar a validade global de uma hipótese nula H_0 quando com uma síntese meta-analítica pretendermos harmonizar os resultados de vários testes independentes sobre essa hipótese (usando uma mesma estatística de teste T, obviamente). Assim, avançamos no problema levantado por Paul (2003): como testar um ajustamento uniforme (nomeadamente quando se dispõe de um número escasso

de observações)?

Na secção 3 discutimos, de forma muito breve, questões que resultam do enviesamento na publicação, nomeadamente a questão de a amostra de valores$-p$ ser censurada à direita. Na secção 4 uma simulação elementar mostra o efeito deste aumento artificioso da dimensão da amostra na decisão de manter ou rejeitar a hipótese nula combinando a evidência obtida em testes independentes.

3.1 Estatísticas ordinais, espaçamentos e caracterizações das populações uniformes.

Se a normal domina a parte de leão da Estatística que se ocupa de resultados assintóticos, a uniforme é incontornável em todas as áreas da Estatística Computacional, e nomeadamente distingue-se pelo seu papel preponderante na geração de números aleatórios. Johnson *et al.* (1995) descrevem muitas caracterizações úteis da uniforme e de variáveis aleatórias com ela relacionadas.

Adiante teremos que fazer uso de algumas propriedades das estatísticas ordinais. Nomeadamente, seja $(X_1, \ldots, X_n)$ uma amostra aleatória de dimensão n de uma população $X \frown Uniform(0,1)$, $(X_{1:n}, \ldots, X_{n:n})$ o correspondente vector de estatísticas ordinais (ascendentes), e defina-se $X_{0:n} := \alpha_{F_X} = 0$ e $X_{n+1:n} := \omega_{F_X} = 1$, onde α_{F_X} e ω_{F_X} são os pontos terminais esquerdo e direito, respectivamente, do suporte de X. O conjunto das estatísticas ordinais divide $(0,1)$ em $n+1$ espaçamentos $S_k := X_{k:n} - X_{k-1:n}$.

A função densidade de probabilidade conjunta do vector das estatísticas ordinais $(X_{1:n}, \ldots, X_{n:n})$ é

$$f_{(X_{1:n},\ldots,X_{n:n})}(x_1, \ldots, x_n) = n!\, I_{\{0<x_1<\cdots<x_n<1\}}$$

e consequentemente a função densidade de probabilidade do vector das estatísticas ordinais dos correspondentes espaçamentos $(S_{1:n+1}, \ldots, S_{n:n+1})$ is, in the support $A = \{(s_1, \ldots, s_n)\}$ com $s_k \in (0,1)$ tais que $\sum_{k=1}^{n} s_k < 1$

$$f_{(S_{1:n+1},\ldots,S_{n:n+1})}(s_1, \ldots, s_n) = n!\,(n+1)!\,I_A.$$

Mas então, considerando as variáveis aleatórias $W_j := (n+2-j)\,(S_{j:n+1} - S_{j-1:n+1})$, $j = 2, \ldots, n+1$ (uma transformação introduzida por Sukhatme (1937)),

$$f_{(W_1,\ldots,W_n)}(s_1, \ldots, s_n) = n!\,I_A,$$

e segue-se que os vectores $(X_{1:n}, \ldots, X_{n:n})$ e $(W_1, W_1 + W_2, \ldots, W_1 + \cdots + W_n)$ são equidistribuídos.

A função de distribuição $k-$ésima estatística ordinal é $F_{X_{k:n}}(x) = \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j}$ (para $x \in (0,1)$), e da expressão geral $\mathbb{E}(X) = \int_0^\infty [1 - F_X(x)]\,\mathrm{d}x$ que, para variáveis aleatórias positivas, reescreve o valor médio como integral de Riemann da cauda direita da distribuição, no caso de $\mathbb{E}(X)$ existir, obtém-se $\mathbb{E}(S_k) = \binom{n}{k} \int_0^1 x^{k-1}(1-x)^{n+1-k}\,\mathrm{d}x$. David and Nagaraja (2003) é uma boa fonte de informação geral sobre estatísticas ordinais e funções de estatísticas ordinais, consulte-se em particular a secção 5.4 no que respeita à divisão aleatória de um intervalo.

Paul (2003) avançou a caracterização: $X \frown Uniforme(a,b)$ se e só se $\mathbb{E}(S_{k:n+1}) = \dfrac{b-a}{n+1}$, $k = 1, \ldots, n$. No que se segue vamos demonstrar um resultado mais forte:

Teorema 3.1.1.

$X \frown Uniforme(a,b)$ se e só se os espaçamentos determinados por uma amostra aleatória $(X_1, \ldots, X_n)$ forem equidistribuídos.

De facto, atendendo à expressão da função densidade de probabilidade $f_{U_{k-1:n},U_{k:n}}$ de duas estatísticas ordinais consecutivas da população $U \frown Uniform(0,1)$[1]

$$f_{U_{k-1:n},U_{k:n}}(x,y) = \frac{n!\,x^{k-2}(1-y)^{n-k-1}}{(k-2)!\,(n-k-1)!}\,I_{\{0<x<y<1\}},$$

imediatamente se estabelece que a função densidade de probabilidade do $k-$ésimo espaçamento originado por uma amostra aleatória de dimensão n,

$$f_{S_k}(z) = \frac{n!}{(k-2)!(n-k-2)!}\int_0^{1-z} x^{k-2}(1-z-x)^{n-k-1}\mathrm{d}x = n(1-z)^{n-1}\,I_{(0,1)}(z),$$

que é a função densidade de probabilidade de uma variável aleatória $Beta(1,n)$, qualquer que seja $k = 1,\ldots,n$.

Por outro lado, se os espaçamentos correspondentes a uma variável aleatória X com suporte em (0,1) forem equidistribuídos, de $\sum_{k=1}^{n+1} S_k = 1$ segue-se, para $n = 1,2,\ldots,$ que $\mathbb{E}(S_k) = \dfrac{1}{n+1}$, e em particular $\mathbb{E}(S_{n+1}) = \int_0^1 F^n(x)\,\mathrm{d}x = \dfrac{1}{n+1} = \mathbb{E}(U^n)$, $n = 1,2,\ldots;$ do critério de Carleman é então imediato que $X \frown Uniform(0,1)$.

3.2 Descendentes uniformes de parentes uniformes

Sejam X e Y variáveis $Uniforme(0,1)$, independentes. A variável aleatória

[1] Como as transformações lineares não alteram a forma das estatísticas ordinais e dos espaçamentos, sem perda de generalidade trabalhamos com (a,b)=(0,1).

$W = \min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right)$ tem suporte $(0,1)$, e para $z \in (0,1)$

$$\mathbb{P}\left[\min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \leq z\right] =$$

$$\int_0^1 zy \, \mathrm{d}y + \int_0^1 z(1-y) \, \mathrm{d}y = z.$$

Mais: para quaiquer $y, z \in (0,1)$,

$$\mathbb{P}\left[\min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \leq z | \, Y = y\right] =$$

$$yz + (1-y)z = \mathbb{P}\left[\min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \leq z\right],$$

e consequentemente $W \frown Uniform(0,1)$, sendo Y e W independentes.

Analogamente, a variável aleatória $V = X + Y - \mathcal{I}[X+Y]$, onde $\mathcal{I}[X+Y]$ é o maior inteiro que não excede $X+Y$ (no caso de argumento $X+Y \geq 0$, podemos dizer simplesmente que é a parte inteira de $X+Y$), é $Uniform(0,1)$. De facto, o seu suporte é $(0,1)$, e para $z \in (0,1)$

$$\mathbb{P}[X + Y - \mathcal{I}[X+Y] \leq z] =$$

$$\mathbb{P}[X + Y \leq z] + \mathbb{P}[1 < X + Y \leq 1 + z] = z.$$

Também neste caso, para quaisquer $y, z \in (0,1)$,

$$\mathbb{P}[X + Y - \mathcal{I}[X+Y] \leq z | Y = y] =$$

$$\max(0, z-y) + \min(1+z-y, 1) - (1-y) = z$$

e neste caso podemos permutar os papéis de X e Y; portanto X, Y e V são independentes.

Johnson *et al.* (1995) creditam a Deng and George (1992) resultados similares com Y contínua com suporte em $(0,1)$. Apenas tivemos acesso ao resumo, que nos

pareceu ter um fôlego mais limitado, mas a questão será aprofundada mais adiante.

Vamos discutir, na secção 4, como usar aqueles resultados para aumentar computacionalmente uma amostra de níveis de significância descritivos $(p_1, \ldots, p_n)$, e investigar se o uso dessses "pseudo-valores$-p$" nos permite meta-analisar os resultados com potência acrescida.

3.3 Uma chamada de atenção para o possível efeito do enviesamento na publicação de resultados

O enviesamento na publicação de resultados é uma questão relevante; transcrevemos de uma célebre conferência de Glass (1999):

"*It's one thing to believe that peer review guarantees truth; it is quite another to believe that all truth appears in peer reviewed journals. (The most important paper on the multiple comparisons problem in ANOVA was distributed as an unpublished ditto manuscript from the Princeton University Mathematics Department by John Tukey; it never was published in a peer reviewed journal).*"

É pois um ponto importante a ter em conta, e que de modo algum pode ser escamoteado: o conjunto de níveis de significância descritiva p disponíveis, resultando de testes independentes, é em teoria, sob validade de H_0, uma amostra da população $Uniform(0, 1)$; mas, de facto, em geral apenas conseguiram passar o apertado crivo de publicação os trabalhos em que o nível de significância descritiva p permitia reclamar "significância estatística". Por isso o que se tem, e isto nos melhores casos[(2)]

[(2)] De facto, e sobretudo nos trabalhos mais antigos, os níveis de significância descritivos ou não

é uma amostra uniforme censurada à direita. No caso de ser credível que apenas os casos em que $p < 0.05$ foram acolhidos em publicações científicas, o remédio parece simples: basta re-escalar os níveis de significância descritivos p *antes* de usar os procedimentos que adiante descrevemos para aumentar computacionalmente a amostra com "pseudo$-p$s".

Por outro lado, quando observamos que foram retidos para a síntese meta-analítica trabalhos em que estão reportados níveis de significância descritiva p superiores a 0.05, $\frac{n+1}{n} P_{n:n}$ é um estimador centrado do "*right-endpoint*" $\omega_{_F} := \sup\{x : \ F(x) < 1\}$, que limita à direita o suporte da população censurada, e podemos usar a variável $\frac{n+1}{n} P_{n:n}$ para fazer um re-escalamento aleatório da amostra disponível.

Note-se no entanto que se $P \frown Uniform(0, a)$ com a estimado por $\frac{n+1}{n} p_{n:n}$, $\{\frac{n\, p_{1:n}}{(n+1)\, p_{n:n}}, \ldots, \frac{n\, p_n}{(n+1)\, p_{n:n}}\}$ já não é uma amostra de uma população uniforme padrão (ou sequer uniforme).

3.4 Um breve estudo de simulação

Como o enviesamento na publicação favorece a disseminação de resultados com níveis de significância descritiva p inferiores a 0.05, pareceu-nos interessante escolher uma alternativa à hipótese nula de os $\{p_{_k}\}$ observados provirem de uma população $U \frown Uniforme(0, 1)$ que favorecesse valores$-p$ baixos. Decidimos por isso estudar alternativas$H_{_{A,m}}$, com $m \in [-2, 0)$, de a população de que se observou a amostra $\{p_{_k}\}$ ter uma função densidade de probabilidade

$$f_{X_m}(x) = \left(mx + 1 - \frac{m}{2}\right) \, I_{(0,1)}(x),$$

estão registados ("$p < 0.05$") ou, para valores muito baixos, estão registados com aproximações muitas vezes grosseiras ("$p < 0.0001$"pode significar $p = 0.23 \times 10^{-17}$. Isto não falar da situação em que se usa a convenção de usar um, dois ou três asteriscos para denotar $p < 0.05$, $p < 0.01$ e $p < 0.001$ respectivamente, por exemplo.

A função de distribuição pode ser explicitamente invertida

$$F_{X_m}^{-1}(y) = \frac{\frac{m}{2} - 1 + \sqrt{\left(\frac{m}{2} - 1\right)^2 + 2my}}{m}$$

e consequentemente a geração de números pseudo-aleatórios de X_m é simples e imediata.

Note-se que H_0 corresponde a $m = 0$.

Fizemos então o breve estudo de simulação que adiante se descreve:

1. Para $m = -2.0, -1.95, \ldots, -0.05, 0$, gerámos uma amostra de dimensão n, $p_1, \ldots, p_n$, da população X_m.

   • Se $p_{1:n} < 1 - (1 - \alpha^{\frac{1}{n}})$, rejeita-se a hipótese nula de uniformidade da população ao nível de significância α (no que se segue, apenas apresentamos resultados para $\alpha = 0.05$); denotamos $\pi_1 = \pi_1(\alpha)$ a proporção de rejeições em N *runs* usando este critério.

   • Se $-2\sum_{k=1}^{n} \ln p_k > \chi^2_{2n;\,1-\alpha}$, rejeita-se a hipótese nula de uniformidade da população ao nível de significância α ; denotamos π_2 a proporção de rejeições em N *runs* usando este critério.

2. Usando o conjunto original $\{p_k\}_{k=1}^{n}$, e pseudo-aleatórios uniformes u_k, obtivemos um novo conjunto de "pseudo$-p$'s

   $$p_{n+k} = p_k + u_k - \mathcal{I}\left[p_k + u_k\right],$$

   $k = 1, \ldots, n$.

   • Se $p_{1:2n} < 1 - (1 - \alpha^{\frac{1}{2n}})$, rejeita-se a hipótese nula de uniformidade da população ao nível de significância α; denotamos $\pi_3 = \pi_1(\alpha)$ a proporção de rejeições em N *runs* usando este critério.

• Se $-2\sum_{k=1}^{2n}\ln p_k > \chi^2_{4n;\,1-\alpha}$, rejeita-se a hipótese nula de uniformidade da população ao nível de significância α ; denotamos π_4 a proporção de rejeições em N *runs* usando este critério.

3. Usando os dois conjuntos anteriores, calculámos um conjunto terciário de "pseudo$-p$"s

$$p_{2n+k} = \min\left(\frac{p_{n+k}}{p_k}, \frac{1-p_{n+k}}{1-p_k}\right),$$

$k = 1, \ldots, n$.

• Se $p_{1:3n} < 1-(1-\alpha^{\frac{1}{3n}})$, rejeita-se a hipótese nula de uniformidade da população ao nível de significância α; denotamos $\pi_5 = \pi_1(\alpha)$ a proporção de rejeições em N *runs* usando este critério.

• Se $-2\sum_{k=1}^{3n}\ln p_k > \chi^2_{6n;\,1-\alpha}$, rejeita-se a hipótese nula de uniformidade da população ao nível de significância α ; denotamos π_6 a proporção de rejeições em N *runs* usando este critério.

As recomendações de van Belle (2008), baseadas na constatação de que com amostras de dimensão muito baixa há instabilidades fortes nas avaliações por simulação, levaram-nos a optar, para um estudo inicial, por amostras de dimensão $n_1 = 4$, a fim de as amostras computacionalmente aumentadas terem dimensões $2n_1 = 8$ e $3n_1 = 12$.

Na Figura 1 representam-se as fracções de rejeições $\pi_1(0.05), \ldots, \pi_6(0.05)$ como função do declive m, para amostras de dimensões $n_1 = 4$, $2n_1 = 8$ and $3n_1 = 12$, respectivamente. ($\alpha = 0.05$, $N = 10\,000$), alisando os resultados obtidos nos pontos $m = -2, -1.95, \ldots, -0.05, 0$.

Observa-se que há um decréscimo de potência com o aumento do número de pseudo$-p$'s, um resultado que nos levou, necessariamente, a aprofundar esta investigação.

Como de qualquer modo podia ser meramente um efeito espúrio, levámos a cabo um estudo de simulação análogo para uma amostra de base de maior dimensão. Na

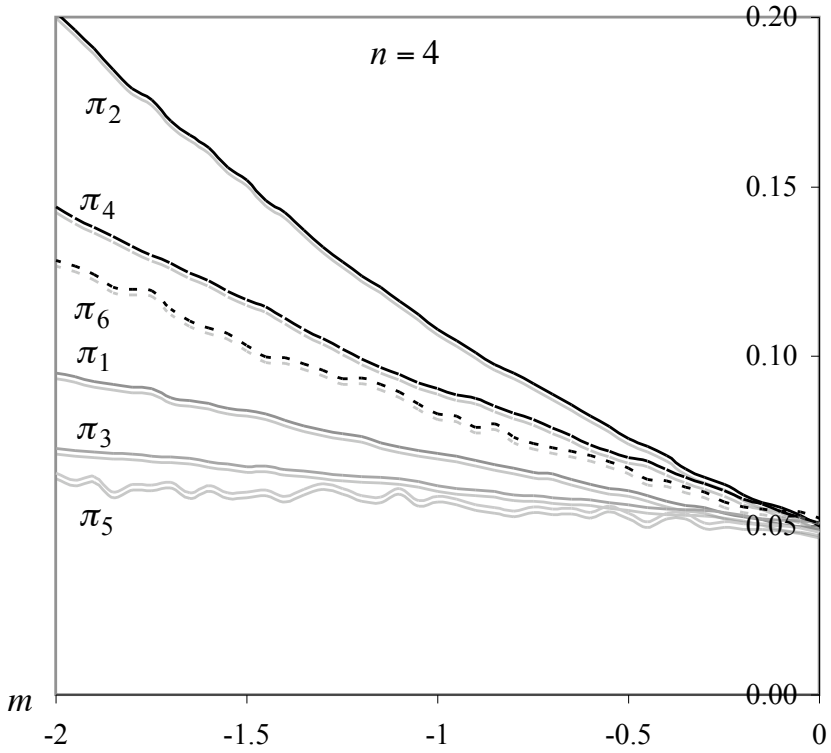

Figura 3.1: Proporções de rejeições em função de $m \in [-2, 0]$, $n = 4$.

Figura 2 mostra-se graficamente o resultado da simulação para $n = 30$, que afinal confirma o resultado inesperado observado.

Descendentes uniformes de parentes não uniformes

Poderia esperar-se que o uso de X_m com função densidade de probabilidade $f_{X_m}(x) = \left(mx + 1 - \frac{m}{2}\right) I_{(0,1)}(x)$,, em vez de $U \frown Uniforme(0, 1)$ na definição das variáveis aleatórias $W_m = \min\left(\frac{U}{X_m}, \frac{1-U}{1-X_m}\right)$ e $V_m = U + X_m - \mathcal{I}[U + X_m]$, revelaria a não-uniformidade de X_m. De facto tal não é o caso, naquelas definições U "massaja" suficientemente X_m, a ponto de o resultado ser uniforme. Neste aspecto, U exerce uma "atracção" quase tão forte para as operações descritas quanto a normal para a soma.

Vamos ver que tal é o caso quando a "outra" variável tem suporte em (0,1):

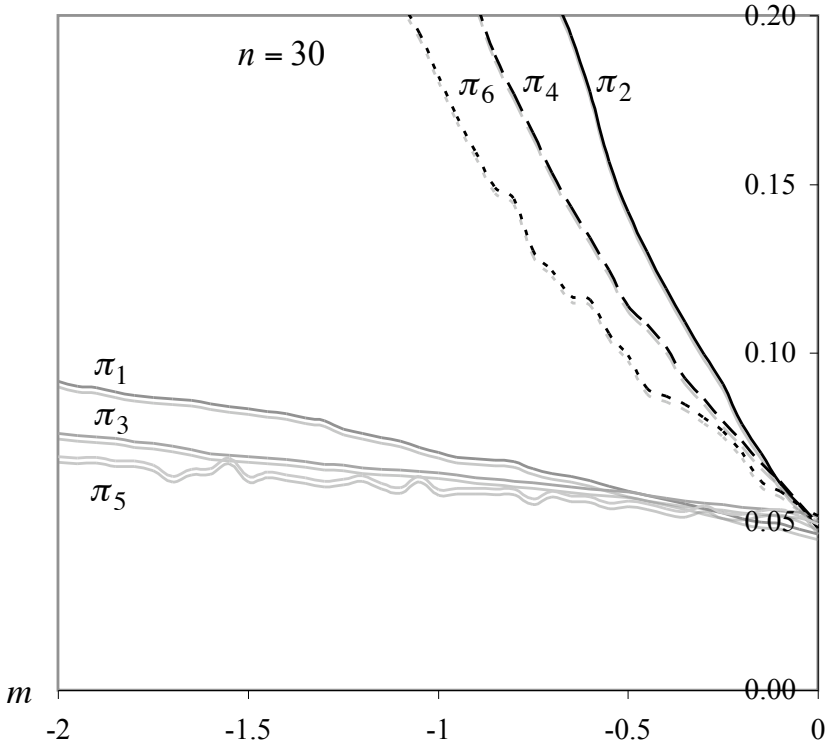

Figura 3.2: Proporções de rejeições em função de $m \in [-2, 0]$, $n = 30$.

Teorema 3.4.1. *Sejam X e U variáveis aleatórias independentes, X absolutamente contínua com suporte em (0,1) e $U \frown Uniforme(0,1)$. Então $W^* = \min\left(\frac{U}{X}, \frac{1-U}{1-X}\right) \frown Uniforme(0,1)$ e $V^* = X + U - \mathcal{I}[X+U] \frown Uniforme(0,1)$.*

Demonstração:

Consideremos X uma variável aleatória absolutamente contínua com suporte em $(0,1)$ e $U \frown Uniforme(0,1)$, independentes uma da outra.

É fácil verificar que $W^* = \min\left(\frac{U}{X}, \frac{1-U}{1-X}\right) \frown Uniforme(0,1)$. É imediato que o suporte desta variável é $\boldsymbol{S} = (0,1)$; para qualquer $w \in (0,1)$

$$F_{W^*}(w) = \mathbb{P}\left[U \leq X, U \leq \left(\frac{X}{w} + \frac{w-1}{w}\right)\right] + \mathbb{P}\left[U > X, U \geq \frac{X}{w}\right] =$$

$$= \int_0^w f_U(u) \left[\int_{\frac{u}{w}}^1 f_X(x)\ \mathrm{d}x \right]\ \mathrm{d}u + \int_{1-w}^1 f_U(u) \left[\int_0^{\frac{u}{w}+\frac{w-1}{w}} f_X(x)\ \mathrm{d}x \right]\ \mathrm{d}u =$$

$$= \int_0^w \left[1 - F_X\left(\frac{u}{w}\right) \right]\ \mathrm{d}u + \int_{1-w}^1 \left[F_X\left(\frac{u}{w} + \frac{w-1}{w}\right) \right]\ \mathrm{d}u$$

Efectuando a substituição $z = u + w + 1$ no segundo integral, obtem-se:

$$F_{W^*}(w) = \int_0^w \left[1 - F_X\left(\frac{u}{w}\right) \right]\ \mathrm{d}u + \int_{1-w}^1 f_U(u) \left[\int_0^{\frac{u}{w}+\frac{w-1}{w}} f_X(x)dx \right]\ \mathrm{d}u =$$

$$= \int_0^w \left[1 - F_X\left(\frac{u}{w}\right) \right]\ \mathrm{d}u + \int_0^w \left[F_X\left(\frac{z-(w-1)}{w} + \frac{w-1}{w}\right) \right]\ \mathrm{d}z = w$$

ficando assim estabelecido que $W^* \frown Uniforme(0,1)$.

No que se refere a $V^* = X + U - \mathcal{I}[X+U]$, é também imediato que o suporte de V^* é $\boldsymbol{S} = (0,1)$; para qualquer $v \in (0,1)$

$$\begin{aligned}
F_{V^*}(v) &= \mathbb{P}\left[U \leq -X+1, U \leq -X+v\right] + \mathbb{P}\left[U > -X+1, U \leq -X+v+1\right] = \\
&= \int_0^v f_X(x) \left[\int_0^{-x+v} \mathrm{d}u \right] \mathrm{d}x + \int_0^v f_X(x) \left[\int_{-x+1}^1 \mathrm{d}u \right] \mathrm{d}x + \int_v^1 f_X(x) \left[\int_{-x+1}^{-x+v+1} \mathrm{d}u \right] \mathrm{d}x = \\
&= \int_0^v \left[f_X(x)(-x+v) \right] \mathrm{d}x + \int_0^v \left[f_X(x)x \right] \mathrm{d}x + \int_v^1 \left[f_X(x)v \right] \mathrm{d}x = \\
&= -\int_0^v \left[f_X(x)x \right] \mathrm{d}x + \int_0^v \left[f_X(x)x \right] \mathrm{d}x + v[\int_v^1 \left[f_X(x) \right] \mathrm{d}x + \int_v^1 \left[f_X(x) \right] \mathrm{d}x = v
\end{aligned}$$

e portanto $V^* \frown Uniforme(0,1)$.

No final do capítulo estuda-se $Z = \frac{1}{W^*}$, e comenta-se no contexto dos testes de uniformidade usando amostras aumentadas.

3.5 Complementos sobre álgebra das variáveis aleatórias

Pretende-se determinar a distribuição da variável aleatória $W = \min\left(\frac{Y}{X}, \frac{1-Y}{1-X}\right)$ no caso de X e Y terem distribuição Beta e serem independentes. Trata-se de generalizar o resultado conhecido no caso de X e Y serem v.a.´s i.i.d. Uniformes standard. Neste caso, sabe-se que W também tem distribuição Uniforme standard e é independente de X e de Y.

Consideremos então X e Y v.a.´s independentes com distribuição Beta(p_1, q_1) e Beta(p_2, q_2) respectivamente. Seja $w \in [0, 1]$,

$$\begin{aligned} F_W(w) &= \mathbb{P}(Y \leq X, Y \leq wX) + \mathbb{P}(Y > X, Y \geq wX + 1 - w) \\ &= \int_0^1 f_X(x) \left[\int_0^{wx} f_Y(y)\mathrm{d}y\right] \mathrm{d}x + \int_0^1 f_X(x) \left[\int_{wx+1-w}^1 f_Y(y)\mathrm{d}y\right] \mathrm{d}x \end{aligned}$$

Se considerarmos no primeiro integral a substituição

$$\begin{cases} z = x \\ v = \dfrac{y}{wx} \end{cases}$$

e no segundo integral a substituição

$$\begin{cases} z = x \\ v = \dfrac{y - (wx + 1 - w)}{1 - (wx + 1 - w)} \end{cases}$$

teremos que, para $w \in [0,1]$,

$$F_W(w) = w\int_0^1 f_X(z)z\left[\int_0^1 f_Y(wzv)\,\mathrm{d}v\right]\mathrm{d}z + \int_0^1 f_X(z)[w(1-z)]\left[\int_0^1 f_Y(w(1-z)(v-1)+1)\,\mathrm{d}v\right]\mathrm{d}z$$

$$= \frac{w}{B(p_1,q_1)\,B(p_2,q_2)}\int_0^1 z^{p_1}(1-z)^{q_1-1}\left[\int_0^1 (wzv)^{p_2-1}(1-wzv)^{q_2-1}\mathrm{d}v\right]\mathrm{d}z +$$

$$+\frac{1}{B(p_1,q_1)\,B(p_2,q_2)}\int_0^1 wz^{p_1-1}(1-z)^{q_1}\left[\int_0^1 (1-w(1-z)(1-v)^{p_2-1}(w(1-z)(1-v)^{q_2-1}\mathrm{d}v\right]\mathrm{d}z =$$

$$= \frac{w^{p_2}}{B(p_1,q_1)\,B(p_2,q_2)}\sum_{k=0}^{q_2-1}\frac{B(p_1+p_2+k,q_1)\binom{q_2-1}{k}(-w)^k}{k+p_2} +$$

$$+\frac{w^{q_2}}{B(p_1,q_1)\,B(p_2,q_2)}\sum_{k=0}^{p_2-1}\frac{B(p_1,q_1+q_2+k)\binom{p_2-1}{k}(-w)^k}{k+q_2}$$

Assim, a função de distribuição é uma função polinomial de grau dependente dos parâmetros da Beta que surge em numerador, dependendo deles apenas através da sua soma. Os parâmetros da Beta em denominador apenas contribuem para a determinação dos coeficientes do polinómio. Se fizermos todos os parâmetros unitários é um simples exercício verificar que se obtém a função de distribuição da uniforme padrão.

Outro ponto a realçar é que a permutação dos parâmetros de qualquer uma das Betas não altera a distribuição de W.

No caso particular das Betas serem idênticamente distribuídas $\{p_1 = p_2 = p,\ q_1 = q_2 = q\}$ a expressão simplifica-se para,

$$F_W(w) = \begin{cases} 0 & w < 0 \\ \dfrac{w^p}{(B(p,q))^2}\displaystyle\sum_{k=0}^{q-1}\frac{B(2p+k,q)\binom{q-1}{k}(-w)^k}{k+p} + & \\ \quad +\dfrac{w^q}{(B(p,q))^2}\displaystyle\sum_{k=0}^{p-1}\frac{B(p,2q+k)\binom{p-1}{k}(-w)^k}{k+q} & 0 \le w < 1 \\ 1 & w \ge 1 \end{cases}$$

e no caso de $p = q$, obtemos

$$F_W(w) = \begin{cases} 0 & w < 0 \\ \dfrac{2\,w^p}{\left(B(p,p)\right)^2} \displaystyle\sum_{k=0}^{p-1} \frac{B(2p+k,p)\binom{p-1}{k}(-w)^k}{k+p} & 0 \leq w < 1 \\ 1 & w \geq 1 \end{cases}$$

Outros casos particulares, que poderão ter alguma relevância, no contexto i.i.d.:

- $q = 1$

$$F_W(w) = \begin{cases} 0 & w < 0 \\ \dfrac{w^p}{2} - \displaystyle\sum_{k=1}^{p} \frac{\binom{p}{k}(-w)^k}{\binom{p+k}{k}} & 0 \leq w < 1 \\ 1 & w \geq 1 \end{cases}$$

- $p = 1$

$$F_W(w) = \begin{cases} 0 & w < 0 \\ \frac{w^q}{2} - \displaystyle\sum_{k=1}^{q} \frac{\binom{q}{k}(-w)^k}{\binom{q+k}{k}} & 0 \leq w < 1 \\ 1 & w \geq 1 \end{cases}$$

- $p = q = 2$

$$F_W(w) = \begin{cases} 0 & w < 0 \\ (9 - 4w)\,\frac{w^2}{5} & 0 \leq w < 1 \\ 1 & w \geq 1 \end{cases}$$

Abordemos agora a questão de determinar a distribuição da variável aleatória $V = X + Y - \mathcal{I}[X + Y]$ no caso de X e Y serem v.a.´s i.i.d. com distribuição Beta$(p, 1)$, com p natural. Trata-se de generalizar o resultado conhecido no caso de X e Y serem v.a.´s i.i.d. Uniformes standard. Neste caso, sabe-se que V também tem distribuição Uniforme padrão e é independente de X e de Y.

Seja $v \in [0,1]$,

$$F_V(v) = \mathbb{P}(0 \leq X+Y \leq 1, X+Y \leq v) + \mathbb{P}(1 \leq X+Y \leq 2, X+Y \leq v+1) =$$

$$= p^2 \int_0^v x^{p-1} \left[\int_0^{-x+v} y^{p-1} \mathrm{d}y \right] \mathrm{d}x + p^2 \int_0^v x^{p-1} \left[\int_{-x+1}^1 y^{p-1} \mathrm{d}y \right] \mathrm{d}x + p^2 \int_v^1 x^{p-1} \left[\int_{-x+1}^{-x+v+1} y^{p-1} \mathrm{d}y \right] \mathrm{d}x$$

$$= p \int_0^v x^{p-1} (-x+v)^p \mathrm{d}x + p \int_0^v x^{p-1} \left[1 - (1-x)^p \right] \mathrm{d}x + p \int_v^1 x^{p-1} \left[(-x+v+1)^p - (1-x)^p \right] \mathrm{d}x$$

$$= -p\, B(p, p+1)(1 - v^{2p}) + v^p + p \sum_{k=0}^{p} \frac{(-1)^k \binom{p}{k} (v+1)^{p-k} (1 - v^{p+k})}{p+k}$$

A função de distribuição de V é um polinómio de grau máximo $2p$ e é um simples exercício verificar que para $p = 1$ dará a função de distribuição da uniforme padrão.

Se desenvolvermos um raciocínio análogo considerando X e Y v.a.´s i.i.d mas com distribuição Beta$(1, q)$, chegamos a resultados similares: para $v \in [0,1]$,

$$F_V(v) = \mathbb{P}(0 \leq X+Y \leq 1, X+Y \leq v) + \mathbb{P}(1 \leq X+Y \leq 2, X+Y \leq v+1) =$$

$$= q^2 \int_0^v (1-x)^{q-1} \left[\int_0^{-x+v} (1-y)^{q-1} \mathrm{d}y \right] \mathrm{d}x + q^2 \int_0^v (1-x)^{q-1} \left[\int_{-x+1}^1 (1-y)^{q-1} \mathrm{d}y \right] \mathrm{d}x + q^2 \int_w^1 (1-x)^{q-1} \left[\int_{-x+1}^{-x+v+1} (1-y)^{q-1} \mathrm{d}y \right] \mathrm{d}x =$$

$$= q \int_0^v (1-x)^{q-1} \left[1 - (1+x-v)^q \right] \mathrm{d}x + q \int_0^v (1-x)^{q-1} x^q \, \mathrm{d}x + q \int_v^1 (1-x)^{q-1} \left[(x)^q - (x-v)^q \right] \mathrm{d}x =$$

$$= q\, B(q+1, q) \left[1 - (1-v)^{2q} \right] + \left[1 - (1-v)^q \right] - q \sum_{k=0}^{q} \frac{(-1)^k \binom{q}{k} (2-v)^{q-k} (1 - (1-v)^{q+k})}{q+k}$$

Repare-se que $F_{V_{(1,q)}}(v) = 1 - F_{V_{(q,1)}}(1-v)$. Analisámos também o caso i.i.d. tendo ambas as variáveis distribuição Beta$(2,2)$: para $v \in [0,1]$,

$$F_V(v) = \mathbb{P}(0 \leq X+Y \leq 1, X+Y \leq v) + \mathbb{P}(1 \leq X+Y \leq 2, X+Y \leq v+1) =$$

$$= 36\int_0^v x(1-x)\left[\int_0^{-x+v} y(1-y)\mathrm{d}y\right]\mathrm{d}x + 36\int_0^v x(1-x)\left[\int_{-x+1}^1 y(1-y)\mathrm{d}y\right]\mathrm{d}x + 36\int_v^1 x(1-x)\left[\int_{-x+1}^{-x+v+1} y(1-y)\mathrm{d}y\right]\mathrm{d}x$$

donde

$$F_V(v) = \begin{cases} 0 & v < 0 \\ -\frac{6}{5}v^5 + 3v^4 - 2v^3 + \frac{6}{5}v & 0 \leq v < 1 \\ 1 & v \geq 1 \end{cases}$$

Uma questão que decerto merece investigação, neste ponto, é o que acontece se trocarmos numeradores com denominadores na expressão definitória de W^*, isto é se no numerador usarmos X com suporte (0,1) e no denominador $U \frown Uniforme(0,1)$, independentes. Atente-se nos resultados que se seguem:

Queremos então investigar a distribuição do mínimo dos inversos das variáveis anteriores, ou seja,

$$Z = \frac{1}{W^*} = \min\left(\frac{X}{U}, \frac{1-X}{1-U}\right).$$

O resultado obtido para o caso particular das Betas indica que a f.d. é polinomial com grau dependente dos parâmetros da Beta que surge em numerador;os parâmetros da Beta que surge em denominador apenas influenciam os coeficientes polinomiais o que sugere a não uniformidade desta variável, pelo menos para a sub-classe das Betas com pelo menos um dos parâmetros diferentes de um. Considerando $z \in (0,1)$,

$$F_Z(z) = \mathbb{P}\left[X \leq zU\right] + \mathbb{P}\left[(U \leq \frac{X+z-1}{z}\right] =$$

$$= \int_0^z f_X(x)\left[\int_{\frac{x}{z}}^1 \mathrm{d}u\right]\mathrm{d}x + \int_{1-z}^1 f_X(x)\left[\int_0^{\frac{x+z-1}{z}} \mathrm{d}u\right]\mathrm{d}x =$$

$$= \int_0^z f_X(x) \left[1 - \frac{x}{z}\right] \, \mathrm{d}x + \int_{1-z}^1 f_X(x) \left[\frac{x+z-1}{z}\right] \, \mathrm{d}x.$$

Efectuando no segundo integral a substituição $y = x + z - 1$, vem:

$$F_Z(z) = \int_0^z f_X(x) \left[1 - \frac{x}{z}\right] \, \mathrm{d}x + \int_0^z f_X(y+1-z) \left[\frac{y}{z}\right] \, \mathrm{d}y =$$

$$= F_X(z) + \int_0^z [f_X(x+1-z) - f_X(x)] \frac{x}{z} \, \mathrm{d}x).$$

Integrando por partes, chegamos à expressão

$$F_Z(z) = 1 - \frac{\int_0^z \{F_X(x+1-z) - F_X(x)\} \, \mathrm{d}x}{z}, \qquad z \neq 0$$

Para que

$$F_Z(z) = \begin{cases} 0 & z < 0 \\ z & z \in [0,1) \\ 1 & z \geq 1 \end{cases}$$

é condição necessária e suficiente que

$$(1-z)z = \int_0^z F_X(x+1-z) \, \mathrm{d}x - \int_0^z F_X(x) \, \mathrm{d}x.$$

Derivando ambos os termos e aplicando o Teorema Fundamental do Cálculo Integral chegamos a uma condição necessária e suficiente para que a distribuição citada seja uniforme standard:

$$f_X(z) + f_X(1-z) = 2.$$

Repare-se que todas as funções não negativas desta família são densidades pois o anterior requisito é suficiente para que o $\int_0^1 f_X(x)\ \mathrm{d}x = 1$. ... também fácil verificar que a própria Uniforme pertence a esta família, e que neste aspecto é unica dentro do grupo das Betas, uma vez que aquela identidade falha se $p \neq 1$ ou $q \neq 1$.

É evidente que as variáveis X_m estão na classe atrás caracterizada. Começa assim a fazer sentido que a potência dos procedimentos inferenciais usando pseudo$-p$'s construídos com aquelas "alternativas" esteja longe do que inicialmente esperávamos.

Mas, por outro lado, põem-se agora problemas de dependência que não existiam com parentes uniformes. Assim, o número de graus de liberdade da estatística de Fisher tem que ser devidamente ajustado.

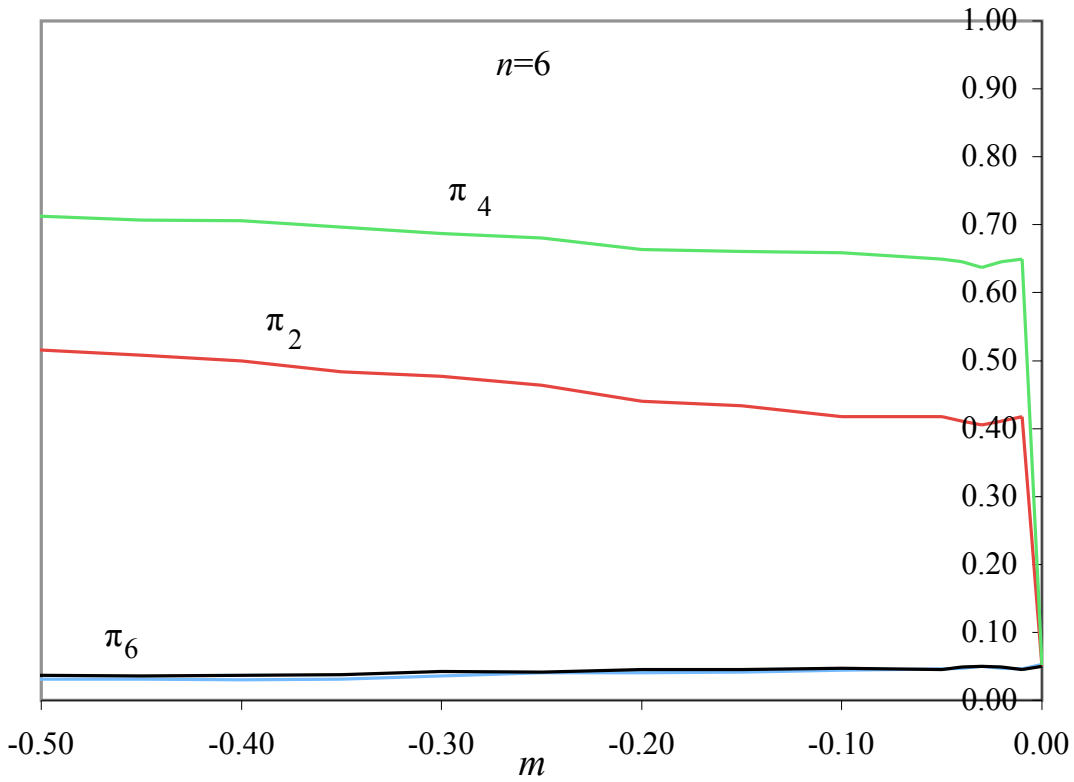

Figura 3.3: Efeito de ajustar o número de graus de liberdade.

Na Figura 3.3 apresenta-se um gráfico descrevendo o que acontece quando $n = 6$, $\alpha = 0.05$ e usamos simulação para ajustar o número de graus de liberdade, confira-se a discussão em Makambi (2003) sobre a combinação de $p - values$.

Capítulo 4

Harmonização de Efeitos, Replicação, Enviesamento na Publicação, e a Construção de Conhecimento Científico

4.1 Níveis de significância descritiva aleatórios

No capítulo anterior uma parte substancial da discussão centrou-se na harmonização de "$p-$values", na perspectiva que sob validade de H_0 podem ser considerados uma amostra da população uniforme padrão.

Ora parece ser esticar excessivamente a corda da plausibilidade admitir que se está a tentar sintetizar um conjunto de níveis de significância descritivos, presumivelmente inferiores a 0.05 e apontando para a rejeição de H_0, e pretender por outro lados que vamos fazer uma síntese no pressuposto que afinal H_0 é verdadeira!

É porventura por discrepâncias como esta que se observa uma crítica persistente ao uso dos valores de prova p como sumário relevante, nomeadamente quando

são usados para medir a *magnitude do efeito observado*. Kulisnkaya *et al.* (2008) defendem que os valores$-p$ não passam de um indicador de a que ponto o valor observado da estatística de teste é surpreendente, e propõem que toda a inferência estatística se faça em termos da *evidência em prol da hipótese alternativa*, ideia vaga que formalizam depois a partir do quantil $z_{0.95} = 1.645$ de $Z \frown Gaussiana(0, 1)$, e usando sempre estatísticas com desvio padrão 1 — por outras palavras, chamam a tenção que usando uma estatística de teste com distribuição amostral gaussiana padrão, observar o um "$p - value$"0.05 corresponde, nessa perspectiva, a ter uma evidência a favor da alternativa a que corresponde um intervalo de confiança de aproximadamente 68% (0.645,2.645). Consideram por isso que uma observação da estatística de teste próxima de 1.7 é uma evidência fraca a favor da alternativa, uma observação próxima de 3.5 uma evidência moderada, e uma observação próxima de 5 uma evidência forte de que H_0 deve ser abandonada para se adoptar H_A.

São ideias com algum mérito, embora obriguem de novo a privilegiar os métodos de Stouffer, transformando quaisquer quantis observados em *scores* gaussianos. Note-se que os autores usam transformações estabilizadoras da variância como base para a construção do que chamam "*key inferential function*", cuja função é exactamente permitir o cálculo da evidência em prol da alternativa usando *scores* gaussianos.

De qualquer forma, como Kulisnkaya *et al.* (2008, p. 123) sumariam

> "When faced with a number of 'significant' results, each of which casts some doubt on the null hypothesis, it is natural to want to combine these results, *and to do so under the alternative hypotesis.*"

Há diversas abordagens possíveis para implementar o desiderato expresso na frase que sublinham. Uma das que consideramos interessante é o uso de $random\ p-values$, que sumariamente descrevemos:

Seja S_r o modelo para o resultado da replicação de uma experiência S; as condições da replicação devem ser iguais às da experiência original.

O nível de significância descritivo aleatório $PV := \mathbb{P}[S_r \geq S] = 1 - F_0(S)$, que portanto tem função de distribuição

$$F_{PV}(p) = 1 - F_1\left(F_0^{-1}(1-p)\right),$$

onde F_1 é a função de distribuição de S.

Condicionalmente à validade de H_0, reencontra-se o conceito usual de $p-value$; sob a validade de H_A, parece o conceito adequado para sintetizar os resultados de testes, nomeadamente sobre a magnitude dos efeitos.

Um exemplo particularmente simples: Testar $H_0 : \mu = 0$ vs. $H_A : \mu = \mu_A > 0$ numa população X gaussiana, sabendo que o desvio padrão $\sigma = 1$, observando $X = x^{(1)}$, a densidade de PV pode ser expressa (Donahue, 1999) como

$$f_{PV}(p) = \frac{\varphi\left(\Phi^{-1}(1-p) - \mu\right)}{\varphi\left(\Phi^{-1}(1-p)\right)},$$

onde, como habitualmente, Φ e φ são a função de distribuição e a função densidade de probabilidade, respectivamente, da gaussiana padrão.

Apesar da vantagem da simplicidade, não passa de um exemplo académico. Em situações mais realistas — mesmo a simples questão $H_0 : \mu = \mu_0$ vs. $H_A : \mu = \mu_A > \mu_0$, no caso de população gaussiana com variância desconhecida, a usual estatística de teste sob H_0 tem distribuição t_{n-1}, mas sob H_A tem distribuição $t_{n-1,\lambda}$ não central, com parâmetro de não centralidade $\lambda = \sqrt{n}\,\dfrac{\mu_A - \mu_0}{\sigma}$, que podemos estimar por $\hat{\lambda} = \sqrt{n}\,\dfrac{\overline{x} - \mu_0}{\sigma}$. Evidentemente nesta situação a avaliação de $F_A\left(F_0^{-1}(1-p)\right)$ é complexa, Kulisnkaya *et al.* (2008, p. 24, e justificações mais detalhadas no capítulo 20) mostram que a evidência em prol da alternativa pode ser calculada à custa da chave inferencial

(1) Estamos a apresentar o mais simples dos exemplos, com uma amostra mais usual, de dimensão n, as expressões viriam um pouco mais complicadas, haveria que trabalhar com $\sqrt{n}\overline{X}$ e com $\sqrt{n}$ $li\mu A$.

$$\mathcal{K}(\delta) = \sqrt{2}\,\sinh^{-1}\left(\frac{\delta}{\sqrt{2}}\right) = \sqrt{2}\,\ln\left(\frac{\delta}{\sqrt{2}} + \sqrt{1+\frac{\delta^2}{2}}\right)$$

onde $\hat{\delta} = \dfrac{\overline{x} - \mu_0}{\sigma}$, "chave"essa a que se chega usando o método delta para obter uma transformação que estabilize a variância.

Kulisnkaya *et al.* (2008, p. 118–119) sumariam algumas propriedades deste níveis de significância aleatórios. Trabalhos anteriores de Dempster and Schatzoff (1965) Hung *et al.* (1999), Sakowitz and Samuel-Cahn(1999) fazem a ponte entre valores esperados e quantis. Veja-se também Bhattacharya and Habtzghi (2002), sobre $p - values$ medianos.

A questão de replicação de resultados experimentais é fundamental na metodologia da investigação científica, e não é demais recomendar o notável trabalho de Utts (1991) — mau grado os seus defeitos, mas a discussão por outros eminentes estatísticos que o editor convidou a comentá-lo é esclarecedora. Vale a pena atentar que na replicação de uma experiência que resultou num $p = 0.05$ se espera observar um $p^* = 0.12$ (Goodman, 1992), um resultado que decerto parece contraintuitivo para muitos cientistas, e que com certeza lança achas para a fogueira da polémica sobre enviesamento na publicação.

4.2 Índices de significância descritiva generalizados

Tsui e Weerahandi (1989), Weerahandi (1993), Gamage and Weerahandi (1998), entre outros, contribuíram para o refinamento de um conceito de níveis de significância descritiva ainda mais sofisticado definindo $p - values$ generalizados, na

situação de haver parâmetros perturbadores que compliquem o uso da estatística de teste.

Suponha-se que X é uma variável aleatória cuja função de distribuição depende não só de um parâmetro escalar θ como de um conjunto de parâmetros perturbadores $\boldsymbol{\eta}$, e queremos testar

$$H_0 : \theta \leq \theta_0 \qquad H_A : \theta > \theta_0$$

com base em x.

Suponha-se que é possível especificar uma variável de teste $T(X; x, \theta, \boldsymbol{\eta})$ tal que

- Para x fixo, a distribuição amostral de $T(X; x, \theta, \boldsymbol{\eta})$ não depende do parâmetro perturbador $\boldsymbol{\eta}$.
- O valor observado de $T(X; x, \theta, \boldsymbol{\eta})$ quando $X = x$, $T(x; x, \theta, \boldsymbol{\eta})$ não depende do parâmetro perturbador $\boldsymbol{\eta}$.
- Para x e $\boldsymbol{\eta}$ fixos, $\mathbb{P}\left[T(X; x, \theta, \boldsymbol{\eta}) \geq t\right]$ é uma função não decresceente de θ.

Nestas condições o nível de significância generalizado, GP, é

$$GP := \mathbb{P}\left[T(X; x, \theta_0, \boldsymbol{\eta}) \geq T(x; x, \theta_0, \boldsymbol{\eta})\right].$$

Verbeke and Molenberghs (1997) e Weerahandi (2003, 2004) contêm usos sofisticados e interessantes do conceito de $p - value$ generalizado, nomeadamente na construção de intervalos de confiança sintetizando resultados, veja-se também Hartung *et al.* (2008, p. 81–83).

É um desafio realizar com PV's ou GP's um estudo distribucional e/ou de simulação semelhante ao levado a cabo no capítulo anterior. O problema encontra-se parcialmente (e "aproximativamente") resolvido no que respeita ao importante problema de testar um efeito por comparação de valores médios de duas gaussianas heterocedásticas com o teste de Welch-Satterthwaite, veja-se Tang and Tsui (2007).

4.3 Escrever para a gaveta

Nas ciências experimentais as conclusões devem provir de evidência factual devidamente enquadrada pela teoria que presidiu á escolha de um planeamento experimental adequado para a recolecção dos dados. Apesar de se assistir frequentemente à publicação de "pesquisas cegas", com resultados aparentemente significativos mas que não resistem a estudos de replicação, a contribuição destes estudos para o avanço da Ciência é muitas vezes efémera ou mesmo nula. Por outro lado, muitos estudos mais sérios, mas que por isso mesmo não produzem os $p-values$ significativos tão do agrado dos utilizadores, nem sempre bem informados e críticos, da Estatística, raramente conseguem publicação no seu estatuto mais nobre, e não passam de documentação interna, que em Ciência tem um estatuto de quase clandestinidade. Os próprios autores se auto-censuram, e não se esforçam por publicar resultados não-significativos, porventura porque têm experiência de que em geral esses resultados não têm em geral acolhimento nas revistas da especialidade.

Assim, grosseiramente, poderíamos caracterizar o panorama da publicação científica em três categorias

- publicação, ou publicações em número escasso, de resultados em que a evidência factual ficou claramente estabelecida, por a magnitude do efeito ser incontroversa e baseada em amostras de dimensão adequada, garantindo a potência dos resultados — neste caso não se sente a necessidade de publicação de estudos adicionais que mais não fazem do que replicar resultados, mas este enviesamento na publicação é venial, no sentido em que também não há qualquer razão para meta-analisar resultados.

- não publicação de trabalhos por terem tido resultados inconclusivos, porventura apenas por a dimensão das amostras ser inferior à necessária para obter magnitudes de efeitos interessantes; a combinação meta-analítica desses diversos estudos inconclusivos poderia ser substancial, mas não existe sequer o

material de base para indicar os benefícios que poderiam advir de uma síntese meta-analítica.

- publicação de um ou mais estudos mostrando um efeito significativo, contrariados por outros estudos que não replicam os resultados; ou vários estudos significativos, mas pouco convincentes por o protocolo de obtenção dos dados, ou a dimensão das amostras usadas, fundamentarem alguma contestação ou pelo menos falta de convicção na comunidade científica.

É neste terceiro contexto que em geral as sínteses meta-analíticas têm um papel de relevo, procurando extrair uma conclusão harmonizando toda a informação disponível. Veja-se por exemplo o interessante trabalho de Utts (1991) sobre parapsicologia e a eventual capacidade de um emissor influenciar "telepaticamente" um receptor de quem está electromagneticamente isolado, ou o relatório de Shekelle et al. (2003) sobre os os efeitos secundários de ephedra ou ephedrina mais cafeína.

Mas este terceiro contexto é exactamente aquele em que se coloca com maior premência a problemática do enviesamento na publicação. Porventura os estudos publicados privilegiam os que trazem alguma novidade à polémica científica em curso, e consequentemente os que reforçam a significância dos resultados, ou a incapacidade de replicação, e porventura é uma situação em existe mesmo uma deficiência na análise de base, faltando a medição das magnitudes dos efeitos a determinados níveis dos "estímulos"[(2)].

Um *gráfico em funil*[(3)] pode contribuir para o diagnóstico de que existe de facto

[(2)] É esta constatação que leva à repetida afirmação nos textos de meta-análise de que um dos seus efeitos colaterais é poder detectar que há estudos que deveriam ter sido executados e ainda estão em falta. Infelizmente este aspecto muito positivo da meta-análise ainda não ganhou a popularidade que a nosso ver merece.

[(3)] (*Funnel plot*): representação gráfica de pontos (x, y), em que a abcissa é a magnitude do efeito e a ordenada a dimensão da amostra (ocasionalmente, em vez da dimensão da amostra usa-se o erro padrão como ordenada), pode revelar a existência de enviesamento na publicação. De facto, se não houver enviesamento na publicação, é de esperar grande dispersão dos pontos correspondentes a estudos baseados em amostras pequenas na base do gráfico, e concentração no topo, e consequentemente um perfil de *funil* com a ponta virada para cima.

enviesamento na publicação. Apesar do nome, a forma nem sempre é a de funil apontado para cima, e é esse formato inesperado que nos leva a desconfiar de que existe enviesamento na publicação.

A existência de enviesamento na publicação é não só difícil de detectar como quase sempre impossível de resolver. A partir do trabalho de Rosenthal (1979) o problema de muitos estudos terem resultado num "escrever para a gaveta" (*file drawer problem*) passou a ser tratado indirectamente: procura determinar-se qual é o número k_0 de "estudos para a gaveta" por não terem resultados significativos levariam, quando combinados com os k estudos usados na síntese meta-analítica, a uma síntese qualitativamente diferente.

Suponha-se por exemplo que estamos a preparar uma síntese meta-analítica de $n = 16$ testes independentes usando a estatística de teste T para testar uma hipótese nula H_0. Suponha-se ainda que usando o método de Stouffer *et al.* (1949) transformamos os níveis de significância descritiva $\{p_k\}_{k=1}^{16}$ em *scores* gaussianos $\left\{z_k = \Phi^{-1}(p_k)\right\}_{k=1}^{16}$, e que a observação extrema $Z(obs.) = \frac{1}{\sqrt{16}} \sum_{k=1}^{16} z_k = -2.7462$ nos leva a rejeitar, globalmente, a hipótese nula ao nível $\alpha = 0.05$, por exemplo (visto que $\mathbb{P}\left[|Z| > Z(obs.) = 2.7462\right] \approx 0.0030$).

Suponha-se que k_0 estudos não reportados por não terem resultados significativos podiam inverter aquela rejeição. Podemos admitir, numa primeira aproximação (veja-se Hartung *et al.* (2008, p. 175–178) refinamentos do método, e descrição de resultados análogos usando as metodologias de Tippett e de Fisher para meta-analisar $p-values$), que a contribuição dos scores gaussianos dos níveis de significância descritiva desses estudos estaria próxima do seu valor esperado 0 (isto é, que $\sum_{k=1}^{k_0} z_k \approx 0$).

Contrariar a decisão inicial de rejeição, ao nível de significância 0.05, por exem-

É muito comum uma barra horizontal ancorada em cada ponto indicar visualmente um intervalo de confiança para a magnitude do efeito.

plo, significaria então que

$$|Z(obs.)| = \left| \frac{1}{\sqrt{16+k_0}} \sum_{k=1}^{16+k_0} z_k \right| < 1.645 = z_{0.95}.$$

O menor valor de k_0 para o qual tal acontece é então $k_0 = 29$ — a questão que se põe é então se se considera credível que haja 29 estudos não publicados devido ao enviesamento na publicação que tende a deixar na gaveta estudos cujos resultados não são, pelo menos na aparência imediata, significativos.

Note-se que na perspectiva simplificada apresentada o que se pretende é determinar o k_0 mínimo que contrarie a rejeição global de k estudos, ao nível de significância α. Quer isto dizer que se teria

$$\frac{1}{\sqrt{k+k_0}} \sum_{n=1}^{k+k_0} \Phi^{-1}(p_n) \approx \frac{1}{\sqrt{k+k_0}} \sum_{n=1}^{k} \Phi^{-1}(p_n) < z_{1-\alpha},$$

o que acontece se

$$k_0 > \left(\frac{\sum_{n=1}^{k} \Phi^{-1}(p_n)}{z_{1-\alpha}} \right)^2 - k.$$

4.4 Efeito do uso de pseudo−p's

Apresentamos agora algumas indicações elementares do que acontece quando se usam os resultados do capítulo 3 para aumentar computacionalmente a amostra de níveis de significância descritivos obtidos em testes independentes com pseudo−p's.

O uso de marcadores imuno-histoquímicos no estudo diagnóstico e prognóstico de melanoma desmoplásico tem permitido progressos importantes, ainda que a maior parte da evidência disponível se baseie em estudos usando um número muito escasso (no limite 2!) de doentes. Até 2008 apenas cinco estudos publicados se baseavam na observação de mais do que 100 doentes: Quinn *et al.* (1988), Posther *et al.* (2006), Skelton *et al.* (1995), Soares de Almeida *et al.* (2008), e Ohsie *et al.* (2008), com respectivamente 280, 129, 128 e 113 casos. Estes foram obviamente retidos na revisão sistemática de Pestana, Sequeira e Soares de Almeida (2008)[(4)], tal como — e como é de rigor — a revisão sistemática de Lens *et al.* (2005).

Suponha-se que os níveis de significância descritiva em quatro dos estudos referidos, no que se refere à relevância do marcador S100 no prognóstico de melanoma desmoplásico são $(0.0412, 0.0436, 0.0324, 0.0273)$, a que correspondem os *scores* normais $(-1.7369, -1.7104, -1.8466, -1.9220)$. Assim, observa-se o valor $\frac{-7.2160}{\sqrt{4}} = -1.8040$, o que quer dizer que o nível de significância descritivo combinado é 0.0356, levando à rejeição da hipótese nula de não haver efeito. Para contrariar esta conclusão, ao nível de significância usual 0.05, seria necessário que houvesse $\left(\frac{-7.2160}{1.645}\right)^2 - 4 = 16$ estudos não publicados por não terem conseguido demonstrar significância.

Suponha-se agora que usamos um gerador de números pseudo-aleatórios uniformes para gerar 4 valores — por exemplo $(0.2555, 0.4861, 0.4835, 0.7546)$. Isso permite-nos construir uma amostra com pseudo$-p$'s com a dimensão 12, no caso dos valores reportados

$$(0.0412, 0.0436, 0.0324, 0.0273, 0.1613, 0.0897, 0.0670, 0.0362, 0.2025, 0.1333, 0.0994, 0.0635),$$

a que corresponde a amostra de *scores* normais negativos com valores absolutos

$$(1.7369, 1.7104, 1.8466, 1.9220, 0.9893, 1.3426, 1.4984, 1.7969, 0.8329, 1.1110, 1.2849, 1.5262).$$

(4) Cuja consulta nos dispensa de sobrecarregar a já longa bibliografia com o arrolamento de trabalhos sobre melanoma desmoplásico que foram retidos para esse estudo.

Passámos portanto a observar um valor $\frac{-17.5982}{\sqrt{12}} = -5.0802$, que leva a rejeitar a hipótese nula ao nível de significância descritivo $1.886 \times ao^{-7}$. Seriam agora necessários $\left(\frac{-17.5982}{1.645}\right)^2 - 12 = 103$ estudos na gaveta, por não serem significativos, para contrariar esta decisão!

Não só a decisão de rejeitar pôde ser tomada com mais convicção, como o número de estudos que m que existir, não publicados, para a invalidarem, mais do que sextuplicou.

Claro que só um trabalho exaustivo de simulação com amostras de várias dimensões, e com diversos vectores de $p-values$ originais permitirá uma avaliação objectiva e concreta. Esse estudo foi feito no caso particular $n = 4$, da seguinte forma

- Foram gerados 120 pseudo-aleatórios uniformes, retendo-se os primeiros quatro que fossem inferiores a 0.05, $\{p_1, \ldots, p_4\}$; calcularam-se os correspondentes *scores* gaussianos, $\{\Phi^{-1}(p_1), \ldots, \Phi^{-1}(p_4)\}$.

- Determinou-se o número de estudos $A = \left(\frac{\sum_{n=1}^{4} \Phi^{-1}(p_n)}{1.645}\right)^2 - 4$ não significativos (ao nível 0.05) seriam necessários para contrariar uma decisão de rejeição de H_0 com base na observação daqueles 4 $p - values$. (Claro que se este resultado for negativo, esse conjunto de valores não é contabilizado no estudo de simulação, porque corresponde a uma situação em que a síntese dos $p - values$ observados não leva a rejeição da hipótese nula).

- Amplia-se a amostra inicial de $p - values$ com base em 4 uniformes, usando $p_{4+k} = \min\left(\frac{u_k}{p_k}, \frac{1-u_k}{1-p_k}\right)$, e calcula-se $B = \left(\frac{\sum_{n=1}^{8} \Phi^{-1}(p_n)}{1.645}\right)^2 - 8$.

  O índice $K_1 = \frac{B}{A}$ mede a que ponto a ampliação da amostra com pseudo$-p$'s, para o dobro, implica no aumento do número de casos não significativos para "virar" a decisão.

- Geram-se mais 4 pseudo$-p$'s com a expressão $p_{8+k} = p_k + p_{4+k} - \mathcal{I}\left(p_k + p_{4+k}\right)$, calcula-se $C = \left(\frac{\sum_{n=1}^{12} \Phi^{-1}(p_n)}{1.645} \right)^2 - 12$.

  O índice $K_2 = \frac{C}{A}$ mede a que ponto a ampliação da amostra com pseudo$-p$'s, para o triplo, implica no aumento do número de casos não significativos para "virar" a decisão.

- Os procedimentos acima esquematizados foram repetidos N=10 000 vezes, calculando as médias parciais dos índices K_1 e K_2, que estabilizaram em torno de 5 e 11, respectivamente.

Assim, conclui-se que também neste aspecto a ampliação do conjunto de $p-$*values*, por muito artificiosa que seja, reforça a nossa convicção no resultado da síntese meta-analítica.

Bibliografia

- Abrahams, M. (2002) *The Ig Nobel Prizes: The Annals of Improbable Research*, Orion.

- Abrahams, M. (2005) *Ig Nobel Prizes: Why Chickens Prefer Beautiful Humans*, Orion.

- Ananda, M. M. A., and Weerahandi, S. (1977). Two-way anova with unequql cell frequencies and unequal variances. *Statistics Sinica* **7**, 631–646.

- Asiribo, O., and Gurland, J. (1990). Coping wth variance heterogeneity. *Commun. Statist. — Theory Methods* **19**, 4029–4048.

- Baillar III, J. C., and Mosteller, F. (1992). *Medical Uses of Statistics*, 2nd ed., New England Journal of Medicine Books, Boston.

- Bangert-Drowns, R. L., Rudner, L. M. (1991). Meta-analysis in educational research. *Practical Assessment, Research & Evaluation* **2**, 8.

- Beecher, H. K. (1955). The powerful placebo, *J. Amer. Medical Assoc.* **159**, 1602–1606.

- Berkey, C. S., Hoaglin, D. C., Mosteller, F., and Colditz, G. A. (1996). A random-effects regression model for meta-analysis, *Statistics in Medicine* **14**, 395–411.

- Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *Ann. Statist.* **8**, 457–487.

- Bhattacharya, B., and Habtzghi, D. (2002). Median of the $p-$ value under the alternative hypothesis. *Amer. Statistician* **56**, 202–206.

- Biggerstaff, B., and Tweedie, R. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16**, 753–768.

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*, Wiley, Chichester.

- Brilhante, M. F. Pestana, D., Rocha, J., e Velosa, S. (2001). *Inferência Estatística Sobre Localização e Escala*, Sociedade Portuguesa de Estatística, Ponta Delgada.

- Brown, M. B., and Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. it Technometrics **16**, 129–132.

- Casella, G., and Berger, R. L. (2002). *Statistical Inference*, 2nd ed., Duxbury Press, Pacific Grove.

- Chalmers, I., and Altman, D. G. (1995). *Systematic Reviews*, British Medical Journal Publ., London.

- Chalmers, I., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., and Ambroz, A. (1981). A method for assessing the quality of randomized control trials, *Controlled Clinical Trials* **2**, 31–49.

- Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments*, Springer, New York.

- Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *J. Roy. Statist. Soc.* (*Supplement*) **4**, 102–118.

- Cochran, W. G. (1954). The combination of estimates from different experiments, *Biometrics* **10**, 101–129.

- Collins, R., Yusuf, S., and Peto, R. (1985). Overview of randomized trials of diuretic pregnancy, *British Medical J.* **290**, 17–23.

- Cooper, H., and Hedges, L. V. (1994). *The Handbook of Research Synthesis*, Sage, Newbury Park.

- Cressie, N., and Read, T. R. C. (1984). Multinomial Goodness-of-fit tests, *J. Roy. Statist. Soc.* **B 46**, 440–464.

- Davey Smith, G., and Egger, M. (1988). Meta-Analysis — unresolved issues and future developments. *British Medical Journal* **315**, 221–225.

- David, H. A., and Nagaraja, H. N. (2003). *Order Statistics*, 3rd ed., Wiley, New York.

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

- Deng, L.-Y., and Chu, Y.-C. (1991). Combining random number generators, in B. L. Nelson, W. D. Kelton and G. N. Clark (Eds.) *Proceedings of the 1991 Winter Simulation Conference*, 1043–1046.

- Deng, L.-Y., and George, E. O. (1992). Some characterizations of the uniform distribution with applications to random number generation. *Ann. Instit. Statistical Mathematics* **44**, 379–385.

- DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177–188.

- Donahue, R. (1999). A note on information seldom reported via the $p-$value. *Amer. statistician* **53**, 303–3006.

- Draper, D., Gaver, D., Goel, P., Greenhouse, J., Hedges, L., Morris, C., and Waternaux, C. (1992). *On Combining Information: Statistical Issues and Opportunities for Research*, Rep. panel convened by the National Academy of Sciences.

- Dunnet, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *J. Amer. Statist. Assoc.* **75**, 796-800.

- Durbin, J. (1973). *Distribution Theory of Tests Based on the Sample Distribution Function*, SIAM, Philadelphia.

- Eddy, D. M., Hasselblad, V., and Schachtner, R. D. (1990). Bayesian method for synthesizing evidence: The confidence profile method, *Internat. J. Techn. Assessment in Health Care* **6**, 31–56.

- Eddy, D. M., Hasselblad, V., and Schachtner, R. D. (1992). *Meta-Analysis by the Confidence Profile Method*, Academic Press, San Diego.

- Egger, M., and Davey Smith, G. (1997). Meta-Analysis — potentials and promise. *British Medical Journal* **315**, 1371–1374.

- Egger, M., and Davey Smith, G. (1997a). Meta-Analysis — principles and procedures. *British Medical Journal* **315**, 1533–1537.

- Egger, M., and Davey Smith, G. (1997b). Meta-Analysis — beyond the grand mean? *British Medical Journal* **315**, 1610–1614.

- Egger, M., and Davey Smith, G. (1998). Meta-Analysis — bias in location and selection of studies. *British Medical Journal* Ê**316**, 61–66.

- Egger, M., Davey Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical J.* **315**, 629–634.

- Egger, M., Schneider, M., and Davey Smith, G. (1998a). Meta-Analysis — spurious precision? Meta-analysis of observational studies. *British Medical Journal* **315**, 140–144.

• Everitt, B. S. (1992). *The Analysis of Contingency Tables*, 2nd ed., Chapman and Hall, London.

- Fisher, R. A. (1932). *Statistical Methods for Research Workers*, 4th ed., Oliver and Boyd, London.

- Fisher, R. A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford Univ. Press, Oxford.

- Fleiss, J. L. (1993). The statistical basis of meta-analysis, *Statist. Meth. Medical Res.* **2**, 121–145.

- Galbraith, R. F. (1988). Graphical display of estimates having differing standard errors, *Technometrics* **30**, 271–281.

- Galbraith, R. F. (1988a). A note on graphical presentation of estimated odds ratios from several clinical trials, *Statistics in Medicine* **7**, 889–894.

- Gamage, J., and Weerahandi, S. (1998). Size performance of some tests in one-way anova. *Commun. Statist. — Simul. Comput.* **27**, 625–640.

- Gans, D. J. (1991). Preliminary tests on variances. *Amer. Statist.* **45**, 258.

- Gaylor and Hopper (1969). Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. *Technometrics* **11**, 699-706.

- Gilbert, N. (1989). *Biometrical Interpretation — Making Sense of Statistics in Biology*, 2nd ed., Oxford Univ. Press, Oxford.

- Glass G. V, (1976). Primary, secondary and meta-analysis of research. *Educational Researcher* **10**, 3–8.

- Glass, G. V. (1999). Meta-Analysis at 25. `http://glass.ed.asu.edu/gene/papers/meta25.html`.

- Gokhale, D. V., and Kullback, S. (1979). *The Information in Contingency Tables*, M. Dekker, New York.

- Goodman, S. (1992). A comment on replication, $p-$values and evidence. *Statististics in Medicine* **11**, 875–879.

- Goodman, S. (1998). $P-$value, in Armitage, P., and Colton, T (eds *Encyclopedia of Biostatistics*, vol. 4, 3233–3237.

- Greenland, S., and Salvan, A. (1990). Bias in the one-step method for pooling study results. *Statist. Medicine* **9**, 247–252.

- Griffiths, W., and Judge, G. (1992). Testing and estimating location vectors when the error covariances matrix is unknown. *J. Econometrics* **54**, 121–138.

- Hartung, J., Argac, D., and Makambi, K. H. (2002). Small sample properties of tests of heterogeneity in the one-way anova and meta-analysis. *Statistical papers* **43**, 197-235.

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

- Hedges, L. V., and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Boston.

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

- Ioannidis, J. P. A. (2005). Why most published research findings are false, *PLoS Med.* **2**(8):e124, 696–701.

- Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, vol. 2, end ed., Wiley, New York.

- Kendall, M. G., and Stuart, A. (1961). *The Advanced Theory of Statistics, II: Inference and Relationship*, Griffin, London.

- Khuri, A. I., Mathew, T., and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*, Wiley, New York.

- Kitchens, L. J. (1998). *Exploring Statistics*. Duxbury Press, Pacific Grove.

- Koch, G. G., Gillings, D. B., and Stokes, M. E. (1980). Biostatistical implications of design, sampling, and measurement to health science data analysis, *Biometrics* **1**, 163–225.

- Koch, G. G., Amara, I. A., Davis, G. W., and Gillings, D. B. (1982). A review of some statistical methods for covariance analysis of categorical data, *Annual Review Public Heath* **38**, 563–595.

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

- Kulinskaya, E., and Staudte, R. G. (2007). Confidence intervals for the standardized effect arising in comparisons of two normal populations, *Statistics in Medicine* **26**, 2853–2871.

- Kuritz, S. J., Landis, J. R., and Koch, G. G. (1988). A general overview of Mantel–Haenszel methods: applications and recent developments, *Annual Review Public Heath* **9**, 123–160.

- Larholt, K. M. (1989). *Statistical Methods and Heterogeneity in Meta-Analysis*, Ph. D. dissertation, Harvard School of Public Health.

- Larholt, K. M., Tsiatis, A. A., and Gelber, R. D. (1989). Variability of coverage probability when applying a random effects methodology for meta-analysis, in Larholt (1989), *Statistical Methods and Heterogeneity in Meta-Analysis*, Ph. D. dissertation, Harvard School of Public Health.

- Lens, M. B., Newton-Bishop, J. A., and Boon, A. P. (2005). Desmoplastic malignant melanoma: a systematic review. *British J. Dermatology* **152**, 673–678.

- Light, R. J. , and Pillemer, D. B. (1984). *Summing Up. The Science of Reviewing Research*, Havard University Press, Cambridge, Mass.

- Littenberg, B., Moses, L., and Rabinowitz, D. (1990). Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method, *Clinical Research* **38**, 415A.

- Longford, N. T. (2008). *Studying Human Populations*. Springer Verlag.

- Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *J. Nat. Cancer Inst.* **22**, 719–748.

- Markowski, C. A., and Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *Amer. Statist.* **44**, 322-326.

- McManus, C. (2003). *Right Hand, Left Hand : The Origins of Asymmetry in Brains, Bodies, Atoms and Cultures*, Harvard University Press.

- Mehrota, D. V. (1997). Improving the Brown-Forsythe solution to the generalized Behrens-Fisher problem. *Commun. Statist. — Simul. Comput.* **26**, 1139–1145.

- Milliken, G. A., and Johnson, D. E. (1992). *Analysis of Messy Data, I: Design of Experiments* Chapman & Hall, London.

- Milliken, G. A., and Johnson, D. E. (1989). *Analysis of Messy Data, II: Nonreplicated Experiments* Chapman & Hall, New York.

- Milliken, G. A., and Johnson, D. E. (2002). *Analysis of Messy Data, III: Analysis of Covariance*, Chapman & Hall, New York.

- 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. *Communications in Statistics — Theory and Methods* **18**

- Mosteller, F., and Bush, R. (1954). Selected quantitative techniques, in G. Lidsey (Ed.), *Handbook of Social Psychology: Theory and Methods*, vol. I, Addison-Wesley, Cambridge, MA.

- Mosteller, F., and Chalmers, T. C. (1992). Some progress and problems in meta-analysis of clinical trials, *Statistical Science* **7**, 227–236.

- Mulrow, C. D. (1994). Rationale for systematic reviews, *British Med. J.* **309**, 597–599.

- Murteira, B. J. F. (1988). *Estatística: Inferência e Decisão.* Imprensa Nacional/Casa da Moeda, Lisboa.

- Nel, D., van der Merwe, C. A.,and Moser, B. (1990). The exact distribution of the univariate and multivariate Behrens-Fisher statistics with a comparison of several solutions in the univariate case, *Comm. Statist. — Theory and Methods* **19**, 279–298.

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

- Ohsie, S. J., Sarantopoulos, G. P., Cochran, A. J., and Binder, S. W. (2008). Immunohistochemical characteristics of melanoma. *Journal of Cutaneous Pathology* **35**, 207–215.

- Olkin, I. (1995). Statistical and theoretical considerations in meta-analysis, *J. Clinical Epidemiology* **48**, 133–146.

- Oxman, A. D. (1994). Preparing and maintaining systematic reviews, in *Cochrane Collaboration Handbook*, Sec. VI, D. Sackett, ed., Cochrane Collaboration, Oxford.

- Oxman, A. D., Clarke, M. J., and Stewart, L. A. (1995). From science to practice: meta-analysis using individual patient data are needed, *J. Amer. Medical Assoc.* **274**, 845–846.

- Paul, A. (2003). Characterizations of the uniform distribution via sample spacings and nonlinear transformations. *J. Mathematical Analysis and Applications* **284**, 397–402.

- Pestana, D., Brilhante, F., and Rocha, J. (1999). The analysis of variance revisited. In *Extreme Values and Additive Laws*, 73–77, Lisboa.

- Pestana, D., Rocha, J., Sequeira, F., e Vasconcelos, R. (2006). Meta-Análise, a Síntese de Evidência Estatística, *Estatística e Qualidade na Saúde*, **IV**, 48–74.

- Pestana, D., Sequeira, F., e Soares de Almeida, L. (2008). Meta-analytical assessment of predictive values of positive and negative immunohistochemi-

cal markers of desmoplastic melanoma, II Iberian Mathematical Meeting, http://imm2.unex.es/poster/list.html.

- Pestana, D. D., and Vasconcelos R. (1999). Power divergence statistics, *Extreme Values and Additive Laws*, CEAUL, Lisboa, 78–81.

- Pestana, D., e Velosa, S. (2002). *Introdução à Probabilidade e à Estatística*, vol. I, Fundação Gulbenkian, Lisboa.

- Pettiti, D. (1994). *Meta-Analysis, Decision Analysis and Cost Effectiveness Analysis*, Oxford Univ. Press, Oxford.

- Posther, K. E., Selim, M. A., Mosca, P. J. (2006). Histopathologic characteristics, recurrence patterns, and survival of 129 patients with desmoplastic melanoma. *Ann. Surg. Oncol.* **13**, 728–739.

- Pyke, R. (1965). Spacings. *J. Royal Statist. Soc.*Ê**B27**, 395–436.

- Quinn, M. J., Crotty, K. A., Thompson, J. F. (1998). Desmoplastic and desmoplastic neurotropic melanoma: experience with 280 patients. *Cancer* **83**, 1128–1135.

- Rosenthal, R. (1979). The "file-drawer problem"and tolerance for null results. *Psychological Bull.* **86**, 638–641.

- Rosenthal, R. (1991). Meta-analysis: a review, *Psychosom. Med.* **53**, 247–271.

- Sackowitz, H., and Samuel-Cahn,, E. (1999). $P-$values as random variables, expeected $p-$values. *Amer. Statistician*Ê**53**, 326–331.

- Samuels, M. L., and Witmer, J. A. (1999). *Statistics for the Life Sciences.* Prentice Hall, Upper Saddle River.

- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110-114.

- Senn, S. (2000). The many modes of Meta. *Drug Information J.* **34**, 535-549.

- Senn, S. (2007). *Statistical Issues in Drug Development*, 2nd ed. Wiley, Chichester.

- Shannahoff-Khalsa, D., Boyle, M. R., and Bubel, M. E. (1991). The effect of unilateral forced nostril breathing on cognition, J. Neuroscience **57**, 239–249.

- Shekelle, P., Morton, S., Maglione, M., *et al.* (2003). Ephedra and Ephredine for weight loss and athletic performance enhancement: clinical efficacity and side effects. California Agency for Healthcare Research and Quality Publication no. 03–E022, Rockville.

- Simes, R. J. (1986). Publication bias: the case for an international registry of clinical trials, *J. Clin. Oncology* **4**, 1529–1541.

- Simpson, E. H. (1951). The interpretation of interaction in contingency tables, *J. Roy. Statist. Soc.* **B13**, 238–241.

- Skelton, H. G., Smith, K. J., Laskin, W. B. (1995). Desmoplastic malignant melanoma. *J. Am. Acad. Dermatol.* **32**, 717–725.

- Smith, H. (1936). The problem of comparing the results of two experiments with unequal means. *J. Council Sci. Industr. Res.* **9**, 211-212.

- Soares de Almeida, L., Requena, L., Rütten, A., Kutzner, H., Garbe, C., Pestana, D., and Marques Gomes, M. (2008). Desmoplastic malignant melanoma: a clinicopathologic analysis of 113 cases. *Am. J. Dermapathol.* **30**, 151-174.

- Stangl, D. K., and Berry, D. A. (2000). *Meta-Analysis in Medicine and Health Policy*, M. Dekker, New York.

- Stangl, D. K., and Berry, D. A. (2000). Meta-analysis: past and present challenges, in Stangl and Berry, eds. (2000). *Meta-Analysis in Medicine and Health Policy*, M. Dekker, New York, 1–28.

- Sterne, J. A. C., and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis, *Journal of Clinical Epidemiology*Ê**54**, 1046–1055.

- Steven, S. S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.

- Stouffer, S. A., Schuman, E. A., DeVinney, L. C., Star, S. and Williams, R. M. (1949). *The American Soldier*, vol. I: *Adjustment During Army Life*, Princeton University Press, Princeton.

- Sukhatme, P. V. (1937). Tests of significance for samples from the χ^2 population with two degrees of freedom. *Annals of Eugenics* **8**, 52–56.

- Thompson, S. G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction, *Statist. Meth. Medical Res.* **2**, 173–192.

- Tang, S., and Tsui, K.-W. (2007). Distributional properties for the generalized $p-$value for the Behrens-Fisher problem, *Statistics & Probability Letters*Ê**77**, 1–8.

- Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated, *British Med. J.* **309**, 1351–1355.

- Thompson, S. G., and Pocock, S. J. (1991). Can meta-analysis be trusted? *Lancet* **338**, 1127–1130.

- Thompson, S. G., and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods, *Statistics in Medicine*Ê**18**, 2693–2708.

- Thursby, J. G. (1992). A comparison of several exact and approximate tests for structural shift under heteroscedasticity. *J. Econometrics* **53**, 363–386.

- Tippett, L. H. C. (1931). *The Methods of Statistics*, Williams & Norgate, London.

- Tsui, K., and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses. *J. Amer. Statist. Assoc.* **84**, 602-607.

- Tufte, E. R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Conn.

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

- van Belle, G. (2008). *Statistical Rules of Thumb*, 2nd ed., Wiley.

- Velosa, S. (2003). *O Problema de Behrens-Fisher*, Escolar Editora, Lisboa.

- Verbecke, G., and Molenberghs, G. (1997). *Linear Mixed Models in Practice*, Springer, New York.

- Wakefield, J. (1996). The bayesian analysis of population pharmacokinetic models, *J. Amer. Statist. Assoc.* **91**, 62–75.

- Wakefield, J., and Rahman, N. (2000). The combination of population pharmacokinetic studies, *Biometrics* **56**, 263–270.

- 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.

- Weerahandi, S. (1993). Generalized confidencee intervals. *J. Amer. Statist. Assoc.* **88**, 899–905.

- Weerahandi, S. (2004). *Generalized Inference in Repeated Measures*, Wiley, New York.

- 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. (1943). The generalization of "Student's"problem when several different population variances are involved. *Biometrika* **34**, 28-35.

- Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika* **38**, 330-336.

- Welch, B. L. (1956). On linear combinations of several variances. *J. Amer. Statist. Assoc.* **51**, 132-148.

- Wilkinson, B. (1951). A statistical consideration in psychological research.*Psychological Bulletin* **48**, 156–158.

- Woodward, M. (2005). *Epidemiology — Study Design and Data Analysis*, 2nd ed. Chapman & Hall/CRC.

- Yule, G. U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–134.

- Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blockage during and after myocardial infarction: an overview of the randomized trials, *Progress in Cardiovascular Disease* **27**, 335–371.

- Zar, J. H. (1999). *Biostatistical Analysis*, 4th ed., Prentice Hall, Upper Saddle River.

Indice Onomástico