Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Combinação de Valores de Prova e de Valores de Prova Aleatórios

Catarina Cláudia Abreu Gouveia Monteiro

Dissertação
Mestrado em Estatística

2013

Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Combinação de Valores de Prova e de Valores de Prova Aleatórios

Catarina Cláudia Abreu Gouveia Monteiro

Dissertação
Mestrado em Estatística

Orientadores: Professor Doutor Fernando Sequeira
e Professora Doutora Maria de Fátima Brilhante

2013

Agradecimentos

Aos meus orientadores, a Professora Doutora Maria de Fátima Brilhante que aceitou colaborar na orientação, não só pelo apoio fundamental mas também por toda a disponibilidade que teve ao longo deste processo, e ao Professor Doutor Fernando Sequeira que de igual forma facultou apoio, contribuindo sempre com palavras encorajadoras em momentos de maior dificuldade.

Ao Professor Doutor Dinis Pestana, que desde cedo motivou o meu interesse pela Probabilidade e Estatística, tendo sido um dos principais intervenientes no meu percurso académico, com contínua disponibilidade e amizade ao longo destes anos.

À Leonor a quem tanto custa aceitar os longos períodos de trabalho da mãe, esperando ansiosamente pelo fim desta fase.

Ao Sérgio que com muita paciência e amor tanto me apoiou de forma a ultrapassar os obstáculos que encontrei.

Aos meus pais e avó que sempre me apoiaram incondicionalmente e motivaram para na vida, fazer sempre mais e melhor.

Resumo

A meta-análise é uma subdisciplina da estatística que analisa e sintetiza estudos com um objetivo comum, através de uma metodologia que toma em consideração toda a informação disponível. Expõem-se algumas metodologias utilizadas em sínteses meta-analíticas nomeadamente na estimação dos efeitos combinados e com especial ênfase, na combinação de testes independentes com recurso aos valores de prova reportados em estudos primários. Neste tipo de abordagem, os testes de uniformidade são cruciais, uma vez que os valores de prova reportados em estudos independentes, sob a hipótese nula global, formam uma amostra aleatória proveniente duma população uniforme padrão.

Existem porém duas preocupações fundamentais em meta-análise, o número reduzido de valores de prova reportados, que em última análise afetam a potência dos testes combinados e a possível existência de enviesamento na publicação. Consequentemente apresentam-se alguns resultados referentes à ampliação computacional de amostras e alguns métodos que podem ser utilizados de forma a minimizar os efeitos do enviesamento na publicação, como o método *file-drawer*. Realizou-se um estudo por simulação com recurso a este método por forma a compreender melhor os impactos da possível existência de enviesamento na publicação para os métodos de Fisher, Logit, Stouffer e média geométrica, tendo-se concluído que o método de Fisher é provavelmente o método mais robusto perante o viés na publicação.

A utilização de valores de prova de forma a combinar informação tem sido amplamente criticada por alguns autores, neste sentido expõe-se o conceito de valor de prova aleatório sobre a alternativa, bem como uma aplicação desta abordagem de forma a testar uniformidade.

Aborda-se o conceito valor de prova generalizado, uma extensão do valor de prova usual, utilizado na situação de existência de parâmetros perturbadores que compliquem o uso da estatística de teste, apresentando-se dois resultados importantes deste tipo de abordagem.

Palavras-chave: meta-análise, combinação de valores *p*, valores *p* aleatórios, valores *p* generalizados, viés de publicação, método estudos na gaveta, testes de uniformidade.

Abstract

Meta-analysis is a subfield of statistics which analyses and synthesizes studies with a common goal through a methodology which takes in consideration all the available information. We expose some of the methodology used in meta-analysis reports, namely in estimating the combined effect with a special emphasis on combining independent tests through the use of the reported *p*-values from primary studies. In this type of approach, the uniformity tests are crucial once the reported *p*-values from independent studies, under the overall null hypothesis, are observations from a standard uniform random variable.

However, there are two major concerns in meta-analysis. The number of reported *p*-values is usually small therefore affecting the power of combined tests, other issue is the possible presence of publication bias. Thus we present some results on data augmentation and some methods that can be used in order to minimize the effects of publication bias, as the file-drawer method. We conducted a study by simulation using this method in order to better understand the impacts of the possible existence of publication bias applying Fisher, Logit, Stouffer and geometric mean methods. We conclude that Fisher's method is probably the most resistant method in the presence of publication bias.

The use of reported *p*-values in order to combine information has been criticized by some authors, hence we shall expose the concept of random *p*-value under the alternative and a result of such approach in uniformity tests.

We address the concept of generalized *p*-value, an extension of the usual *p*-value, used in the presence of nuisance parameters that compromise the use of test statistics, presenting two important results obtained through this type of approach.

Key-words: meta-analysis, combining *p*-values, random *p*-values, generalized *p*-values, publication bias, file-drawer method, uniformity tests.

Conteúdo

Capítulo 1

Introdução

Desde os primórdios, o Homem procura a obtenção de conhecimento. A curiosidade e a necessidade, motores dessa demanda, associados ao desenvolvimento tecnológico têm possibilitado grandes avanços do conhecimento, tentando saciar a sede de respostas.

A estatística, de forma transversal a todas as áreas científicas, fornece inúmeras ferramentas e metodologias de apoio à investigação, substanciando os resultados dos investigadores. Associada à probabilidade, a estatística apresenta-se assim, como um dos principais ramos da matemática no apoio à investigação, procura do conhecimento e do saber.

Na procura de respostas ou resultados que substanciem (ou não) outros já alcançados, existem diversas questões com as quais os investigadores se deparam, não desejáveis, inerentes aos projetos de investigação. Uma delas é a escassez de dados em análise, que poderá levar a resultados inconclusivos ou contraditórios. Razões intrínsecas ao objeto em estudo ou de teor económico, podem levantar sérios problemas na obtenção de uma amostra com a dimensão desejável de modo a possibilitar a aplicação das metodologias estatísticas convencionais.

É deste modo natural que na procura do conhecimento, tenham sido desenvolvidas técnicas matemáticas mais refinadas, de modo a obter respostas mais credíveis aos olhos da comunidade científica, ou pelo menos, mais substanciadas. Surge desta forma na estatística, a necessidade da análise para além da análise, hoje de forma comum, denominada meta-análise, estudo meta-analítico ou estudo síntese.

O estatístico Karl Pearson terá sido, provavelmente, o primeiro a combinar dados de diferentes estudos médicos utilizando técnicas formais. Em 1904, o seu estudo sobre o efeito preventivo de inoculações contra a febre entérica, terá sido um dos trabalhos pioneiros de maior relevância na história da meta-análise. Do mesmo modo, Tippett (1931), Fisher (1932), Cochran (1937), Mantel e Haenszel (1959), contribuíram desde cedo no desenvolvimento de métodos estatísticos para lidar com a síntese de diversos estudos.

A denominação e definição formal de meta-análise deve-se a um psicólogo Gene Glass (1976), que considera que se está perante a "análise da análise". Tendo-a definindo como sendo a "análise estatística de grandes coleções de resultados de estudos individuais, com o propósito de integrar os achados desses estudos", sublinhando que se trata de uma alternativa mais rigorosa e substanciada do que a narrativa de revisão:

> *"Meta-analysis refers to the analysis of analysis. 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)

A meta-análise pode ser definida como a análise e síntese de estudos com objetivo comum, através de uma metodologia que toma em consideração toda a informação disponível, fundamentada num sólido enquadramento de natureza estatística. Analisar de forma credível e bem substanciada resultados sumarizados de estudos (denominados primários), torna a meta-análise uma ferramenta de utilidade extrema para o trabalho e desenvolvimento científico em diversas áreas, com especial ênfase na medicina, farmacologia, educação e ciências sociais. Como afirma Sequeira (2009), "As ideias da meta-análise são de facto estimulantes, e correspondem a uma evolução natural da estatística".

É importante salientar que podem de facto existir diferenças entre os diversos estudos primários, sobre os quais incide o estudo meta-analítico. Diferentes metodologias amostrais, diferentes covariáveis, entre outras situações que podem ocorrer, mesmo quando existe um objetivo comum aos estudos.

No Capítulo 2 expõem-se algumas metodologias utilizadas em sínteses meta-analíticas, nomeadamente na estimação de efeitos combinados, fornecendo alguns exemplos, de forma a ilustrar não só o conceito subjacente às sínteses meta-analíticas, mas também a importância das mesmas no progresso da investigação científica.

Em sínteses meta-analíticas é habitual combinar testes independentes, utilizando os valores de prova *p* reportados nos estudos primários, que sob a hipótese nula formam uma amostra de observações independentes provenientes de uma população Uniforme padrão. Alguns métodos para combinar testes independentes são apresentados no Capítulo 3, salientando-se que em determinadas situações, a independência dos valores de prova *p* pode não ser assegurada, existindo uma estrutura de dependência entre os valores observados. Apesar de não se terem formalizado métodos para lidar com esta situação, salienta-se que estes existem e são extensões dos métodos aplicados a uma situação de independência.

O viés de publicação, que será abordado no Capítulo 3, é um dos problemas que podem estar associados às sínteses meta-analíticas. Verifica-se que existe uma maior tendência para publicar estudos com resultados estatisticamente significativos,

desprezando outros resultados não significativos. Deste modo, existe o risco de se ter em análise observações enviesadas, que apesar de poderem proporcionar conclusões estatisticamente significativas, poderão ter carácter discutível.

No entanto, existem metodologias para atenuar os possíveis efeitos do viés de publicação. Através do gráfico de funil é possível detetar a existência deste fenómeno, podendo-se recorrer ao método *file-drawer* (estudos na gaveta) que lida com esta situação através do cálculo de quantos estudos não publicados (não significativos), seriam necessários para inverter uma decisão de rejeição da hipótese nula global para uma situação de não rejeição. O viés de publicação e o método *file-drawer* serão objeto de estudo no Capítulo 5, tendo-se realizado um estudo de simulação de forma a tentar compreender melhor este fenómeno e o seu impacto.

Além da possível existência de enviesamento na publicação, existe outra questão preocupante em meta-análise, o número de estudos disponíveis costuma ser habitualmente muito baixo. De forma a lidar com esta situação, pode recorrer-se à geração de valores de prova *p* adicionais, ou *pseudo-p's* por forma a aumentar a dimensão da amostra, bem como a potência dos testes. Tendo em mente a possível existência do enviesamento na publicação, aborda-se a questão do aumento computacional de amostras no Capítulo 3, com resultados obtidos em investigações de Gomes *et al.* (2009), Brilhante *et al.* (2010a) e Brilhante *et al.* (2010b).

Porém, o uso dos valores de prova *p* tem sido amplamente criticado por diversos autores. Kulinskaya *et al.* (2008) defendem que os valores de prova *p* não passam de um indicador de até que ponto o valor observado da estatística de teste é surpreendente, propondo que toda a inferência estatística se faça em termos da evidência em prol da hipótese alternativa. Salientam que de modo a realizar comparações, algo muito comum em sínteses meta-analíticas, os valores de prova *p* usuais não devem ser utilizados, sugerindo o recurso ao valor de prova aleatório sobre a alternativa. Repare-se que os valores de prova *p* reportados em estudos independentes podem ser usados para combinar informação. Recorrendo a esta abordagem, os testes de uniformidade são cruciais uma vez que os valores de prova *p* observados formam uma amostra proveniente de uma população Uniforme padrão, deste modo, apresentam-se investigações de Brilhante (2013) que utilizou o conceito de valor de prova aleatório para testar uniformidade.

Tsui e Weerahandi (1989) introduzem o conceito de valor de prova *p* generalizado, uma extensão do conceito de valor *p* usual, utilizado na situação de existência de parâmetros perturbadores que compliquem o uso da estatística de teste. Ao realizar inferência sobre homogeneidade entre populações, a heterocedasticidade pode ser um problema. Através do valor de prova generalizado, é possível realizar um teste exato de homogeneidade com recurso a valores de prova generalizados. Um resultado ilustrativo da aplicação do conceito de valor de prova generalizado pode encontrar-se em Brilhante (2013), onde se procura testar uniformidade quando a alternativa é uma mistura de Beta(1,2) e Uniforme padrão. Estes resultados serão expostos no Capítulo 4.

Como vemos, são várias as questões associadas a sínteses meta-analíticas que podem ser vistas como problemáticas. Como em qualquer situação, os caminhos não são livres de obstáculos, desde que a crença que motiva a demanda seja substanciada e determinada, estes obstáculos são ultrapassados tendo em vista o ponto de chegada. As metodologias estatísticas desenvolvidas na meta-análise têm fundamentos robustos, com utilidade bastante relevante, sendo neste momento, uma área da estatística em franca expansão.

Capítulo 2

Meta-análise

No seu livro *Baby and Child Care*, o Dr. Benjamin Spock escreveu: "*I think it is preferable to accustom a baby to sleeping on his stomach from the beginning if he is willing.*" Esta afirmação foi incluída na grande maioria das edições deste livro, e nas mais de 50 milhões de cópias vendidas entre as décadas de 50 e 90. O conselho não era pouco habitual na época, de facto, eram vários os pediatras que faziam recomendações semelhantes.

Durante o mesmo período, mais de 100 000 bebés morreram da síndrome de morte súbita infantil (SMSI), também denominada síndrome de morte súbita de lactentes (SMSL), uma doença que costuma atingir bebés aparentemente saudáveis, entre um mês e um ano de vida.

No início da década de 90, investigadores aperceberam-se que o risco de SMSI, diminuía pelo menos 50%, quando os bebés eram deitados de costas em vez de virados para baixo. Os governos de diversos países lançaram iniciativas de campanhas educativas, levando a uma queda acentuada no número de mortes por SMSI.

> *"Advice to put infants to sleep on the front for nearly half a century was contrary to evidence available from 1970 that this was likely to be harmful. Systematic review of preventable risk factor for SIDS, from 1970 would have led to earlier recognition of the risks of sleeping on the front and might have preserved over 10 000 infant deaths in the UK and at least 50 000 in the Europe, the USA and Australasia."*
>
> Gilbert *et al.* (2005)

Referido por Borenstein *et al.* (2009), este exemplo é um dos muitos citados por Chalmers (2005), que enfatizava o facto de ser necessário observar um conjunto de experiências, em vez de serem observados casos particulares. Antes da década de 90, a tarefa de combinar informação proveniente de diversos estudos, era realizada por via da narrativa de revisão: um especialista de uma determinada área, lia diversos estudos realizados sobre um assunto em particular, sumarizava a informação, chegando a uma conclusão final. Esta metodologia é considerada subjetiva (Borenstein *et al.*, 2009),

dado que existem várias limitações inerentes à mesma, desde a decisão de quais os estudos a incluir ou quais os estudos a dar maior ênfase, por exemplo, por serem de maior dimensão. A narrativa de revisão, apesar de proporcionar uma visão global sobre um determinado assunto, não dispõe ferramentas objetivas de forma a poder fornecer conclusões substanciadas. Estes autores salientam que outra limitação da narrativa de revisão é o facto de quanto mais informação se encontra disponível, maior é a dificuldade em obter resultados, dado que a quantidade de informação torna-se difícil de analisar, podendo levar a resultados, até certo ponto, inúteis.

Devido a estas questões, em meados da década de 80 e início da década de 90, os investigadores começaram a afastar-se das narrativas de revisão, adotando narrativas sistemáticas, de forma a possibilitar a reunião de toda a informação disponível mesmo que por vezes contraditória, na tentativa de alcançar respostas mais credíveis e robustas.

Nas revisões sistemáticas um conjunto claro de regras é utilizado de forma a pesquisar estudos e depois determinar quais serão incluídos ou excluídos da análise. Dado que existe um elemento de subjetividade ao adotar critérios de seleção de informação, não se pode afirmar que este método seja totalmente objetivo. Porém, dado que todas as decisões são especificadas de forma clara, o mecanismo da seleção de informação, torna-se transparente. Um elemento chave na narrativa de síntese é a síntese estatística da informação, onde a importância atribuída a cada estudo é consequência de critérios matemáticos, contrariamente à narrativa de revisão (onde a relevância de cada estudo depende do critério de quem realiza a revisão).

Nos estudos primários normalmente são reportadas as dimensões dos efeitos, médias, variâncias, coeficientes de correlação, proporções, razões (riscos relativos, *odds ratio*), entre outros, bem como a significância dos resultados obtidos, através de um teste de hipóteses ou por via dos valores de prova *p*, podendo ser realizada uma análise de variância ou regressão linear multivariada. As metodologias utilizadas na meta-análise, são consideradas extensões das utilizadas nos estudos primários.

Pode-se considerar a existência de dois grandes tipos de análise estatística utilizados em sínteses meta-analíticas: a combinação das magnitudes dos efeitos dos diversos estudos primários e a síntese de resultados de testes, nomeadamente, sobre a magnitude dos efeitos.

Apesar de se abordar mais detalhadamente a metodologia de combinação de testes independentes através dos seus valores de prova *p*, destaca-se a relevância da combinação dos efeitos dado que é uma metodologia muito utilizada em sínteses meta-analíticas.

2.1. Adição de informação em tabelas de contingência 2x2

A combinação de informação recolhida por diversos observadores, eventualmente mesmo com metodologias diferentes, é uma das questões com grande importância em estatística. Tendo sido uma das primeiras áreas da meta-análise a desenvolver-se, muito antes de se ter reconhecido a sua importância, ou mesmo de ter ganho nome próprio.

No que respeita a combinação de informação de várias tabelas 2x2 existem diversos métodos (c.f. Everitt, 1992), apresenta-se seguidamente o método da "raiz do Qui-quadrado" como descrito por Pestana e Velosa (2010).

Considere-se k tabelas 2x2,

	$Tratado$	$Controlo$	
$Sucesso$	a	b	$a+b$
$\overline{Sucesso}$	c	d	$c+d$
	$a+c$	$b+d$	n

1) Para cada tabela

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

é aproximadamente (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 com um grau de liberdade, $Y = X^2$, onde $X \cap Gaussiana(0,1)$.

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 $ad - bc < 0$, ocorre a situação inversa: $\frac{a}{b} < \frac{c}{d}$;

2) Para cada uma das k tabelas, calcula-se: $Y_j(obs.) = \sqrt{X^2_{2,2}(obs.)},\ j = 1, \dots, k.$ Dando-se sinal "+" se $ad - bc > 0$ e sinal "-" se $ad - bc < 0$, de modo a regressar a observações Gaussianas;

3) Soma-se os valores, sendo a soma, o valor observado de uma variável aleatória $\sum_{j=1}^{k} Y_i$, com distribuição amostral $gaussiana\left(0, \sqrt{n}\right)$, donde resulta que: $\frac{1}{\sqrt{k}} \sum_{i=1}^{k} Y_i \cap N(0,1)$.
Deste modo, obtém-se uma regra de decisão clara.

2.2. Combinação de efeitos em meta-análise

Modelo de efeitos fixos

Considere-se k estudos independentes onde T_i é o estimador do parâmetro populacional θ_i, referente ao i-ésimo estudo ou população. Suponha-se que $\hat{\sigma}_i^2(T_i)$ é a variância estimada de $T_i, i = 1, \dots, k$. Habitualmente, T_i tem por base uma amostra aleatória de dimensão n_i proveniente da i-ésima população. Para grandes amostras, T_i tem uma distribuição aproximadamente Normal, com média θ_i e variância $\hat{\sigma}^2(T_i) = \sigma^2_{(\theta_i; n_i)}$. Na grande maioria dos casos, a variância $\sigma^2_{(\theta_i; n_i)}$, depende de θ_i, sendo deste modo desconhecida, utilizando-se $\hat{\sigma}^2(T_i)$, uma estimativa de $\sigma^2_{(\theta_i; n_i)}$. Assuma-se a existência de homogeneidade

$$\theta_1 = \cdots = \theta_k = \theta ,$$

onde θ representa o efeito comum das populações.

Uma estimativa combinada para θ, é dada por uma combinação ponderada (com pesos) dos T_i's,

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}$$

onde w_i é um peso não negativo, atribuído ao efeito do estudo i. Este método geral de combinar linearmente os T_i's, de modo a calcular uma estimativa para o valor médio comum é atribuído a Cochran (1937). Para qualquer escolha dos pesos não estocásticos w_i, $\hat{\theta}$ é um estimador centrado de θ (Definição A.5, Apêndice A), sendo os pesos que fazem com que $Var(\hat{\theta})$ seja mínima, são dados por

$$w_i = \frac{1}{\sigma^2_{(\theta_i; n_i)}} , i = 1, \dots, k.$$

No entanto, os pesos considerados ótimos, são tipicamente desconhecidos, dado que as variâncias $\sigma^2_{(\theta_i; n_i)}$ são desconhecidas, não podendo ser utilizadas no cálculo dos w_i's.

Quando $\sigma^2_{(\theta_i;n_i)}$ é estimado por $\hat{\sigma}^2(T_i)$,

$$\tilde{\theta} = \frac{\sum_{i=1}^{k} T_i/\hat{\sigma}^2(T_i)}{\sum_{i=1}^{k} 1/\hat{\sigma}^2(T_i)},$$

com variância estimada (assintoticamente) por

$$\hat{\sigma}^2(\tilde{\theta}) = \widehat{Var}(\tilde{\theta}) = \frac{1}{\sum_{i=1}^{k} 1/\hat{\sigma}^2(T_i)},$$

sendo

$$\hat{\sigma}(\tilde{\theta}) = \sqrt{\hat{\sigma}^2(\tilde{\theta})}.$$

Para qualquer situação e considerando $\tilde{\theta}$, uma estimativa combinada de θ_i, é possível obter um intervalo de confiança a um nível $(1-\alpha)100\%$ de confiança,

$$LI \approx \tilde{\theta} - z_{1-\frac{\alpha}{2}}\,\hat{\sigma}(\tilde{\theta}), \qquad LS \approx \tilde{\theta} + z_{1-\frac{\alpha}{2}}\,\hat{\sigma}(\tilde{\theta})$$

onde LI e LS são os limites inferior e superior, respetivamente e $z_{1-\alpha/2}$, o quantil $1-\alpha/2$, de uma Gaussiana padrão. Sendo que se o intervalo não contém 0, rejeita-se a hipótese nula: $H_0: \theta = 0$ ao nível de significância α, a favor da alternativa: $H_A: \theta \neq 0$. De forma equivalente, é possível testar esta hipótese rejeitando-se H_0 ao nível de significância α se

$$|Z| = \frac{|\tilde{\theta}|}{\hat{\sigma}(\tilde{\theta})} > z_{1-\frac{\alpha}{2}}.$$

Por fim, se for desejável testar $\theta_1 = \cdots = \theta_k$(homogeneidade), pode-se recorrer a um teste assintótico do Qui-quadrado. Usando $\tilde{\theta}$, este teste é baseado no teste Qui-quadrado, para grandes amostras:

$$X^2 = \sum_{i=1}^{k} \frac{(T_i - \tilde{\theta})^2}{\hat{\sigma}^2(T_i)} = \sum_{i=1}^{k} \frac{{T_i}^2}{\hat{\sigma}^2(T_i)} - \frac{\left[\sum_{i=1}^{k} T_i/\hat{\sigma}^2(T_i)\right]^2}{\sum_{i=1}^{k} 1/\hat{\sigma}^2(T_i)},$$

rejeitando-se H_0, se

$$X^2 > \chi^2_{k-1;1-\alpha}.$$

Havendo o perigo de um dos estudos dominar as conclusões, nomeadamente por ser de maior dimensão, é preferível basear as conclusões em caraterísticas amostrais relativas. Em dados binários, é usual usar riscos relativos (RR) ou razões de vantagens (*odds ratio* – *OR*, também designados por razões de possibilidades). É também frequente uma transformação logarítmica, que ao simetrizar tem o efeito de

proporcionar uma melhor aproximação à Gaussiana para a distribuição amostral do estimador pretendido.

Considere-se como exemplo os dados reportados por Collins *et al.* (1985), provenientes de nove estudos primários sobre o uso de um diurético durante a gravidez na prevenção da pré-eclampsia, reproduzidos na Tabela 2.1.

De forma a ilustrar a metodologia acima descrita, optou-se por basear a análise no cálculo de *odds ratio*, uma característica amostral muito utilizada na área das ciências biomédicas.

Considere-se k tabelas 2x2, cada uma referente ao i −ésimo estudo. Repare-se que para cada estudo, o cálculo do *OR* é dado pela expressão

$$OR = \frac{ad}{bc}.$$

Como já mencionado, é conveniente realizar uma transformação logarítmica. Deste modo,

$$\ln OR = \ln \frac{ad}{bc},$$

e consequentemente a variância é aproximadamente,

$$Var(\ln OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

Sendo

$$\ln LI = \ln OR - z_{1-\frac{\alpha}{2}} \sqrt{Var(\ln OR)}$$

e

$$\ln LS = \ln OR + z_{1-\frac{\alpha}{2}} \sqrt{Var(\ln OR)}$$

os limites inferior e superior respetivamente, de um intervalo de confiança a um nível $(1-\alpha)100\%$ de confiança numa escala logarítmica. De forma a retomar os valores à escala unitária, realiza-se a operação inversa do logaritmo neperiano,

$$OR = \exp\{\ln OR\}$$

e

$$Var(OR) = exp\{Var(\ln OR)\}.$$

Com,

$$LI = \exp\{\ln LI\}$$

e

$$LS = \exp\{\ln LS\},$$

os limites inferior e superior respetivamente, dos intervalos de confiança a um nível $(1-\alpha)100\%$ de confiança na escala pretendida. Esta foi a metodologia utilizada por forma a estimar os valores apresentados dos *OR* e respetivos IC para cada um dos nove estudos apresentados na Tabela 2.1.

Tabela 2.1. Uso de diuréticos durante a gravidez na prevenção da pré-eclampsia.

Estudo		**Tratado**	**Controlo**		***OR***	**IC95%**	***Valor p***
Weseley	PE	14	14	28	1,04	(0,48; 2,28)	0,5417
	$\overline{PE}$	117	122	239			
		131	136	267			
Flowers	PE	21	17	38	0,40	(0,20; 0,78)	0,0036
	$\overline{PE}$	364	117	481			
		385	134	519			
Menzies	PE	14	24	38	0,33	(0,14; 0,74)	0,0039
	$\overline{PE}$	43	24	67			
		57	48	105			
Fallis	PE	6	18	24	0,23	(0,08; 0,67)	0,0035
	$\overline{PE}$	32	22	54			
		38	40	78			
Cuadros	PE	12	35	47	0,25	(0,13; 0,48)	<0,0001
	$\overline{PE}$	999	725	1724			
		1011	760	1771			
Landesman	PE	138	175	313	0,74	(0,59; 0,94)	0,0071
	$\overline{PE}$	1232	1161	2393			
		1370	1336	2706			
Kraus	PE	15	20	35	0,77	(0,39; 1,52)	0,2258
	$\overline{PE}$	491	504	995			
		506	524	1030			
Tervila	PE	6	2	8	2,97	(0,59; 15,07)	0,9056
	$\overline{PE}$	102	101	203			
		108	103	211			
Campbell	PE	65	40	105	1,14	(0,69; 1,91)	0,6982
	$\overline{PE}$	88	62	150			
		153	102	255			

PE – manifestação de qualquer forma de pré-eclampsia.

Por uma questão de coerência com as notações anteriormente utilizadas, denote-se $ln(OR_i), i = 1, \dots, 9$ como T_i e a variância de T_i por $\hat{\sigma}^2(T_i)$ por forma a estimar o efeito combinado dos nove estudos.

Obteve-se

$$\sum_{i=1}^{9} \frac{T_i}{\hat{\sigma}^2(T_i)} = -49{,}862 \ e \sum_{i=1}^{9} \frac{1}{\hat{\sigma}^2(T_i)} = 125{,}283,$$

deste modo,

$$\ln \tilde{\theta} = \frac{\sum_{i=1}^{9} T_i/\hat{\sigma}^2(T_i)}{\sum_{i=1}^{9} 1/\hat{\sigma}^2(T_i)} = -0{,}398$$

e

$$\widehat{Var}(\ln \tilde{\theta}) \approx \frac{1}{\sum_{i=1}^{9} 1/\hat{\sigma}^2(T_i)} = 0{,}008.$$

Obteve-se $Z = 4{,}45$, rejeitando-se a hipótese nula ($H_0: \theta = 0$) ao nível de significância $\alpha = 0{,}05$. Considerando o mesmo nível de significância obteve-se

$$LI \approx -0{,}573 \ e \ LS \approx -0{,}223 ,$$

salientando-se que estes limites encontram-se à escala logarítmica.

Retomando a escala unitária,

$$\tilde{\theta} = 0{,}67 \ e \ \widehat{Var}(\tilde{\theta}) = 1{,}01,$$

com $(0{,}56; 0{,}80)$ o IC a 95% de confiança. Obteve-se $X^2 = 27{,}265$, dado que $\chi^2_{9;0,95} = 16{,}919$, rejeitou-se a hipótese de homogeneidade, ou seja, a amostra não fornece evidência estatisticamente significativa, a um nível de significância $\alpha = 0{,}05$, de que os efeitos provenientes dos diversos estudos sejam iguais.

Observe-se que o valor estimado do efeito combinado é 0,67 e o intervalo de confiança com coeficiente de confiança 0,95 é $(0{,}56; 0{,}80)$. Obteve-se um *OR* inferior a 1, sendo indicativo de que o tratamento é de facto eficaz. 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.

Salienta-se que a importância desta investigação não reside em saber qual o diurético ou dosagem que produz melhores resultados. A questão central é saber se o uso de um diurético durante a gravidez produz efeito na prevenção de pré-eclampsia – sim ou não?

Optou-se por aplicar o método descrito anteriormente para um modelo de efeitos fixos, também denominado "*inverse variance method*". Porém, outro método muito recomendando neste contexto (frequências em tabelas 2x2) é o método de Mantel-

Haenszel (1959), podendo-se também recorrer ao método de Peto, conhecido por "*one-step (Peto) method*" (Borenstein *et al.*, 2009), por forma a calcular uma estimativa combinada para o *OR*. Salienta-se que ambos os métodos recorrem a transformações logarítmicas retomando a escala unitária no fim, sendo que o método de Peto pode ser estendido a um modelo de efeitos aleatórios. Comparativamente ao método que se utilizou ("*inverse variance method"*), estes têm a vantagem de poderem ser utilizados quando se observam tabelas com células vazias, isto é, a 0.

Modelo de efeitos aleatórios

O modelo de efeitos fixos começa por assumir que a dimensão ou tamanho dos efeitos são iguais (homogéneos) entre os diferentes estudos. Na formulação clássica de meta-análise, o modelo de efeitos fixos goza de algum favoritismo, porventura por ser de formulação e implementação menos trabalhosa. No entanto, na meta-análise e não só, pode não existir motivo plausível para assumir homogeneidade dos efeitos entre populações. Podem existir diferenças nos tamanhos dos efeitos subjacentes aos diversos estudos, deste modo fará todo o sentido optar por um modelo de efeitos aleatórios. O objetivo de usar modelos com efeitos aleatórios é tornar as conclusões extensivas a uma maior generalidade de situações.

Repare-se que no modelo de efeitos fixos, o objetivo é utilizar a informação dos estimadores (T_i) do parâmetro populacional (θ_i) proveniente do $i-$ésimo estudo, de forma a estimar o parâmetro combinado ou média global. De forma a obter a estimativa mais precisa do parâmetro combinado (variância mínima), calcula-se uma estimativa ponderada dessa média global onde a ponderação (ou peso) atribuída a cada estudo, é o valor inverso da variância desse mesmo estudo.

De forma a calcular a variância de um estudo no modelo de efeitos aleatórios, é necessário conhecer tanto a variância dentro dos estudos, como τ^2, a variância entre estudos uma vez que a variância total, é dada pela soma destes dois valores.

O parâmetro τ^2 é a variância entre os estudos, por outras palavras, se de alguma forma fosse possível saber o verdadeiro tamanho do efeito para cada estudo e calculássemos a variância desses efeitos (de um número infinito de estudos), τ^2 seria o valor dessa variância.

Uma das formas de estimar τ^2, é através do método dos momentos onde se obtém

$$T^2 = \frac{Q - g.l.}{C},$$

onde, k é o número de estudos, $g.l. = k - 1$ e

$$Q = \sum_{i=1}^{k} w_i T_i^2 - \frac{\left(\sum_{i=1}^{k} w_i T_i\right)^2}{\sum_{i=1}^{k} w_i},$$

com,

$$C = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}$$

No modelo de efeitos fixos, cada estudo é ponderado pelo valor inverso da sua variância, o mesmo verifica-se no modelo de efeitos aleatórios. A diferença é que neste último modelo, a variância inclui a variância dentro dos estudos mais a variância estimada entre os estudos, T^2 (o estimador de τ^2).

É relevante mencionar que de forma a estabelecer um paralelismo entre o modelo de efeitos fixos e o modelo de efeitos aleatórios, utilizar-se-á a mesma notação, distinguindo o modelo de efeitos aleatórios com *.

Neste modelo, supõe-se que os tamanhos dos efeitos subjacentes aos diversos estudos, têm distribuição Gaussiana (identidade distribucional dos efeitos). Uma estimativa combinada para θ, num modelo de efeitos fixos, é dada por uma combinação ponderada dos T_i's:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}\ , w_i = \frac{1}{\hat{\sigma}^2\left(T_i\right)}\ , i = 1, \ldots, k.$$

No modelo de efeitos aleatórios tem-se de forma análoga

$$\hat{\theta}^* = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*},$$

onde os pesos w_i^* são calculados da seguinte forma

$$w_i^* = \frac{1}{\hat{\sigma}^{*2}(T_i)}, \qquad \hat{\sigma}^{*2}(T_i) = \hat{\sigma}^2\left(T_i\right) + T^2.$$

Deste modo, o estimador pode ser calculado com recurso a

$$\hat{\theta}^* = \frac{\sum_{i=1}^{k} T_i / \hat{\sigma}^{*2}(T_i)}{\sum_{i=1}^{k} 1 / \hat{\sigma}^{*2}(T_i)},$$

com variância estimada (assintoticamente) por

$$\hat{\sigma}^2(\hat{\theta}^*) = \widehat{Var}(\hat{\theta}^*) = \frac{1}{\sum_{i=1}^{k} 1/\hat{\sigma}^{*2}(T_i)},$$

sendo

$$\hat{\sigma}(\hat{\theta}^*) = \sqrt{\hat{\sigma}^2(\hat{\theta}^*)}.$$

Para qualquer situação e considerando $\hat{\theta}^*$, uma estimativa combinada de θ_i, é possível obter um intervalo de confiança a um nível $(1-\alpha)100\%$ de confiança com

$$LI \approx \hat{\theta}^* - z_{1-\frac{\alpha}{2}}\,\hat{\sigma}(\hat{\theta}^*), \qquad LS \approx \hat{\theta}^* + z_{1-\frac{\alpha}{2}}\,\hat{\sigma}(\hat{\theta}^*)$$

onde LI e LS são os limites inferior e superior, respetivamente.

De forma equivalente ao modelo de efeitos fixos, é possível testar a hipótese se o efeito combinado é zero ou diferente de zero, rejeitando H_0 ao nível de significância α se

$$|Z| = \frac{|\hat{\theta}^*|}{\hat{\sigma}(\hat{\theta}^*)} > z_{1-\frac{\alpha}{2}}.$$

Por forma a ilustrar os resultados obtidos quando se considera um modelo de efeitos aleatórios e possibilitar a comparação com os resultados calculados anteriormente, utilizou-se os dados da Tabela 2.1 referentes a nove estudos primários sobre o uso de um diurético durante a gravidez na prevenção da pré-eclampsia (Collins *et al.*, 1985).

Obteve-se uma estimativa combinada para o *OR* de 0,60, com $(0{,}40; 0{,}89)$ o IC a 95% de confiança. Repare-se que se obteve um valor combinado inferior ao obtido no modelo de efeitos fixos e um intervalo de confiança com maior amplitude.

Salienta-se que o modelo de efeitos aleatórios apresenta um maior erro padrão do efeito combinado comparativamente ao obtido através do modelo de efeitos fixos. Do mesmo modo, os IC obtidos através do modelo de efeitos aleatórios são mais amplos do que os obtidos através do modelo de efeitos fixos.

As estimativas síntese calculadas para o modelo de efeitos fixos e modelo de efeitos aleatórios são em última análise, médias ponderadas das estimativas provenientes de cada estudo. Sendo que em cada modelo (ou método) essa ponderação é realizada de forma substancialmente diferente. Observe-se a Tabela 2.2, onde se calculou em percentagem o peso relativo (ou contribuição) do efeito de cada estudo no cálculo do efeito combinado, utilizando as ponderações do modelo de efeitos fixos e aleatórios, respetivamente. Considerou-se igualmente relevante, para efeitos de comparação, o cálculo da dimensão amostral relativa à amostra global para cada estudo.

É de facto notório o contraste entre as ponderações do modelo de efeitos fixos e o modelo de efeitos aleatórios atribuídos a cada estudo. Observe-se o contraste entre o peso relativo do estudo de Landesmann (55%) e o estudo de Tervila (1%).

Quando se considera o modelo de efeitos aleatórios, embora a ponderação atribuída ao estudo de Landesman seja a mais elevada, esta não apresenta tantas discrepâncias relativamente aos restantes estudos em análise. De facto, as ponderações são mais equilibradas num modelo de efeitos aleatórios comparativamente a um modelo de efeitos fixos. Não obstante, verifica-se que ambas as estimativas síntese obtidas para os dois modelos, encontram-se mais próximas do efeito do estudo de Landesmann. O que sugere que estudos primários de maior dimensão, poderão ter um papel mais relevante na obtenção de estimativas síntese, independentemente do modelo utilizado, repare-se que o estudo de Landesmann tem uma amostra que representa cerca de 38% da amostra global.

Tabela 2.2. Peso relativo - modelo de efeitos fixos vs modelo de efeitos aleatórios.

	Peso relativo		**Dimensão amostral relativa**
Estudo	**Modelo de Efeitos Fixos**	**Modelo de Efeitos Aleatórios**	
Weseley	5%	11%	4%
Flowers	7%	12%	7%
Menzies	3%	10%	2%
Fallis	3%	8%	1%
Cuadros	7%	12%	26%
Landesman	55%	17%	38%
Kraus	7%	12%	15%
Tervila	1%	4%	3%
Campbell	12%	14%	4%

Que modelo escolher – um modelo de efeitos fixos ou de efeitos aleatórios? Borenstein *et al.* (2009) referem que a escolha de um modelo de efeitos fixos deve estar assente em duas premissas: primeiro deve-se acreditar que todos os estudos incluídos na análise são funcionalmente idênticos, em segundo lugar o objetivo deve ser o de estimar um efeito combinado para a população em questão e não pretender generalizar o resultado obtido para populações mais vastas. Consideram que a grande maioria das

situações não se enquadra nestas condições, preferindo por isso em termos gerais, a utilização de um modelo de efeitos aleatórios.

Estes autores enfatizam o facto de que num modelo de efeitos aleatórios quando o número de estudos é muito pequeno, origina uma estimativa de τ^2 com pouca precisão. Deste modo, embora possa ser o modelo correto para sumarizar determinados resultados, não existe informação necessária para aplica-lo corretamente. Perante uma situação destas a escolha de um modelo é difícil dado que existem várias opções, no entanto qualquer uma delas levantará problemas. Numa situação deste tipo, pode-se optar por um modelo de efeitos fixos, mas neste caso as conclusões não poderão ser generalizadas a uma população mais vasta, outra opção será calcular os efeitos de cada estudo separadamente, não calculando o sumário do efeito. Uma terceira opção é recorrer a uma abordagem Bayesiana onde a estimativa de τ^2 é baseada numa amostra obtida fora dos estudos em análise.

Alguns investigadores têm por hábito iniciar a síntese meta-analítica com um modelo de efeitos fixos, mudando a análise para um modelo de efeitos aleatórios caso o teste de homogeneidade seja estatisticamente significativo. Na opinião de Borenstein *et al.* (2009) este procedimento deve ser fortemente desencorajado. A decisão de usar um modelo de efeitos aleatórios deve ser baseada no entendimento de que todos os estudos partilham o mesmo efeito, ou não, salientando que os testes de homogeneidade muitas vezes têm pouca potência, não devendo ser utilizados para uma tomada de decisão de qual o modelo a adotar.

Thompson e Pocock (1991) referem que a existência de heterogeneidade detetada estatisticamente (ou não), afeta a forma como se interpreta a síntese meta-analítica. Em última análise salientam que qualquer conjunto de estudos é inevitavelmente heterogéneo dado que existem sempre diferenças nos processos de amostragem, seleção dos intervenientes ou políticas de tratamento. Estes autores consideram que em meta-análise é mais realista acreditar que os verdadeiros efeitos variam até certo ponto. Do mesmo modo que Borenstein *et al.* (2009), estes autores consideram que a escolha de um método de efeitos aleatórios não deve depender do resultado de um teste de homogeneidade dado que ao realizar-se este tipo de teste, é necessária prudência na sua interpretação: mesmo que haja um modesto sinal de heterogeneidade genuína, o teste pode não ser estatisticamente significativo, ou seja, a falha em demonstrar heterogeneidade não significa que os estudos sejam de facto homogéneos. Deste modo consideram que é necessário algum cuidado na escolha de um modelo para sintetizar estudos bem como alguma prudência na interpretação dos resultados quantitativos fornecidos pelas sínteses meta-analíticas.

2.3. Homogeneidade em populações Gaussianas

A investigação científica requer inúmeras comparações de efeitos médios, que frequentemente se admitem provenientes de populações Gaussianas, quer na situação em que se tenham medições repetidas, ou não. A comparação simultânea de diversos tratamentos com base em amostras independentes, pode realizar-se com recurso à análise de variância. A análise de variância usa o facto da variância ser o menor de todos os momentos de segunda ordem para verificar a influência que "alterações amostrais" (médias dentro dos grupos) induzem nas somas de quadrados e deste modo testar uma hipótese geral de igualdade de todas as médias (Fisher,1995).

Na teoria clássica, é possível realizar-se a comparação de efeitos médios dos diversos tratamentos sob a exigência de homocedasticidade (as dispersões dos diversos grupos são homogéneas), assumindo-se a igualdade de variâncias nos diversos grupos que estão a ser comparados. Este pressuposto faz sentido na análise de dados obtidos em experiências com planeamento tradicional, dado que é de admitir a existência de um protocolo rígido sobre o que é medido e como se mede. Espera-se que a medição dos efeitos ou da variável resposta com os mesmos equipamentos, resulte em variâncias iguais nos diversos grupos, quando se condiciona à hipótese nula de homogeneidade populacional. Mais ainda, pressupõe-se que os grupos foram constituídos por atribuição aleatória das unidades amostrais, diminuindo o risco de confundimento.

Na meta-análise, onde analisar resultados de diferentes experiências ou estudos com o mesmo objetivo é algo muito comum, estes pressupostos experimentais deixam de fazer sentido. Não é razoável supor que as diferentes experiências seguiram procedimentos amostrais iguais. A heterocedasticidade (as dispersões dos diversos grupos não são homogéneas) é naturalmente uma das questões importantes em estudo, dentro das metodologias estatísticas utilizadas em sínteses meta-analíticas.

Como referido, em sínteses meta-analíticas de modo a testar os efeitos médios de duas populações Gaussianas fará mais sentido pressupor que a escala varia com a localização, não obstante, considerou-se relevante expor igualmente a metodologia para a comparação de efeitos médios em populações Gaussianas com pressuposto de homocedasticidade, recorrendo às demostrações realizadas por Pestana e Velosa (2010).

Considere-se duas amostras aleatórias independentes $\boldsymbol{X_1} = (X_{11}, \dots, X_{1n_1})$ e $\boldsymbol{X_2} = (X_{21}, \dots, X_{2n_2})$, com $X_1 \cap N(\mu_1, \sigma_1)$ e $X_2 \cap N(\mu_2, \sigma_2)$, onde σ_1, σ_2 desconhecidos.

Pretende-se testar se o efeito médio é o mesmo em cada tratamento,

$$H_0: \mu_1 = \mu_2 = \mu \; vs \; H_A: \mu_1 \neq \mu_2.$$

Ao realizar comparações entre duas populações Gaussianas, cujas variâncias são desconhecidas, como já mencionado é necessário ter em conta diferentes situações:

1) Admite-se que há homocedasticidade, isto é que a escala não varia com a localização. Deste modo, considera-se à partida que $\sigma_1 = \sigma_2 = \sigma$, onde σ é desconhecido. Repare-se que a manter a hipótese nula equivale a considerar que as duas populações são homogéneas;

2) Não se admite à partida a existência de homocedasticidade, isto é que a escala varia com a localização. Neste caso tem-se $\sigma_1 \neq \sigma_2$.

Veja-se a primeira situação e admita-se a existência de homocedasticidade. Deste modo, considere-se $\sigma_1 = \sigma_2 = \sigma$ desconhecido.

Recorrendo a S'^2_1 e S'^2_2, estimadores de σ_1^2 e σ_2^2 respetivamente, as amostras fornecem alguma informação sobre σ,

$$S'^2_1 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_{1k} - \bar{X}_1)^2$$

e

$$S'^2_2 = \frac{1}{n_2 - 1}\sum_{k=1}^{n_1}(X_{2k} - \bar{X}_2)^2.$$

É natural pensar que a amostra de maior dimensão tenha mais informação sobre σ^2. Deste modo, faz sentido ponderar os estimadores $S'^2_{X_1}$ e $S'^2_{X_2}$ de σ^2 pelos respetivos números de graus de liberdade, usando-se como estimador de σ^2,

$$S'^2 = \frac{(n_1 - 1)S'^2_1 + (n_2 - 1)S'^2_2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)S'^2_1 + (n_2 - 1)S'^2_2}{n_1 + n_2 - 2}.$$

Ao admitir independência e homocedasticidade introduz-se uma simplificação no cálculo da distribuição amostral de $\frac{(n_1+n_2-2)S'^2}{\sigma^2}$.

De facto S'^2_1 e S'^2_2 são independentes com,

$$\frac{(n_1 - 1)S'^2_1}{\sigma^2} \cap \chi^2_{n_1-1}$$

e

$$\frac{(n_2 - 1)S'^2_2}{\sigma^2} \cap \chi^2_{n_2-1},$$

deste modo,

$$\frac{(n_1-1)S'^2_1+(n_2-1)S'^2_2}{\sigma^2} \cap \chi^2_{n_1+n_2-2}\,.$$

Repare-se que

$$\bar{X}_1 \cap N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \; e \; \bar{X}_2 \cap N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right),$$

com $\bar{X}_1$ e $\bar{X}_2$ independentes por serem calculadas com base em duas amostras independentes. Então sob $H_0: \mu_1 = \mu_2$,

$$\bar{X}_1 - \bar{X}_2 \cap N\left(0, \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right),$$

estandardizando

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \cap Gaussiana(0,1).$$

A média e a variância empírica são estatísticas independentes. A t de Student com n graus de liberdade é o quociente entre uma Gaussiana padrão e a raiz quadrada de um Qui-quadrado com n graus de liberdade dividido pelo seu valor médio (número de graus de liberdade), sendo a Gaussiana e o Qui-quadrado independentes, tem-se a studentização,

$$\frac{\dfrac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}}{\sqrt{\dfrac{(n_1-1)S'^2_1+(n_2-1)S'^2_2}{\sigma^2(n_1+n_2-2)}}.} = \frac{\bar{X}_1 - \bar{X}_2}{S'\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \; |H_0 \; \cap t_{n_1+n_2-2}.$$

Deste modo, de forma a testar a homogeneidade de duas populações Gaussianas, com base em amostras independentes e admitindo à partida homocedasticidade, utiliza-se o teste t para realizar inferência sobre igualdade de valores médios.

Perante a segunda situação, onde se assume a existência de heterocedasticidade (a escala varia com a localização) isto é, à partida não se assume que $\sigma_1 = \sigma_2$, a clássica abordagem que se baseia na aditividade dos Qui-quadrados independentes, deixa de ser possível, sendo necessário recorrer à aproximação de Welch-Satterthwaite.

Considere-se duas amostras aleatórias independentes $\boldsymbol{X_1} = (X_{11}, \dots, X_{1n_1})$ e $\boldsymbol{X_2} = (X_{21}, \dots, X_{2n_2})$, com $X_1 \cap N(\mu_1, \sigma_1)$ e $X_2 \cap N(\mu_2, \sigma_2)$, então

$$\bar{X}_1 \cap N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \; e \; \bar{X}_2 \cap N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right).$$

Se o objetivo for realizar inferência sobre a diferença das médias, o natural será pensar em utilizar

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \cap Gaussiana(0,1).$$

Observe-se que este resultado não é útil, dado que sob H_0, Z não é uma boa estatística de teste por depender dos parâmetros perturbadores σ_1^2 e σ_2^2. Deste modo, é natural recorrer aos estimadores centrados $S_1^2 = \frac{1}{n_1}\sum_{k=1}^{n_1}(X_{1k} - \bar{X}_1)^2$ e $S_2^2 = \frac{1}{n_2}\sum_{k=1}^{n_2}(X_{2k} - \bar{X}_2)^2$ de σ_1^2 e σ_2^2 respetivamente e usar a "studentização"

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

uma vez que a distribuição exata desta variável não parece factível, aproxima-se o quadrado $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ do denominador por uma variável $\sigma^2 \frac{Y_\upsilon}{\upsilon}$, onde $Y_\upsilon \cap \chi_\upsilon^2$, e υ e σ^2 são escolhidos de tal forma que

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

e

$$Var\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) = Var\left(\sigma^2 \frac{Y_\upsilon}{\upsilon}\right) = \frac{2\sigma^4}{\upsilon}.$$

Tem-se que,

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

e de $\frac{(n_1-1)S_1^2}{\sigma^2} \cap \chi_{n_1-1}^2$ e $\frac{(n_2-1)S_2^2}{\sigma^2} \cap \chi_{n_2-1}^2$ segue-se que

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

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

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

e

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

Consequentemente, procede-se à aproximação

$$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\approx\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\sigma^2\frac{Y_\upsilon}{\upsilon}}}=\frac{\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sigma}}{\sqrt{\frac{Y_\upsilon}{\upsilon}}}.$$

Como $\sigma^2=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$, no numerador tem-se uma variável aleatória Gaussiana padrão e no denominador a raiz quadrada de uma variável Qui-quadrado dividida pelo seu número de graus de liberdade, sendo as mesmas mutuamente independentes. Deste modo,

$$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\approx t_\upsilon,$$

onde υ pode ser estimado por

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

sendo que na prática, dado que é difícil ter acesso à tabela de t_υ com υ fracionário, aproxima-se υ por um número de graus de liberdade natural.

Esta abordagem pode ser reformulada por forma a comparar $k\geq 2$ médias (Welch, 1951).

Uma solução exata para testar a homogeneidade entre populações numa situação de heterocedasticidade, será apresentada no Capítulo 4, onde se utiliza um resultado que se considera relevante, com recurso ao conceito de valor de prova generalizado (Tsui e Weerahandi, 1989).

Capítulo 3

Combinação de testes independentes

Independentemente do estudo e dos seus objetivos, uma das questões centrais e de maior enfoque é a significância dos resultados. É uma peça de informação fundamental, muitas vezes utilizada para avaliar a credibilidade das conclusões apresentadas.

Ao realizar um teste de hipóteses, o resultado desse teste será rejeitar a hipótese nula (H_0) se a amostra observada pertence à região crítica, para um determinado nível de significância α, ou não rejeitar, caso a amostra observada não pertence a essa região. Quando se realiza um teste de hipóteses, tudo se resume a afirmar se a hipótese nula é rejeitada, ou não. Deste modo, não se tem em conta se a amostra observada está muito ou pouco distante da fronteira da região crítica, ou se o valor observado da estatística de teste se situa longe ou perto dos limites de rejeição.

O valor de prova *p*, também chamado de valor *p*, nível de significância observado ou nível de significância descritivo, permite uma forma alternativa de reportar o resultado de um teste, ultrapassando-se a escassa e de certo modo, limitada conclusão de rejeitar, ou não, determinada hipótese nula.

Considera-se relevante começar por definir o que se entende por valor de prova *p*, apresentando primeiro o conceito de teste de significância, bem como de estatística de teste e variável de teste. Conceitos basilares para a definição de valor de prova *p* generalizado que se definirá no Capítulo 4.

Teste de significância é a denominação atribuída à prática de tomar uma decisão sobre uma determinada hipótese, tendo por base um valor de prova *p*, sendo algo efetivamente comum especialmente em estudos da área biomédica.

Pearson (1900), Gosset sob o pseudónimo Student (1908) e Fisher (1956) terão sido pioneiros neste tipo de abordagem, que ao contrário de um teste de hipóteses de nível fixo, permite obter uma maior informação e não requer a especificação de um valor nominal, tal como 0,01 ou 0,05. De um modo geral, testar hipóteses utilizando valores *p*, bem como através de intervalos de confiança, permite uma visão mais global e esclarecedora do que testar hipóteses através dos testes de hipóteses de nível fixo convencionais.

3.1. O valor de prova *p*

De modo a abordar de uma forma mais formal o conceito de valor de prova *p*, considera-se relevante definir o que se entende por estatística de teste e variável de teste, conceitos importantes (Weerahandi, 2003) que serão estendidos a outros resultados, apresentados no Capítulo 4.

Considere-se $\boldsymbol{X} = (X_1, \dots, X_n)$, um vetor aleatório, proveniente de uma determinada família de distribuições, parametrizada por um vetor de parâmetros $\boldsymbol{\zeta} = (\theta, \boldsymbol{\eta})$. Onde θ representa o parâmetro de interesse e $\boldsymbol{\eta}$ é o vetor dos parâmetros perturbadores. Considere-se $F_{\boldsymbol{X}}(\cdot\,; \boldsymbol{\eta})$ a função distribuição de $\boldsymbol{X}$ e seja Θ o espaço paramétrico dos possíveis valores de θ.

A hipótese subjacente a um teste especifica subconjuntos de Θ, ou seja, regiões às quais os parâmetro de interesse θ pertence. Seja Ξ o espaço amostral dos valores que $\boldsymbol{X}$ pode tomar e considere-se $\boldsymbol{x} = (x_1, \dots, x_n)$ o valor observado de $\boldsymbol{X}$.

Os níveis de significância observados (valores de prova *p*) são definidos por uma região extrema amostral, que se encontra em Ξ. Tipicamente, uma região extrema corresponde às caudas de uma função distribuição, com limites determinados pelo valor observado de uma variável aleatória.

Uma definição formal de região extrema requer uma ordenação do espaço amostral, de acordo com a magnitude do parâmetro de interesse. Com diversas aplicações, isto pode ser obtido através da utilização de estatísticas de teste, como se define em seguida.

Definição 3.1.1.

A estatística $T(\boldsymbol{X})$, é chamada estatística de teste para θ se tem as seguintes propriedades:

Propriedade 1. A distribuição de $T = T(\boldsymbol{X})$ é livre do parâmetro perturbador, $\boldsymbol{\eta}$.

Propriedade 2. A função distribuição de T, $F_T(t) = P[T \leq t]$, é uma função monótona de θ para todo o t.

Se $P[T > t]$, é uma função não decrescente, de θ, diz-se que T é estocasticamente crescente em θ.

Deste modo, chama-se a uma quantidade aleatória observável, que tenha estas duas propriedades de estatística de teste.

A uma função real de $\boldsymbol{X}$ e de θ, da forma $T = T(\boldsymbol{X}; \theta)$, chama-se variável de teste, se $T(\boldsymbol{X}; \theta_0)$ é uma estatística de teste quando se tem $\theta = \theta_0$.

Representaremos por $t_{obs} = T(\boldsymbol{x}; \theta)$, o valor observado da variável de teste, quando $\boldsymbol{x}$ é o valor observado de $\boldsymbol{X}$. A existência de uma variável de teste, é suficiente para definir medidas de quão bem a amostra suporta determinada hipótese.

De forma a ilustrar os conceitos desta definição, apresenta-se o seguinte exemplo (Weerahandi, 2003).

Distribuição Exponencial

Seja $X_1, \dots, X_n$ uma a.a. de dimensão n, proveniente de uma distribuição Exponencial de parâmetro θ, com função densidade de probabilidade dada por

$$f_{\boldsymbol{X}}(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x > 0$$

onde θ é o valor médio da distribuição.

Considere-se $T = \sum_{i=1}^{n} X_i$, que se sabe ser uma estatística suficiente para θ (Teorema A.3, Apêndice A). Como

$$X_i \cap Exponencial(\theta),$$

tem-se

$$T = \sum_{i=1}^{n} X_i \cap Gama\ (n, \theta),$$

ou seja, T tem função densidade de probabilidade

$$f_T(t) = \frac{1}{\Gamma(n)\theta^n} t^{n-1} e^{-\frac{t}{\theta}}, t > 0.$$

Considerando

$$W = \frac{T}{\theta} \cap Gama(n, 1),$$

então a função distribuição de T é

$$F_T(t) = P\left[W \leq \frac{t}{\theta}\right] = F_W\left(\frac{t}{\theta}\right),$$

onde F_W é a função distribuição de W. Como F_T é uma função não crescente de θ, temos que T é uma estatística de teste, que pode ser usada para testar hipóteses sobre o parâmetro de interesse θ.

Em diversos problemas, a procura de uma estatística de teste está restringida a uma coleção de estatísticas suficientes (Definição A.1, Apêndice A). Sem perda de informação, pode basear-se a inferência estatística, incluindo testes de hipóteses, a um conjunto de estatísticas suficientes mínimas (Definição A.2, Apêndice A). Em situações que envolvam o parâmetro perturbador, $\boldsymbol{\eta}$, pode ser possível encontrar uma única estatística, por exemplo $T(\boldsymbol{X})$ que seja suficiente para θ. Nestes casos, basta que se verifique se T satisfaz a propriedade 2 (Definição 3.1.1) de uma estatística de teste.

Se a distribuição de $\boldsymbol{X}$ tiver parâmetros perturbadores, o problema de encontrar uma estatística de teste ou uma variável de teste torna-se mais complicado. Métodos para lidar com estas questões serão abordados no Capítulo 4.

Valores de prova p

Na literatura sobre testes de significância, o valor p é usualmente definido como a probabilidade de se observar um resultado tão extremo (ou mais) do que o observado. Como sugerido por Weerahandi (2003), suponha-se que é pretendido testar

$$H_0: \theta \leq \theta_0 \; vs \; H_A: \theta > \theta_0 \; .$$

Assuma-se, sem perda de generalidade, que $T(\boldsymbol{X}, \theta)$ é uma estatística de teste ou variável de teste, estocasticamente crescente em θ, isto é, $P(T > t)$ é uma função de θ, não decrescente. É possível verificar que maiores valores de T, podem ser considerados valores extremos da distribuição de T, sob H_0.

Neste momento, é possível definir mais claramente o que se considera por "região extrema do espaço amostral" e por conseguinte, o que é um valor p.

Definição 3.1.2.

Se a variável de teste $T = (\boldsymbol{X}, \theta)$ é estocasticamente crescente em θ, então uma região extrema baseada na amostra, para $H_0: \theta \leq \theta_0$, é o subconjunto $C_{\boldsymbol{x}}$ do espaço amostral, definido da seguinte forma

$$C_{\boldsymbol{x}} = \{\boldsymbol{X} \, \epsilon \, \Xi : T(\boldsymbol{X}, \theta) \geq T(\boldsymbol{x}, \theta)\}.$$

Definição 3.1.3.

Se $C_{\boldsymbol{x}}$ define uma região extrema, o valor de prova p é definido como

$$p = Sup_{\theta \leq \theta_0} P(\boldsymbol{X} \in C_{\boldsymbol{x}} | \theta).$$

Uma vez que $T(\boldsymbol{X}, \theta)$ é a variável de teste, na qual se baseia a região extrema, o valor p pode ser convenientemente calculado da seguinte forma

$$p = P[T(\boldsymbol{X}, \theta) \geq t_{obs} | \theta = \theta_0],$$

desde que T aumente estocasticamente em θ, onde $t_{obs} = T(\boldsymbol{x}, \theta)$.

As funções

$$\pi(\theta) = P[T(\boldsymbol{X}, \theta) \geq t_{obs} | \theta]$$

e

$$\pi_0(\theta) = [T(\boldsymbol{X}, \theta_0) \geq t_{obs} | \theta],$$

são as chamadas funções potência (baseadas na amostra) da variável de teste $T(\boldsymbol{X}, \theta)$ e da estatística de teste $T(\boldsymbol{X}, \theta_0)$, respetivamente. Estas funções são de utilidade relevante, dado que podem ser utilizadas para realizar comparações de regiões extremas alternativas, caso existam.

Salienta-se que se a distribuição de $T(\boldsymbol{X}, \theta_0)$ não depende de parâmetros perturbadores, tem-se normalmente, $\pi(\theta) = \pi_0(\theta)$ quando θ é um parâmetro de localização ou escala.

Dado que a hipótese nula é a hipótese que está a ser testada, perante uma escolha de valores p baseados em variáveis de teste, a variável de teste com menor valor p é preferida em detrimento das restantes (Thompson, 1985). Se existirem duas estatísticas de teste com o mesmo valor p, a que tiver maior função potência (se existir), para todo o $\theta > \theta_0$, é escolhida. No caso das estatísticas de teste, as comparações podem apenas ser feitas por via da função potência $\pi_0(\theta)$.

Uma vez que a otimalidade dos testes mais potentes, em testes de nível fixo, está presente na definição de variável de teste (baseada em estatísticas suficientes) as funções potência não desempenham um papel muito relevante no caso particular dos testes de significância.

Considerando o teste unilateral esquerdo

$$H_0: \theta \geq \theta_0 \; vs \; H_A: \theta < \theta_0,$$

o valor de prova *p*, pode ser calculado de forma análoga

$$p = P[T(\boldsymbol{X}, \theta) \leq t_{obs} | \theta = \theta_0]$$

ou

$$p = P[T(\boldsymbol{X}, \theta) \geq t_{obs} | \theta = \theta_0]$$

se T é estocasticamente crescente ou decrescente em θ, respetivamente.

Considere-se agora o seguinte teste,

$$H_0: \theta = \theta_0 \; vs \; H_A: \theta \neq \theta_0$$

onde θ_0 é uma constante especificada. Suponha-se $T = T(\boldsymbol{X})$, uma estatística de teste que satisfaça as propriedades 1 e 2 da Definição 3.1.1.

Quando a distribuição de T não é simétrica em torno de θ, não existe consenso sobre a definição do valor de prova *p*. Foram discutidas definições alternativas do valor *p* por Gibbons e Pratt (1975), nas quais podem ser baseados testes de significância sobre H_0 (veja-se também Cox e Hinkley, 1974).

Será fornecida uma extensão da Definição 3.1.3 quando existe uma função de T, que tenda a ter valores mais elevados para maiores discrepâncias entre θ e θ_0.

Definição 3.1.4.

Seja T uma estatística de teste para θ e seja C_t um subconjunto de τ, o espaço amostral de T. Suponha-se que C_t tem a seguinte propriedade:

Propriedade 3. Dados quaisquer t e δ fixos, a probabilidade $P[T \in C_t]$ é uma função não decrescente de (i) $\theta - \theta_0$ quando $\theta \geq \theta_0$ e (ii) $\theta_0 - \theta$ quando $\theta \leq \theta_0$, isto é, existe uma função de T que aumenta estocasticamente em $|\theta - \theta_0|$.

Por vezes esta propriedade pode ser muito restritiva para algumas aplicações, sendo que pode ser aligeirada através da propriedade seguidamente enunciada, que é considerada adequada na definição de uma região extrema. No entanto, salienta-se que um teste que tenha a propriedade 3, é igual ou mais potente do que um teste com a propriedade:

Propriedade 3'. Dados quaisquer t e δ fixos, $P[T \in C_t | \theta] \geq P[T \in C_t | \theta_0]$ para todo o $\theta \in \Theta$.

Considera-se uma região extrema para o teste apresentado, um subconjunto do espaço amostral de T que tenha pelo menos a propriedade 3'. O valor de prova p para testar $H_0: \theta = \theta_0 \; vs \; H_A: \theta \neq \theta_0$ é definido como:

$$p = P\left[T \in C_{t_{obs}} | \theta = \theta_0\right],$$

onde t_{obs} é o valor observado da estatística T. A função potência correspondente é

$$\pi_0(\theta; t_{obs}) = P\left[T \in C_{t_{obs}} | \theta\right].$$

Teorema 3.1.1.

Seja T uma variável aleatória contínua com função probabilidade de densidade unimodal f_T, com moda em μ. Seja M o parâmetro de espaço e seja τ o suporte de $f_T(t) = f_T(t; \mu)$. Assuma-se que dado qualquer $w > 0$ tal que, $t = \mu + w \in \tau$ e que $f_T(\mu - k(w)) = f_T(\mu + w)$.

Então, dado $\mu_0 \in M$, a probabilidade $P[-k(w) \leq T - \mu_0 \leq w]$ é uma função não decrescente de $|\mu - \mu_0|$.

Corolário 3.1.1.

Suponha-se que X é uma variável aleatória observável e $Y = X - \theta$ é uma variável aleatória contínua com função densidade de probabilidade unimodal f_Y, que não depende do parâmetro θ. Dado um valor observado x de X, se $\tilde{x}$ é escolhido de tal modo que $f_Y(\tilde{x} - \theta) = f_Y(x - \theta)$, então $P[\tilde{x} - \theta \leq Y \leq x - \theta | \theta = \theta_0]$ é uma função não decrescente de $|\theta - \theta_0|$.

Quando θ é um parâmetro de localização, pode encontrar-se com ajuda do Teorema 3.1.1, uma região extrema baseada em T que satisfaça a propriedade 3 da Definição 3.1.4, desde que T tenha uma distribuição unimodal. Quando θ é um parâmetro de escala, o método pode ser aplicado talvez utilizando transformações logarítmicas. Noutras situações, um valor de prova *p* para o teste bilateral pode ser calculado tendo por base a algumas propriedades da respetiva família de distribuições, como por exemplo a propriedade razão de verosimilhanças monótona.

Seguidamente apresenta-se um exemplo ilustrativo da aplicação do Teorema 3.1.1 no cálculo do valor de prova *p* (Weerahandi, 2003).

Gaussiana uniparamétrica

Suponha-se que $X_1, \ldots, X_n$ formam uma amostra aleatória proveniente de uma população Gaussiana com função densidade de probabilidade

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2},$$

onde μ é o valor médio da distribuição. Pretende-se testar a seguinte hipótese

$$H_0: \mu = \mu_0 \; vs \; H_A: \mu \neq \mu_0,$$

onde μ_0 é um valor especificado de μ.

A função densidade de probabilidade conjunta de $X_1, \dots, X_n$ pode ser escrita como

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^{n} x_i^2/2} e^{-n\mu^2/2 + \mu \sum_{i=1}^{n} x_i}.$$

Pelo critério da factorização, $T = \sum_{i=1}^{n} X_i$ é uma estatística suficiente (Teorema A.1, Apêndice A) para μ. É conhecido que a distribuição de T é Gaussiana com média $n\mu$ e variância n. Consequentemente, T aumenta estocasticamente em μ, logo T é de facto uma estatística de teste. Sob H_0, a distribuição de T/n é uma curva em forma de sino, simétrica em torno de μ_0. Deste modo, recorrendo ao Teorema 3.1.1 os testes de significância para H_0 podem ser baseados na variável aleatória $Y = |\bar{X} - \mu|$ dado que Y tende a tomar valores mais elevados para afastamentos de μ_0 a partir de μ, onde $\bar{X} = \sum_{i=1}^{n} x_i/n$ é a média amostral. O valor de prova *p* para testar H_0 é

$$\begin{aligned} p &= P[|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|] = \\ &= 1 - P\left[-\sqrt{n}(\bar{x} - \mu_0) < Z < \sqrt{n}(\bar{x} - \mu_0)\right] = \\ &= 2\Phi\left(-\sqrt{n}(\bar{x} - \mu_0)\right) = 2\left[1 - \Phi\left(\sqrt{n}(\bar{x} - \mu_0)\right)\right]. \end{aligned}$$

Independentemente das especificidades inerentes a cada teste, verifica-se que o valor *p* é uma medida (na escala $[0,1]$), que mede a evidência que os dados fornecem a favor de H_0. Deste modo, quanto menor for *p*, maior é a evidência em prol da hipótese alternativa (H_A) e quanto maior for *p*, maior será a evidência a favor da hipótese nula (H_0). Salienta-se que Kulinskaya *et al.* (2008) consideram esta interpretação ingénua, apesar de até certo ponto, correta. São de facto mais cautelosos nas ilações que tiram, sobre determinado nível de significância observado. Salientam que o valor de prova *p*, é condicional a uma determinada amostra e deste modo, quando calculado, torna-se apenas relevante para essa experiência em particular. Afirmam que de forma a se poderem realizar comparações, o valor de prova *p*, deve ser encarado como uma variável aleatória. Estes autores vão mais longe na sua interpretação dos valores de prova *p* e consequente metodologia para lidar com estes valores, dado que propõem que toda a inferência estatística seja feita sobre a hipótese alternativa (esta questão será explorada no Capítulo 4). Seguidamente apresentam-se alguns resultados de forma a combinar testes independentes através de valores de prova *p*.

3.2. Combinação de valores de prova

Desde cedo foram desenvolvidas metodologias para combinar resultados de estudos repetidos, esta ideia terá tido início na década de 30, com notáveis contributos de Tippett (1931), Fisher (1932) e Pearson (1933).

Combinar testes independentes é uma das grandes preocupações na meta-análise, uma forma de o fazer, é combinar os valores de prova *p*. Existem diversos métodos para combinar valores *p*, baseados no facto de que estes são independentes e provenientes de populações Uniforme padrão.

Apesar de serem objeto deste capítulo, métodos para combinar testes independentes, salienta-se que este pressuposto de independência pode não ser verificado, podendo existir uma estrutura de dependência entre os valores de prova *p*. Existem métodos alternativos aos que aqui serão descritos, métodos propostos por Hartung (1999) e Makambi (2003) de modo a lidar com uma situação de dependência.

Os métodos de combinação de valores de prova *p*, podem ser classificados em dois grupos, o primeiro recorre-se de propriedades do modelo Uniforme: método de Tippett, Wilkinson, média aritmética e média geométrica. Outro grupo de métodos, recorre a transformações de probabilidades na distribuição dos valores de prova *p*: método de Fisher, Stouffer, Stouffer modificado com pesos (*weighted*) e Logit.

Repare-se que os valores de prova *p* observados, formam uma a amostra proveniente de uma população Uniforme padrão, sob a validade de H_0. De facto, seja T uma estatística de teste absolutamente contínua, com função distribuição F_T, pelo teorema da transformação uniformizante (Teorema A.4, Apêndice A) tem-se,

$$P = F_T(T) \cap Uniforme(0,1).$$

Deste modo, um conjunto observado de valores de prova *p*, $\boldsymbol{P} = (p_1, \dots, p_k)$, de k testes independentes, usando a estatística T, formam uma amostra proveniente de uma população Uniforme padrão. Atendendo a este resultado, a questão de combinar testes independentes está intrinsecamente ligada a testar a uniformidade dos valores de prova *p* observados.

Assumindo que os valores de prova *p* (p_k), são conhecidos para testar H_{0_i} *versus* H_{A_i}, $i = 1, \dots, k$, provenientes de k estudos independentes sobre um objetivo comum, o intuito é alcançar uma decisão sobre a hipótese global (ou hipótese composta):

$$H_0^*: todas\ as\ H_{0_j}\ são\ verdadeiras \quad vs \quad H_A^*: algumas\ das\ H_{A_j}\ são\ verdadeiras.$$

Combinar os valores p_i de modo que $T(p_1, p_2, \dots, p_k)$ seja o valor observado de uma variável aleatória, cuja distribuição amostral sob H_0^* seja conhecida, é algo que se pode tornar relativamente fácil (dependendo da expressão funcional de T), dado que sob H_0^*, p é um valor observado de uma amostra aleatória $\boldsymbol{P} = (p_1, p_2, \dots, p_k)$ proveniente de uma população com distribuição Uniforme padrão. Testar H_0^*, resume-se a testar a uniformidade da amostra dos valores *p* observados.

Existem dois princípios que um método para combinar testes deve garantir, a admissibilidade e a monotonia:

- Um método para combinar testes é dito admissível, se proporciona o teste mais potente (não necessariamente único), contra outra qualquer hipótese alternativa, de combinar testes;

- Um método para combinar testes é dito monótono, se o teste combinado rejeita H_0^* para um conjunto de valores *p*, rejeita igualmente a hipótese nula global para qualquer conjunto menor de valores *p*.

Birnbaum (1954) mostrou que todos os métodos de combinação de testes monótonos são admissíveis e deste modo, ótimos para alguma situação de teste. Os métodos apresentados seguidamente para combinar valores de prova *p*, satisfazem o princípio da monotonia, sendo por isso ótimos para testar alguma situação de teste.

Os trabalhos pioneiros são de Tippett (1931), este sugere que se considere o mínimo da amostra dos valores de prova observados, como estatística de teste. Utilizando o método de Tippett, rejeita-se H_0^* ao nível de significância α, quando

$$\min(P_1, \dots, P_k) = P_{1:k} < \alpha^* = 1 - (1 - \alpha)^{1/k}.$$

Note-se que $P \cap U(0,1)$, logo $P_{1:k_{|H_0^*}} \cap Beta(1, k)$, tendo-se mais genericamente, $P_{r:k_{|H_0^*}} \cap Beta(r, k - r + 1)$. Resultado este utilizado por Wilkinson (1951).

Wilkinson, apresenta um método mais geral do que Tippett, dado que considera a r-ésima estatística ordinal. Neste método, rejeita-se H_0^* ao nível de significância α, quando $p_{r:k} < c$, onde

$$\int_0^c \frac{u^{r-1}(1-u)^{k-r}}{B(r, k-r+1)} du = \alpha\,.$$

Stouffer *et al*. (1949) introduzem a utilização de *scores* gaussianos: o valor Z, baseado no valor *p*, que é definido como

$$Z = \Phi^{-1}(P),$$

sendo Z uma variável Gaussiana padrão, sob a hipótese nula.

Quando os valores *p* são convertidos em valores Z, obtêm-se variáveis independentes e identicamente distribuídas, provenientes de Gaussianas padrão.

Sob H_0^*, a soma dos valores Z, tem distribuição Gaussiana, com valor médio 0 e variância k. Utiliza-se a estatística de teste

$$Z = \sum_{i=1}^{k} \frac{\Phi^{-1}(P_i)}{\sqrt{k}},$$

que é uma variável com distribuição Gaussiana padrão sob a hipótese nula global, podendo deste modo ser comparada com os valores críticos da Gaussiana *standard*.

Uma vez que valores pequenos de valores de prova *p*, correspondem a pequenos (e negativos) *valores* Z, o teste rejeita H_0^* quando Z é menor do que z_α, ou de forma equivalente, $|Z| > z_{1-\alpha}$.

Este método também é conhecido pelo método da transformação Normal inversa e é vastamente utilizado nas ciências sociais, sendo recomendado neste contexto da meta-análise por Mosteller e Bush, 1954.

Um dos métodos mais usados em meta-análise foi proposto por Fisher (1932) na quarta edição do seu famoso tratado *Statistical Methods for Research Workers*.

Este método baseia-se no facto de que a variável $-2 \ln P$ tem uma distribuição Qui-quadrado com dois graus de liberdade, em que P tem uma distribuição Uniforme padrão. Sob H_0^*, a soma de k valores P_i, terá uma distribuição Qui-quadrado com $2k$ graus de liberdade.

O teste rejeita H_0^* quando a estatística de teste

$$F = -2 \sum_{i=1}^{k} \ln\ P_i$$

excede o valor crítico $100(1-\alpha)\%$ de uma Qui-quadrado com $2k$ graus de liberdade.

George (1977) sugere outro método, investigado posteriormente por Mudholkar e George (1979). De forma a combinar $(p_1, \ldots, p_k)$, cada valor p é transformado em $\ln[P/(1-P)]$, que sob H_0^* é uma variável com distribuição Logística, que convenientemente normalizada, é aproximadamente uma distribuição t. A estatística de teste utilizada é

$$G = -\sum_{i=1}^{k} \ln\left[\frac{P_i}{1-P_i}\right]\left[\frac{k\pi^2(5k+2)}{3(5k+4)}\right]^{-1/2}.$$

Deste modo, considera-se que G tem distribuição t com $5k+4$ graus de liberdade. O teste baseado nesta aproximação, rejeita a hipótese nula se G excede o valor crítico $100(1-\alpha)\%$ de uma distribuição t com $5k+4$ graus de liberdade.

É relevante salientar que sob a hipótese nula, $\ln[P/(1-P)]$ pode ser encarada como aproximadamente Normal, com valor médio zero e variância $\pi^2/3$. Outra aproximação que pode ser utilizada em vez de se aproximar a distribuição de G, sob a hipótese nula, à distribuição t é o teste baseado na estatística

$$G^* = -\sum_{i=1}^{k} \ln\left[\frac{P_i}{1-P_i}\right]\left[\frac{3}{k\pi^2}\right]^{1/2},$$

devendo-se rejeitar H_0^* se G^* exceder $z_{1-\alpha}$.

Outro método para combinar valores de prova *p* independentes, é o método da média aritmética. À primeira vista, recorrer a esta medida é algo que se pode considerar intuitivo. No entanto, este método levanta problemas para pequenas amostras. Para grandes amostras, pode ser utilizado o teorema do limite central (Teorema A.5, Apêndice A) de modo a testar H_0^*.

Como abordagem alternativa ao método da média aritmética, é sugerido o método da média geométrica, pois além de se conseguir obter a função densidade de probabilidade da média geométrica, este método proporciona um teste mais potente do que o método da média aritmética (Pestana, 2011).

Considere-se a média geométrica (G_k) dos valores de prova *p*,

$$G_k = \left(\prod_{i=1}^{k} p_i\right)^{1/k}.$$

Pestana (2011), apresenta a função densidade de probabilidade da variável G_k, que sob a hipótese nula global é dada pela expressão

$$f_{G_k}(x) = \frac{k(-kxlnx)^{k-1}}{\Gamma(k)} I_{(0,1)}(x),$$

rejeitando-se H_0^* quando o valor observado da média geométrica, for superior ao quantil (c) de probabilidade $1-\alpha$ da função distribuição, valor esse que pode ser obtido através da resolução da equação

$$\int_0^c \frac{k(-kxlnx)^{k-1}}{\Gamma(k)} dx = 1 - \alpha\,, \qquad 0 < x < 1.$$

De facto existe uma forma mais expedita e computacionalmente preferível, para o cálculo deste quantil (Demonstração A.1, Apêndice A). Observe-se que $-2\sum_{i=1}^{k} \ln P_i$ é uma variável que tem uma distribuição Qui-quadrado com $2k$ graus de liberdade (à semelhança do que sucede no método de Fisher). Dado que,

$$-2\sum_{i=1}^{k} \ln(P_i) = -2ln\left(\prod_{i=1}^{k} P_i\right)$$

torna-se efetivamente mais simples a aplicação deste método, rejeitando-se H_0^* se

$$G_k = \left(\prod_{i=1}^{k} p_i\right)^{1/k} < exp\left\{-\frac{\chi^2_{2k;1-\alpha}}{2k}\right\}.$$

Selecionar um método para combinar estudos independentes, apesar de importante, não é tarefa fácil. Teoricamente, não existe um teste uniformemente mais potente (Definição A.8, Apêndice A). Alguns estudos têm demonstrado que sob diferentes situações, o melhor método varia (Hedges e Olkin, 1985; Loughin, 2004; Pestana 2011).

Hartung *et al.* (2008) referem os estudos de Hedges e Olkin (1985) que sumarizam alguns resultados sobre a performance dos vários métodos de combinação descritos nesta secção (à exceção do método da média geométrica) que tendo como critérios a admissibilidade, monotonia, e a eficiência de Bahadur, concluíram que o teste de Fisher será porventura o melhor a ser utilizado quando não exista uma alternativa em mente. Estes autores (Hartung *et al.*, 2008) referem também investigações de Marden (1991), que introduziu noções de sensibilidade e robustez (*sturdiness*), de forma a comparar a performance dos procedimentos de testes combinados que, tendo como base cinco dos métodos de combinação (Tippett, Wilkinson - para o máximo, Stouffer, Fisher e soma dos valores de prova *p*), concluiu que o método de Fisher revelou ser o melhor.

Alguns autores criticam os métodos de combinação de valores de prova *p* e questionam a sua utilidade. Como amplamente referido por Borenstein *et al.* (2009), mesmo que a magnitude do efeito seja substancial, o valor *p* não será significativo, a não ser que a dimensão da amostra seja adequada. A ausência de significância estatística da magnitude do efeito, não fornece evidência de que o mesmo não esteja presente, quando esta situação ocorre, sugerem a realização de mais estudos, de modo a clarificar os resultados.

Por outro lado, em síntese meta-analíticas quando se combinam valores de prova *p* num único resultado, ignoram-se as dimensões dos estudos primários. Apesar de se

obter uma maior potência nesse único teste, do que nos testes em separado, corre-se o risco de um dos estudos dominar as conclusões, nomeadamente por ser de maior dimensão.

Estes autores consideram que só se devem realizar métodos de combinação de valores de prova *p* em três situações:

- É pretendido testar a hipótese nula de que o tratamento não produz efeito em qualquer um dos ensaios (efeitos são zero);
- Existe apenas informação sobre os valores de prova *p*, sem acesso às respetivas dimensões das amostras;
- Perante estudos de tal forma diversificados (questões populacionais, entre outras características), que a combinação da magnitude dos efeitos não é expressiva.

Hartung *et al.* (2008), salientam que os valores de prova *p* por si próprios, não fornecem tanta informação como as estimativas (e respetivas variâncias) que o originam. No entanto, referem que quando esta informação tão detalhada não se encontra disponível, os métodos de combinação de valores de prova *p*, devem ser utilizados.

Kulinskaya *et al*. (2008), consideram que tendo uma amostra de valores de prova *p* pequenos (significativos), cresce a convicção de que a hipótese nula é de facto falsa. Sendo desejável um método de combinação de evidência que funcione, quer a hipótese nula seja verdadeira ou não. Propõem que toda a inferência estatística se faça em termos da evidência em prol da hipótese alternativa. Estes autores usam transformações estabilizadoras da variância como base do que chamam *"key inferential function"*, cuja função é precisamente permitir o cálculo da evidência em prol da alternativa, usando *scores* gaussianos. No Capítulo 4 abordaremos esta questão de forma um pouco mais detalhada.

Em sínteses meta-analíticas é geralmente preferido basear as conclusões na combinação dos efeitos. Os métodos de combinação de forma a avaliar a magnitude dos efeitos, apresentados neste capítulo são muitas vezes utilizados como complemento à estimativa combinada dos mesmos.

Volta-se a salientar que a independência dos valores *p* por vezes não é assegurada, estes podem encontrar-se de facto, correlacionados. Como indicado por Hartung (1999), devido a restrições na aleatorização dos ensaios, por exemplo, na escolha da inclusão dos mesmos, podem provocar a existência de uma estrutura de correlação entre os valores de prova *p*. O mesmo autor apresenta um método alternativo de forma a lidar com esta situação, utilizando uma extensão do método modificado de Stouffer, onde a dependência dos resultados é parametrizada através de um único coeficiente de correlação. Makambi (2003), utiliza um resultado semelhante, mas aplica-o ao método de Fisher.

3.3. Viés de publicação

Uma das preocupações em meta-análise é a possível existência do fenómeno conhecido por viés de publicação. Verifica-se que existe uma tendência para publicar estudos com resultados estatisticamente significativos, ignorando outros que não tenham obtido resultados com significância estatística. Uma vez que estudos publicados são mais propensos a ser incluídos nas sínteses meta-analíticas, existe o risco considerável de se ter em análise uma coleção de dados enviesados.

Este problema não é exclusivo das sínteses de revisão ou sínteses meta-analíticas, pode ocorrer na narrativa de revisão ou até mesmo quando um profissional de determinada área, ao realizar uma investigação, pesquisa informação sobre a mesma em estudos anteriores, encontrando apenas os que obtiveram resultados significativos. Como consequência, estudos com resultados estatisticamente significativos, também têm maior probabilidade de serem citados noutros com objetivo comum. Repare-se que um aparente excesso de confiança na informação disponível na internet, também se apresenta como fator agravante deste fenómeno. O acesso facilitado à informação por via da rede global, proporciona que se encontrem mais facilmente estudos com resultados estatisticamente significativos, não só por terem sido publicados, mas também por serem mais citados por outros autores.

Ao planear uma revisão sistemática ou síntese meta-analítica, são determinados critérios de inclusão dos estudos primários. Idealmente, o desejável seria incluir todos os estudos que reúnam esses critérios, sendo na prática algo impossível de conseguir. Se os estudos não incluídos, por não estarem publicados, formam uma subamostra aleatória de todos os estudos relevantes, a sua não inclusão resulta em menos informação disponível, maior erro nas estimativas e menor potência dos testes realizados. A possível ocorrência deste fenómeno pode apresentar de facto consequências nefastas ao trabalho dos investigadores.

Algumas linhas de investigação (revistas por Dickersin, 2005) estabeleceram que estudos estatisticamente significativos, são mais prováveis de serem encontrados na literatura publicada. Sendo que qualquer que seja a dimensão da amostra, é mais provável o resultado ser estatisticamente significativo, quando a magnitude do efeito é maior. Esta tendência tem o potencial de produzir grandes enviesamentos na magnitude dos efeitos combinados, especialmente em sínteses meta-analíticas de pequenas dimensões.

Uma investigação em particular foi seguida por diversos autores: Easterbrook *et al.* (1991), Dickersin *et al.* (1992) e Dikersin e Min (1993), tendo identificado grupos de estudos (quando estes foram iniciados), estes autores seguiram ao longo de vários anos, desenvolvimentos desses estudos iniciais, de modo a identificar os que seriam efetivamente publicados. Concluíram que estudos com resultados não significativos

eram menos prováveis de serem publicados (61 - 86%) e quando publicados estavam sujeitos a grandes demoras.

Em que medida a existência do viés de publicação (quando ocorre) pode condicionar os resultados das sínteses meta-analíticas, seja na estimação combinada da magnitude dos efeitos ou na combinação de testes por via dos valores de prova *p*? A única forma de avaliar o impacto deste fenómeno, seria por comparação dos resultados de sínteses meta-analíticas afetadas pelo viés de publicação, com resultados de outras que incluíssem todos os estudos existentes. Dado que conhecer ou obter esses estudos apresenta-se como tarefa impossível, a resposta a esta pergunta é de facto difícil.

Uma das questões que deve ser colocada por quem realiza uma síntese meta-analítica, é se existe ou não evidência de enviesamento na publicação. Um método para detetar a presença do viés de publicação e que aparenta reunir algum consenso por parte de alguns autores (Hartung *et al.*, 2008; Borenstein *et al.*, 2009), é através da análise de um gráfico de funil (*Funnel plot*), uma representação gráfica de pontos (x, y), em que a abcissa é a magnitude do efeito e a ordenada a dimensão da amostra. Em vez da dimensão da amostra, por vezes usa-se o erro padrão como ordenada. Este gráfico 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. Consequentemente, caso não se esteja perante enviesamento na publicação, espera-se observar um perfil de funil com a ponta virada para cima.

Borenstein *et al.* (2009) além de sugerirem o gráfico de funil para detetar a presença (ou não) de enviesamento na publicação, sugerem que se recorra ao método "*Trim and fill*" (Duval e Tweedie, 2000a; Duval e Tweedie, 2000b), caso se esteja perante uma situação de enviesamento na publicação, de forma a corrigi-lo. Este método é um processo iterativo que tem por base o pressuposto que não existindo enviesamento na publicação, as observações do gráfico de funil, não só devem ter a forma de um funil com a ponta virada para cima, mas também devem ser simétricas, isto é, não deve existir maior concentração de observações à esquerda ou à direita. Através deste método é possível gerar observações (estudos) que se supõem em falta, de modo a obter a desejada simetria. Estes autores defendem esta abordagem, uma vez que indicam que a mesma não altera o valor da estimativa, corrigindo apenas a variância. Por outro lado, Kulinskaya *et al.* (2008) desaconselham vivamente a deteção e correção do viés de publicação através do gráfico de funil, dado que consideram esta metodologia pouco segura.

Existem dois tipos de abordagem para lidar com o enviesamento na publicação, tentando atenuar os seus efeitos: métodos de amostragem e métodos analíticos. Os métodos de amostragem são desenhados de modo a que se elimine o enviesamento na publicação tanto quanto possível, através do método pelo qual os estudos são selecionados para inclusão numa análise síntese, tentando por todos os meios obter estudos não publicados sobre o assunto em análise. Este método sugerido por Peto e

colegas (ver Collins *et al.*, 1987), tem tido algumas críticas dado que ao se considerar todos os estudos, a qualidade de alguns pode não ser aceitável, comprometendo os resultados da síntese meta-analítica.

Nos métodos analíticos, surge o método *file-drawer* (estudos na gaveta), também conhecido como "*Rosenthal's Fail-safe N*" (Rosenthal, 1979). Suponha-se que se está perante uma síntese meta-analítica, onde são reportados k estudos, tendencialmente com valores de prova p significativos. A preocupação é que os estudos que obtiveram resultados não significativos, não se encontrem presentes na análise enviesando deste modo os resultados. Considerando o nível usual de significância $\alpha = 0{,}05$, a amostra de valores de prova *p*, será composta por valores tendencialmente pequenos, aumentando deste modo, a probabilidade de se obter uma rejeição da hipótese nula global, quando se aplica um método de combinação de valores de prova *p*, ou analogamente se testa a uniformidade da amostra. Este autor sugere que se calculem quantos valores não significativos, seriam necessários acrescentar à amostra inicial, de forma a inverter a decisão de rejeição da hipótese nula global, para uma decisão de não rejeição. Repare-se que se este número de estudos não incluídos (não significativos) a acrescentar à amostra inicial for elevado, a decisão de rejeição pode ser tomada com maior convicção.

A metodologia proposta por Rosenthal (1979) incide sobre os métodos de combinação de valores de prova apresentados na secção anterior. De forma a descrever o método *file-drawer*, suponha-se que foi utilizado o método de Stouffer (Stouffer *et al.*,1949) numa amostra de k valores de prova independentes. Como descrito anteriormente, este método sugere que se convertam os valores de prova $(P_1, \dots, P_k)$, em *scores* gaussianos $(Z_1, \dots, Z_k)$ definidos por

$$Z_i = \Phi^{-1}(P_i)\,, i = 1, \dots, k\,.$$

Seguidamente utiliza-se

$$Z = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} Z_i\,,$$

de modo a testar a significância da hipótese nula global, rejeitando-se H_0^* se $|Z| > z_{1-\alpha}$.

Suponha-se que H_0^* foi rejeitada e considere-se k_0, o número de estudos não publicados de modo a inverter a decisão de rejeição para uma decisão de não rejeição. Para determinar o valor de k_0, assuma-se que o valor médio dos valores de prova *p* não publicados (ou indisponíveis) nos $k_0$estudos, é 0. Com esta suposição, mesmo que esses k_0 estudos estivessem disponíveis, o valor da soma combinada dos Z_i's mantém-se inalterado,

$$\sum_{i=1}^{k} Z_i = \sum_{i=1}^{k+k_0} Z_i.$$

De facto, é através desta suposição que se consegue determinar analiticamente o valor de k_0. De forma a não rejeitar a hipótese nula global,

$$|Z| \leq z_{1-\alpha} \Leftrightarrow \left(\frac{\sum_{i=1}^{k+k_0} Z_i}{\sqrt{k+k_0}}\right)^2 \leq z_{1-\alpha}^2 \Leftrightarrow k+k_0 \geq \frac{\left(\sum_{i=1}^{k+k_0} Z_i\right)^2}{z_{1-\alpha}^2} \Leftrightarrow$$

$$\Leftrightarrow k_0 \geq -k + \frac{\left(\sum_{i=1}^{k+k_0} Z_i\right)^2}{z_{1-\alpha}^2}.$$

Considere-se agora o método de Fisher (1932), onde se rejeita H_0^* se

$$F = -2\sum_{i=1}^{k} \ln P_i > \chi_{2k;1-\alpha}^2.$$

Dados os valores de prova $P_1, \dots, P_k$, suponha-se que H_0^* foi rejeitada, os valores de prova $P_{k+1}, \dots, P_{k+k_0}$ não publicados invertem a decisão se

$$-2\sum_{i=1}^{k+k_0} \ln P_i \leq \chi_{2(k+k_0);1-\alpha}^2.$$

Pode-se também relaxar o pressuposto de que a soma dos valores desconhecidos (não publicados) é 0, assumindo que

$$P_{k+1} = \cdots = P_{k+k_0} = \tilde{P},$$

deste modo

$$-2\sum_{i=1}^{k+k_0} \ln P_i = -2\sum_{i=1}^{k} \ln P_i - 2\sum_{i=k+1}^{k+k_0} \ln P_i = -2\sum_{i=1}^{k} \ln P_i - 2k_0 \ln \tilde{P},$$

assim, inverte-se a decisão de rejeição para não rejeição se

$$-2\sum_{i=1}^{k} \ln P_i - 2k_0 \ln \tilde{P} \leq \chi_{2(k+k_0);1-\alpha}^2 ,$$

considerando k_0^*, o menor valor que verifica a desigualdade supra indicada.

No método de Tippett (1931), rejeita-se H_0^* ao nível de significância α, quando

$$\min(P_1, \dots, P_k) < 1 - (1-\alpha)^{1/k}.$$

Suponha-se que tendo por base os valores de prova publicados $P_1, \dots, P_k$, H_0^* foi rejeitada. A decisão será alterada para uma decisão de não rejeição, considerando os $P_{k+1}, \dots, P_{k+k_0}$ valores não publicados, se

$$\min\left(P_1, \dots, P_k, P_{k+1}, \dots, P_{k+k_0}\right) \geq 1 - (1-\alpha)^{1/(k+k_0)}.$$

Dado que os valores de prova $P_{k+1}, \dots, P_{k+k_0}$ não publicados, correspondem a estudos não significativos, é possível assumir que qualquer um destes valores de prova, é maior do que o mínimo de $P_1, \dots, P_k$,

$$\min\left(P_1, \dots, P_k, P_{k+1}, \dots, P_{k+k_0}\right) = \min(P_1, \dots, P_k).$$

Deste modo, não é rejeitada a hipótese nula global se

$$\min(P_1, \dots, P_k) \geq 1 - (1-\alpha)^{1/(k+k_0)},$$

resolvendo esta desigualdade em ordem a k_0 tem-se

$$k_0 \geq -k + \frac{\ln(1-\alpha)}{\ln[1 - \min(P_1, \dots, P_k)]}.$$

Borenstein *et al.* (2009), do mesmo modo que tecem críticas à combinação de testes independentes por via dos valores de prova *p* observados, não aconselham o método *file-drawer*. Salientam que os pressupostos para aplicar este método podem não se verificar, como por exemplo o valor médio dos valores de prova *p* não publicados dos k_0 estudos desconhecidos, pode ser diferente de 0. Referem igualmente que este método é aplicado sobre valores de prova *p* observados e métodos de combinação de testes, considerando que o preferível seria aplicar um método que estimasse quantos estudos não publicados, seriam necessários para reduzir a dimensão do efeito a um valor sem expressão. Tendo em mente estas considerações sugerem o método de Orwin e Boruch (1983), conhecido como "*Orwin's Fail-safe N*". Este método é uma variante do método *file-drawer*, que considera tanto a significância estatística dos efeitos como a dimensão dos mesmos.

Kulinskaya *et al.* (2008) sugerem um método onde se recorre à função verosimilhança de uma distribuição Gaussiana truncada, por forma a corrigir o efeito do viés na publicação, calculando o que deveria ser o "verdadeiro" efeito combinado. Dado que este método corrige o efeito combinado, sem certeza de que tenha existido ou não enviesamento na publicação, sugerem outro processo baseado na função verosimilhança de uma Gaussiana truncada. Este segundo método requer o conhecimento não só do número de estudos em falta, bem como da dimensão das amostras de cada um desses estudos. De forma a ultrapassar estas questões, para os segundos valores em falta é sugerido ao utilizador que utilize a média das dimensões dos estudos publicados, de forma a estimar a dimensão dos estudos desconhecidos. Relativamente ao número de estudos desconhecidos, é sugerido ao utilizador que adivinhe este valor.

Apesar de não se ter gerado um gráfico de funil por forma a detetar a existência, ou não, de enviesamento na publicação, utilizaram-se os valores de prova *p* apresentados na Tabela 2.1 relativos ao uso de diuréticos durante a gravidez na prevenção da pré-eclampsia, de modo a ilustrar o potencial do método *file-drawer*.

Este método pode ser aplicado analiticamente, no entanto, atendendo a que os pressupostos da aplicação analítica do método podem não se verificar (o valor médio dos valores de prova *p* não publicados dos k_0 estudos desconhecidos, pode ser efetivamente diferente de 0), optou-se por uma abordagem por via da simulação.

Tendo por base uma amostra de dimensão 9, formada pelos valores de prova *p* observados na Tabela 2.1, realizou-se uma simulação para estimar o número médio de valores de prova não significativos a acrescentar a esta amostra inicial, por forma a alterar a decisão de rejeição da hipótese nula global para uma decisão de não rejeição. Realizou-se este estudo para o método de Fisher, Logit, Stouffer e média geométrica.

A hipótese nula global foi rejeitada pelos quatro métodos em análise, concluindo-se que a amostra observada não apresenta evidência estatística a um nível de significância $\alpha = 0{,}05$, de que todas as hipóteses nulas formuladas nos testes dos estudos primários sejam verdadeiras $\left(H_{0_i}: OR = 1\,, i = 1, \dots, 9\right)$. Apresentam-se os resultados da simulação realizada na Tabela 3.3.1, tendo-se considerado relevante o cálculo do desvio padrão amostral, rácio entre $\bar{k}_0$ e a dimensão da amostra, bem como do coeficiente de variação para efeitos de comparação dos métodos considerados.

Tabela 3.3.1. Resultados obtidos por simulação referentes ao uso de diuréticos durante a gravidez como prevenção da pré-eclâmpsia – método *file-drawer*.

Método	$\bar{k}_0$	**Rácio**	**Desvio padrão**	**Coeficiente de Variação (CV)**
Fisher	75	8,3	25,70	0,34
Média geométrica	72	8,0	25,10	0,35
Stouffer	34	3,8	24,69	0,72
Logit	48	5,3	29,67	0,62

Pelos resultados obtidos considera-se que seria necessário um número expressivo de estudos não significativos a serem incluídos na análise de forma a mudar a decisão de rejeição da hipótese nula global para uma decisão de não rejeição. Observe-se que aplicando o método de Fisher (com menor CV), seriam necessários valores (estudos) adicionais (não significativos), na ordem de 8,3 vezes a dimensão inicial da amostra (9). Por outro lado, aplicando o método de Stouffer, seriam necessários estudos adicionais

(não significativos) na ordem de 3,8 vezes a dimensão inicial da amostra por forma a mudar a decisão de rejeição.

No Capítulo 5, o viés de publicação e o método *file-drawer* serão objeto de um estudo por simulação mais detalhado, de forma a tentar compreender melhor este fenómeno e o seu impacto nas sínteses meta-analíticas.

3.4. Aumento computacional de amostras

Como já referido anteriormente, combinar testes independentes utilizando valores de prova *p* reportados, que sob H_0, formam uma amostra de observações independentes provenientes de uma população Uniforme padrão, é um procedimento usual em meta-análise. Deste modo, combinar testes independentes está intrinsecamente relacionado com testar a uniformidade dos valores de prova *p* observados.

No entanto, o número de valores de prova *p* disponíveis é habitualmente muito baixo. De forma a lidar com esta situação, pode recorrer-se à geração de valores de prova *p* adicionais por forma a aumentar a dimensão da amostra, bem como a potência dos testes.

Gomes *et al.* (2009), sugeriram que se aumentasse computacionalmente a dimensão das amostras, gerando *pseudo-p's*, recorrendo a resultados que incidem sobre estatísticas ordinais e espaçamentos (*spacings*). Quando presente, o viés de publicação favorece o aparecimento de amostras de valores de prova *p* pequenos (originando amostras truncadas à direita), pelo que quando se procede ao aumento computacional de amostras, este fator deve ser tido em consideração. Estes autores, aumentaram as amostras iniciais, com recurso a *pseudo-p's*, provenientes de populações com maior probabilidade (do que a Uniforme) em gerar valores pequenos, na expectativa de aumentar a potência dos testes.

Geraram uma amostra inicial, considerando a família de densidades

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

como alternativa à Uniforme dado que membros desta família, são mais propensos a gerar valores perto de zero. Observe-se que para $m \in [-2,0)$, X_m é uma mistura de uma $Beta(1,2)$ e Uniforme com pesos $-\frac{m}{2}$ e $1 + \frac{m}{2}$, respetivamente, sendo que para $m = 0$, obtém-se a Uniforme padrão (H_0).

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 pseudoaleatórios de X_m é simples e imediata.

A proporção de rejeições de uniformidade, foi determinada por simulação para conjuntos de *pseudo-p's*, quando a amostra inicial $(p_1, \dots, p_n)$, era gerada por uma população da família de densidades supra mencionada.

De forma a aumentar a dimensão inicial das amostras, foi aplicado o seguinte resultado (Deng e George, 1992).

Sejam X e Y variáveis $Uniforme(0,1)$, independentes. A variável aleatória $W = min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right)$ tem suporte $(0,1)$, e para $z \in (0,1)$,

$$P\left[min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \leq z\right] = \int_0^1 zy\, dy + \int_0^1 z\,(1-y)dy = z\,.$$

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

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

e consequentemente, $W \cap Uniforme(0,1)$, sendo Y e W independentes.

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

$$P[X + Y - I[X+Y] \leq z] = P[X+Y \leq z] + P[1 < X+Y \leq 1+z] = z\,.$$

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

$$P[X + Y - I[X+Y] \leq z | Y = y] = max(0, z-y) + min(1+z-y, 1) - (1-y) = z$$

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

Gomes *et al.* (2009) aplicaram apenas os métodos de Tippett e Fisher, tendo concluído que o aumento da amostra inicial, com recurso à família de densidades $f_{X_m}, m \in [-2, 0)$, diminuía a potência do teste de ajustamento Uniforme, contrariamente ao que seria inicialmente espectável. Estes resultados foram explicados posteriormente por Sequeira (2009), devendo-se à generalização do resultado que se enunciou anteriormente.

Sejam X_{m_1} e X_{m_2}, $m_1, m_2 \in [-2,0]$, variáveis aleatórias independentes, provenientes da família com densidade

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

então

$$W_{m_1,m_2} = min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1 - X_{m_1}}{1 - X_{m_2}}\right) \overset{d}{=} X_{\frac{m_1 m_2}{6}} .$$

Deste modo, quando utilizaram variáveis auxiliares Uniformes, Gomes *et al.* (2009) estavam a adicionar mais observações Uniformes à amostra computacionalmente aumentada, contribuindo para esbater as características que contradiziam a uniformidade.

Brilhante *et al.* (2010b) deram continuidade a esta investigação, utilizando os métodos de Tippett e Fisher e apresentando dois estudos de simulação de modo a obter a potência dos testes de uniformidade: em primeiro lugar, estudaram o comportamento dos *pseudo-p's*, no aumento computacional das amostras, utilizando uma população $Beta(1, q), q \in [0{,}5; 3]$, com $H_0: q = 1\ (uniformidade)\ vs\ H_1: q \in [0{,}5; 3]\backslash\{1\}$. Seguidamente, foi usada uma extensão da família de densidades $f_{X_m}, m \in [-2,0)$, utilizada por Gomes *et al.* (2009), de modo a que $m \in [0,2]$ fosse também considerado, com $H_0: m = 0\ (uniformidade)\ vs\ H_1: m \in [-2,2]\backslash\{0\}$.

Ambas as estratégias apresentadas para combinar valores de prova *p*, foram consideradas gratificantes, especialmente quando aplicado o método de Fisher. Com a primeira abordagem, obtiveram bons resultados, em particular com o modelo $Beta(1{,}3)$, onde se favorece a geração de valores *p* próximos de zero (tendo em mente a possível existência do viés de publicação). A segunda estratégia apresentou uma menor potência do que a primeira, tendo obtido os melhores resultados quando $-1 < m < 1$.

Investigações posteriores destes autores (Brilhante *et al.*, 2010a), confirmaram o papel atractor da Uniforme, despoletando a necessidade de investigar os conceitos de valores de prova generalizados (Tsui e Weerahandi, 1989; Weerahandi, 2003) e de valores de prova aleatórios (Kulinskaya *et al.*, 2008).

Capítulo 4

Valores de prova aleatórios e valores de prova generalizados

4.1. Valores de prova aleatórios

O valor de prova *p* é encarado como uma medida de evidência a favor, ou contra, a hipótese nula. No entanto, Kulinskaya *et al.* (2008) chamam a atenção para o facto de que este é obtido de forma condicionada à amostra observada de uma experiência particular, sendo por isso apenas relevante para essa mesma experiência.

Estes autores consideram que a evidência estatística reside na estatística de teste e não nos valores de prova *p*, não obstante, observam que existem vantagens na utilização destes valores em sínteses meta-analíticas, pois não só existe um vasto leque de aplicações para os valores de prova *p*, como é habitual o seu valor numérico ser referido nos estudos primários. Tendo em mente estas vantagens e de forma a comparar valores de prova *p*, e combinar evidência (situação recorrente em sínteses meta-analíticas), deve-se ter em conta as propriedades distribucionais dos valores de prova *p*, encarando os mesmos como variáveis aleatórias.

Quando a estatística de teste tem distribuição contínua, os valores de prova *p* observados formam uma amostra proveniente de uma distribuição Uniforme padrão e quase Uniforme quando a estatística de teste tem distribuição discreta. Não será pois incorreto combinar valores de prova *p*, tendo por base a sua distribuição e assumindo a independência, ou não, das experiências. No entanto, os autores acima observam algo relevante: quando se tem um número de valores de prova *p* (pequenos), cada um deles significativo, cresce a convicção de que a hipótese nula é de facto falsa. Defendem que o desejável é ter uma combinação de evidência que funcione, seja H_0 verdadeira ou falsa. Por outras palavras, a combinação de evidência não pode ser realizada com base no pressuposto de que H_0 seja verdadeira, ou seja, no facto de que os valores de prova *p* observados têm uma densidade retangular. Concluem que os valores de prova *p* usuais se encontram na escala errada para interpretar evidência estatística, propondo que toda a inferência estatística se faça em termos da evidência em prol da alternativa. São

utilizadas funções estabilizadoras da variância como base para a construção do que chamam de"*inferential key functions*", cuja função é precisamente permitir o cálculo da evidência em prol da alternativa usando *scores* Gaussianos.

Estes autores chamam à atenção de que $T = 1{,}645$ é um valor habitualmente utilizado em inferência estatística quando a estatística de teste tem distribuição amostral Gaussiana padrão (correspondente ao quantil $z_{0,95}$). Referem que este valor separa a obtenção de um resultado "significativo" de um "não significativo", estes autores também salientam que não existe diferença entre se obter $t_{obs} = 1{,}644$ ou $t_{obs} = 1{,}646$ uma vez que ao colocar T na escala correta: somando e subtraindo o erro unitário, obtém-se $(0{,}645; 2{,}645)$, um intervalo de confiança de aproximadamente 68%. Logo o facto de $T = 1{,}645$ não é fiável para uma tomada de decisão.

Deste modo, consideram valores observados da estatística de teste próximos de $T = 1{,}645$, "evidência fraca" a favor da alternativa, sendo valores próximos do dobro desse valor, $T = 3{,}3$, "evidência moderada" e próximos do triplo, $T = 5$, "evidência forte" de que H_0 deve ser abandonada para se adotar H_A.

Em teoria, devem utilizar-se grandes amostras de modo a que a estatística de teste ou o estimador tenha distribuição Gaussiana, no entanto para pequenas amostras quando não é possível aplicar o teorema do limite central (Teorema A.5, Apêndice A), utilizam-se funções estabilizadoras da variância, que possibilitam a aproximação à distribuição à Gaussiana.

De forma a ilustrar estes conceitos, considere-se X uma v.a. proveniente de uma população Gaussiana de valor médio μ desconhecido e desvio-padrão $\sigma = \sigma_0$ conhecido (situação pouco habitual, no entanto serve o propósito de simplificar o exposto). Considere-se que se pretende testar

$$H_0: \mu = \mu_0 \; vs \; H_A: \mu > \mu_0 \, .$$

Seja $\boldsymbol{X} = (X_1, \dots, X_n)$ uma amostra aleatória com observações independentes, em que

$$X_i \cap N(\mu, \sigma_0).$$

Tradicionalmente, o estimador de μ é a média amostral. Considere-se deste modo

$$S = \bar{X} = \sum_{i=1}^{n} X_i.$$

Neste caso em particular rejeita-se H_0 quando o valor de S é elevado, observando-se que valores elevados de S mostram evidência a favor da H_A.

Qual será a escala de calibração para a evidência que S fornece contra H_0?

Observe-se que

$$\bar{X}_n \cap N\left(\mu, \frac{\sigma_0}{\sqrt{n}}\right),$$

logo o erro decresce à escala de $1/\sqrt{n}$, pelo que é necessário quadruplicar o esforço da experiência para duplicar a precisão do estimador de μ. Considera-se que a evidência a favor da H_A encontra-se na mesma escala.

Defina-se o efeito $\theta = \mu - \mu_0$, sendo o efeito estandardizado $\delta = \frac{\theta}{\sigma_0} = \frac{\mu - \mu_0}{\sigma_0}$. Para estimar θ: $\hat{\theta} = \bar{X}_n - \mu_0$, H_0 e H_A podem ser baseadas em termos de θ ou δ.

Considerando o teste

$$H_0: \theta = 0 \; vs \; H_A: \theta > 0,$$

a evidência unilateral contra H_0 a favor de H_A, tem que ser uma transformação monótona crescente: $T = T(S)$, da estatística de teste S, para a qual $T \cap N(E[T], 1)$.

A consequência desta transformação T é que a evidência tem sempre uma distribuição Gaussiana com erro 1, facilitando a combinação e (comparação) de evidência estatística.

Tem-se T aproximadamente identificado com o seu valor esperado $\tau = E[T] = \sqrt{n}\frac{\theta}{\sigma_0}$, sendo 1, o erro de T na estimação de τ. Com

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} \cap N(\tau, 1),$$

que é também conhecida por estatística$-Z$.

Vejamos o caso prático: $\mu_0 = 5$ e $\sigma_0 = 5$, $n = 4$ e $\bar{X} = 10$, obtendo-se $T = 2$. Com $n = 36$ e $\bar{X} = 10$, obtém-se $T = 6$, ambos com erro 1. Repare-se que o primeiro indica uma "evidência fraca" e o segundo, "evidência forte" contra H_0, sugerindo que se deve adotar H_A.

Se perante uma $H_A: \mu < \mu_0$, em vez de T (evidência positiva contra H_0), utiliza-se $-T$ (evidência negativa contra H_0).

Este exemplo foi utilizado de forma a ilustrar com alguma simplicidade o conceito relativamente novo, de medir a evidência em prol da hipótese alternativa. Em situações mais complexas, em que σ_0 seja desconhecido e X_i não seja proveniente de uma Gaussiana, ou S (estatística de teste) não tenha uma distribuição Gaussiana ou que dependa de parâmetros perturbadores desconhecidos, estes autores afirmam ser possível escolher a transformação T de forma a estabilizar a variância para 1 e simultaneamente obter uma distribuição aproximadamente Gaussiana, mesmo para pequenas amostras.

Retomando a questão central desta secção, os valores de prova p para um valor observado da estatística S, são obtidos para testar $H_0: \theta = 0 \; vs \; H_A: \theta > 0$, através de $p = P[S \geq s | \theta = \theta_0]$.

Considerando a transformação T, as propriedades desejáveis para a evidência unilateral são:

Propriedade 1. T é função crescente de S (teste unilateral direito);

Propriedade 2. A distribuição de T é Gaussiana, para todos os valores dos parâmetros desconhecidos;

Propriedade 3. A variância, $Var[T] = 1$, para todos os valores dos parâmetros desconhecidos;

Propriedade 4. A evidência esperada $\tau = \tau(\theta) = E_\theta[T]$ é monótona crescente em θ a partir de $\tau(0) = 0$.

Observa-se que para o exemplo simples do modelo gaussiano com variância conhecida, todas as propriedades acima mencionadas são iguais às definidas pela evidência que é fornecida pela estatística $-Z$. Geralmente as propriedades 2 e 4 verificam-se aproximadamente, mesmo no caso de pequenas amostras.

O valor de prova p para um valor observado de $S = s$, é calculado através de $p = P_0[S \geq s]$, onde P_0 é a distribuição sob H_0 de S. Mais: se $T = T(S)$ satisfaz as propriedades 1 a 3, então o valor de prova *p*, também pode ser calculado a partir do valor observado de $T = t$ através de,

$$p = P_0[T \geq t] = \Phi(-t) = 1 - \Phi(t)$$

e deste modo,

$$t = p(t) = \Phi^{-1}(1 - p).$$

O valor observado $T = t$ é uma função monótona do valor de prova *p*. Salienta-se que a diferença entre t e p é que sobre as alternativas, a distribuição do valor p é assimétrica, mudando com a dimensão da amostra, o que torna a interpretação e combinação de evidência difícil.

Definição 4.1.1.

Seja S_0 o resultado da replicação de uma experiência S, sob idênticas condições às da experiência original, onde se mantenha a hipótese nula. Dado $S = s$, defina-se o valor de prova *p* aleatório observado, como: $PV = P[S_0 \geq s] = 1 - F_0(s)$.

Assumindo que a função distribuição de S_0, F_0 é contínua, então para $0 < p < 1$, a função distribuição do valor de prova *p* aleatório é

$$F_{PV}(p) = P[S_0 \geq s] = P[1 - F_0(S) \leq p] =$$
$$= P[F_0(S) \geq 1 - p] =$$
$$= P[S \geq F_0^{-1}(1-p)] =$$
$$= 1 - F_1(F_0^{-1}(1-p)),$$

onde F_1 é a função distribuição de S. Verifique-se que esta definição não requer que F_1, a distribuição da estatística de teste original S, seja a mesma que F_0. Quando tal sucede, $F_1 = F_0$, tem-se $F_{PV}(p) = p$, para $0 < p < 1$ e deste modo, a distribuição do valor de prova *p* aleatório é uma Uniforme padrão, tendo-se uma distribuição diferente para o valor de prova *p* aleatório, quando $F_1 \neq F_0$. Deste modo, salienta-se que condicionalmente à validade de H_0, é reencontrado o conceito usual de valor *p*.

Interpretação dos valores de prova aleatórios

Seja $z_q = \Phi^{-1}(q)$ o quantil de ordem q da Gaussiana padrão, isto é, $q = P[Z \leq z_q]$. Defina-se para cada x, o valor da função densidade de probabilidade da Gaussiana padrão por $\varphi(x) = exp\{-x^2/2\}/\sqrt{2\pi}$. Então para $H_0: \theta = 0\ vs\ H_A: \theta > 0$, o valor de prova aleatório PV, tem as seguintes propriedades:

Propriedade 1. O quantil de ordem q da distribuição de $PV(x)$ é $p_q = p_q(\mu) = \Phi(z_q - \mu)$. Sendo que a notação $p_q = p_q(\mu)$ serve o propósito de enfatizar que o quantil de ordem q depende de μ;

Propriedade 2. O valor esperado do valor de prova aleatório PV é $E_\mu[PV(X)] = \Phi(-\mu/\sqrt{2})$.

Esta propriedade é enunciada dado que é comum descrever-se uma variável aleatória em termos do seu valor médio e desvio padrão. Para o valor de prova transformado $t(PV)$, o valor médio é μ e o desvio padrão 1. No entanto, estas medidas são boas para representar uma variável aleatória, quando a sua distribuição é simétrica (ou quase). A distribuição do valor de prova aleatório sobre a alternativa, é muito assimétrica, deste modo o valor esperado não é uma boa medida representativa da sua distribuição. Como consequência desta propriedade salienta-se um resultado importante, se numa determinada experiência obteve-se um valor de prova $p = 0{,}05$, então a estimativa de μ é $1{,}65$, sendo o estimador de máxima verosimilhança do valor esperado do valor de prova aleatório: $\Phi(-1{,}65/\sqrt{2}) = 0{,}122$. Deste modo, numa repetição da mesma

experiência, os investigadores devem esperar um valor de prova de $0{,}122$. Resultados semelhantes foram reportados por Goodman (1992);

Decorre das propriedades anteriores:

Propriedade 3. O valor esperado do valor de prova aleatório PV é igual ao quantil de ordem q da sua distribuição, onde q é dado por: $q = \Phi\left(\mu\left(\sqrt{2}-1\right)/\sqrt{2}\right)$.

Por exemplo quando $\mu = 1$, o valor esperado é igual ao quantil $q = 0{,}61$ da sua distribuição e quando $\mu = 2$, o valor esperado é igual ao quantil $q = 0{,}81$. Consequentemente, o valor esperado não uma boa medida para representar a distribuição do valor de prova aleatório sobre a alternativa;

Propriedade 4. A função densidade do valor de prova aleatório é (Donahue, 1999):

$$f_{PV}(p) = \frac{\varphi(t(p)-\mu)}{\varphi\left(t(p)\right)} = \frac{\varphi(\Phi^{-1}(1-p)-\mu)}{\varphi\left(\Phi^{-1}(1-p)\right)}\,, 0 < p < 1.$$

O gráfico de f_{PV} para qualquer $\mu > 0$ é côncavo, monótona decrescente e enviesado à direita. Como os gráficos mudam com μ, são difíceis as comparações entre valores de prova aleatórios sobre a alternativa.

Kulinskaya *et al.* (2008) sumarizam estas propriedades dos valores de prova aleatórios e fornecem consequências dessas propriedades, importantes para a interpretação dos valores de prova aleatórios sob a alternativa. Investigações de Dempster e Schatzoff (1965) e mais recentemente de Hung *et al.* (1997), Sackowitz e Samuel-Cahn (1999) e Brilhante (2013), estudaram o valor esperado dos valores de prova aleatórios. Sackowitz e Samuel-Cahn (1999), também consideraram os quantis da distribuição dos valores de prova aleatórios, fornecendo aplicações para os mesmos. Os valores de prova medianos terão sido investigados por Bhattacharya e Habtzghi (2002).

Brilhante (2013) utiliza o valor de prova aleatório para testar a uniformidade, quando a alternativa é uma mistura de $Beta(1{,}2)$ e Uniforme padrão, uma distribuição mais propensa a gerar valores próximos de zero (como mencionado no Capítulo 3).

Ao contrário de Kulinskaya *et al.* (2008) que considera a distribuição do valor de prova aleatório muito assimétrica e deste modo, o valor de prova esperado não é uma boa medida para representar a sua distribuição, alguns autores (Dempster e Schatzoff, 1965; Sackowitz e Samuel-Cahn, 1999) defendem a utilização do mesmo sob a alternativa, como medida da performance de um teste: quanto menor o valor de prova esperado, melhor é o teste. Os mesmos autores apresentam outras razões a favor da utilização do valor de prova esperado: (1) depende da alternativa e não do nível de significância; (2) permite determinar a dimensão da amostra; (3) permite determinar qual a alternativa que o valor observado do valor de prova representa.

Por exemplo, Brilhante (2013) calcula o valor de prova esperado sob a alternativa para o método de Fisher usando a família de densidades

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

Segue-se uma breve descrição da metodologia utilizada.

Pretende-se testar

$$H_0: m = 0\ (uniformidade)\ vs\ \ H_A: m < 0\ (não\ uniformidade).$$

A função densidade de P sob a alternativa é dada por

$$P_\theta[P \leq p] = 1 - F_\theta\left(F_0^{-1}(1-p)\right), 0 < p < 1.$$

Considere-se T^* e T duas variáveis aleatórias independentes, com função distribuição F_θ e F_0 respetivamente,

$$EPV_\theta = P[T^* \geq T] = \int_{-\infty}^{+\infty} P[T^* \geq T | T = t]\, f_T(t) dt = E[1 - F_0(T)],$$

em que EPV_θ denota o valor esperado para o valor de prova aleatório que sob H_0 tem valor esperado 0,05.

A expressão geral da função de distribuição da estatística de teste do método de Fisher T_n, sob a alternativa m é

$$f_P(p) = \frac{f_m[F_0^{-1}(1-p)]}{f_0[F_0^{-1}(1-p)]} = \frac{f_m\left(\chi^2_{2n;1-p}\right)}{f_m\left(\chi^2_{2n;1-p}\right)}, 0 < p < 1,$$

onde f_m é a função densidade de probabilidade de T_n sob a alternativa m.

Com estes resultados, Brilhante (2013) realizou um estudo de simulação, onde calculou o valor de prova esperado sob algumas hipóteses alternativas ($m = -2; -1{,}5; -1;\ 0{,}5$), tendo verificado que para amostras de dimensão inferior a 20, não era esperado observar valores de prova inferiores a 0,05. Concluindo que aplicar o conceito de valor de prova aleatório p à estatística de Fisher e calcular respetivo valor de prova esperado em sínteses meta-analíticas, requer algum cuidado na presença de amostras de pequena dimensão. Sendo aconselhável utilizar os valores esperados de prova com alguma prudência, especialmente na presença de distribuições de P sob a alternativa muito assimétricas.

4.2. Valores de prova generalizados

Tsui e Weerahandi (1989) introduziram o conceito de valor de prova generalizado. Com contributos de Weerahandi (1993), Gamage e Weerahandi (1998), entre outros autores, é realizada uma extensão do conceito de valor *p* usual, o valor *p* generalizado, utilizado na situação de existência de parâmetros perturbadores que compliquem o uso da estatística de teste. Em muitas situações não é de facto possível, ou fácil, encontrar uma estatística de teste, cuja distribuição não dependa de parâmetros perturbadores.

Pretende-se estender a noção de variável de teste, de modo a que seja possível obter valores de prova *p* que não dependam de parâmetros perturbadores. É importante referir que esta extensão, não afeta a interpretação dos valores de prova *p*, descrita no Capítulo 3, a diferença entre um valor *p* convencional e um valor *p* generalizado, reside no método pelo qual se especifica a região extrema amostral.

Considere-se $\boldsymbol{X}$ um vetor aleatório, com função distribuição $F_{\boldsymbol{X}}(\cdot\,;\boldsymbol{\zeta})$, onde $\boldsymbol{\zeta} = (\theta, \boldsymbol{\eta})$ é o vetor de parâmetros desconhecidos, θ representa o parâmetro de interesse e $\boldsymbol{\eta}$, o vetor dos parâmetros perturbadores.

Seja Ξ o espaço amostral de valores possíveis de $\boldsymbol{X}$ e considere-se $\boldsymbol{x} = (x_1, \dots, x_n)$, o valor observado de $\boldsymbol{X}$, em que $\boldsymbol{x} \in \Xi$.

Definição 4.2.1.

Uma variável aleatória da forma $T = T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$ é chamada variável de teste generalizada, se verifica as seguintes propriedades:

Propriedade 1. O valor observado de T quando $\boldsymbol{X} = \boldsymbol{x}$, isto é, $t_{obs} = T(\boldsymbol{x}; \boldsymbol{x}, \theta, \boldsymbol{\eta})$ não depende de parâmetros desconhecidos;

Propriedade 2. Para θ fixo, a função distribuição de T, não depende do parâmetro perturbador $\boldsymbol{\eta}$;

Propriedade 3. Para $\boldsymbol{x}$ e $\boldsymbol{\eta}$ fixos, $P[T \leq t; \theta]$ é uma função monótona de θ para todo o t.

No caso de $\mathrm{P}[T < t]$ ser uma função não decrescente, de θ, diz-se que T é estocasticamente crescente em θ.

Ao lidar com variáveis de teste, é necessário ter a atenção de garantir que o "papel" de $\boldsymbol{X}$ seja o de encontrar a probabilidade da região extrema, sendo o objetivo de $\boldsymbol{x}$, identificar o ponto de fronteira observado, de forma a definir a referida região extrema.

De modo a clarificar esta definição, apresenta-se o seguinte exemplo ilustrativo, apresentado por Weerahandi (2003).

Segundo momento da distribuição Gaussiana

Seja $X_1, \dots, X_n$ uma amostra aleatória proveniente de uma população Gaussiana, com valor médio μ e variância σ^2. Considere-se o segundo momento da distribuição

$$\theta = E(X^2) = \mu^2 + \sigma^2,$$

o parâmetro de interesse. Uma estatística de teste que tenha as propriedades 1 e 2, da Definição 3.1.1 (Capítulo 3), não existe para testar hipóteses sobre θ.

Consequentemente, deve ser encontrada uma variável de teste generalizada T com base nas estatísticas suficientes

$$\bar{X} = \sum_{i=1}^{n} X_i \ e \ S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

nomeadamente, a média e variância amostral. Será verificado que

$$C_x(\mu, \sigma) = \left\{ \boldsymbol{X}: \left(\bar{x} - \frac{s\,(\bar{X} - \mu)}{S} \right)^2 + \frac{s^2\sigma^2}{S^2} \geq \mu^2 + \sigma^2 \right\},$$

é um subconjunto do espaço amostral, com probabilidade livre de parâmetros perturbadores.

Uma vez que as variáveis de teste generalizadas não podem depender de parâmetros desconhecidos, pode basear-se a construção da variável de teste generalizada, nas quantidades aleatórias, também denominadas quantidades ou variáveis pivotais

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \cap N(0,1) \ e \ U = \frac{n\,S^2}{\sigma^2} \cap \chi^2_{n-1},$$

que tendo distribuições conhecidas, não dependem de parâmetros desconhecidos. Deste modo,

$$\theta = \left(\bar{X} - Z\frac{\sigma}{\sqrt{n}} \right)^2 + \sigma^2 = \left(\bar{X} - Z\frac{S}{\sqrt{U}} \right)^2 + \frac{n\,S^2}{U}.$$

Define-se uma potencial variável generalizada de teste $T = (\bar{X}, S)$ como

$$T = \left(\bar{x} - s\frac{Z}{\sqrt{U}} \right)^2 + \frac{n\,s^2}{U} - \theta,$$

onde $\bar{x}$ e s^2 são os valores observados da média e variância amostral, respetivamente.

Observa-se que T é expresso em termos de $\bar{x}, \bar{X}, s$ e S. O lado direito da desigualdade que define a região extrema $C_{\boldsymbol{x}}$ é $T + \theta$.

Pela definição das variáveis aleatórias Z e U, o valor observado de T é $t_{obs} = T(\bar{x}, s) = 0$. Assim, a distribuição de T depende apenas do parâmetro de interesse θ.

Assim, T satisfaz as propriedades 1 e 2 da Definição 4.2.1, onde a propriedade 3 decorre do facto de θ ser o parâmetro de localização da distribuição de T. Por outras palavras, T é estocasticamente crescente em θ. Considere-se

$$R = \left(\bar{x} - s\frac{Z}{\sqrt{U}}\right)^2 + \frac{n\,s^2}{U},$$

a distribuição de R, F_R, não depende de parâmetros desconhecidos. A função distribuição de T, pode ser expressa como

$$F_T(t) = P[T \leq t] = P[R \leq t + \theta] = F_R(t + \theta)\,,$$

pelo que, a função distribuição de T é uma função crescente de θ, verificando-se a propriedade 3.

Assim, T definido por

$$T = \left(\bar{x} - s\frac{Z}{\sqrt{U}}\right)^2 + \frac{n\,s^2}{U} - \theta$$

é de facto, uma variável de teste generalizada para θ.

À semelhança do que foi apresentado no Capítulo 3, considere-se que se pretende testar

$$H_0\colon \theta \leq \theta_0 \; vs \; H_A\colon \theta > \theta_0,$$

onde θ_0 é um valor especificado do parâmetro de interesse, θ. Nos testes de significância usuais, com base na estatística de teste $T(\boldsymbol{X})$, definiu-se a região extrema baseada na amostra,

$$C_{\boldsymbol{x}} = \{\boldsymbol{X} \in \Xi : T(\boldsymbol{X}) \geq T(\boldsymbol{x})\},$$

em que menores valores de T indicam forte evidência contra H_0. De forma a estender esta definição, seja $T = T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$ uma variável de teste generalizada para o parâmetro de interesse, θ.

Assuma-se sem perda de generalidade que T é estocasticamente crescente em θ. Então T tem a propriedade desejável de que maiores valores de T suportam grandes

valores de θ. Deste modo, é possível generalizar a definição de região extrema de modo a que a propriedade desejada seja preservada e o valor de prova *p* generalizado possa ser definido.

Definição 4.2.2.

Seja $T = T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$ uma variável de teste estocasticamente crescente em θ, então o subconjunto definido por,

$$C_{\boldsymbol{x}}(\boldsymbol{\zeta}) = \{\boldsymbol{X} \in \Xi : T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta}) \geq T(\boldsymbol{x}; \boldsymbol{x}, \boldsymbol{\zeta})\},$$

é uma região extrema generalizada para testar $H_0 vs\ H_A$.

Definição 4.2.3.

Se $C_{\boldsymbol{x}}(\boldsymbol{\zeta})$ é uma região extrema generalizada, então

$$PV_G = Sup_{\theta \leq \theta_0} P(\boldsymbol{X} \in C_{\boldsymbol{x}}(\boldsymbol{\zeta}) | \theta)$$

é chamado de valor de prova generalizado para testar H_0.

Seja $T = T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$ uma variável de teste generalizada, estocasticamente crescente em θ, então o valor de prova *p* generalizado pode ser calculado da seguinte forma

$$PV_G = P[T \geq t_{obs} | \theta = \theta_0],$$

onde $t_{obs} = T(\boldsymbol{x}; \boldsymbol{x}, \boldsymbol{\zeta_0})$, com $\boldsymbol{\zeta_0} = (\theta_0, \boldsymbol{\eta})$.

Repare-se que se T for estocasticamente decrescente em θ, o nível observado de significância para testar H_0 é dado por

$$PV_G = P[T \leq t_{obs} | \theta = \theta_0].$$

As funções

$$\pi(\theta) = P[T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta}) \geq t_{obs} | \theta]$$

e

$$\pi_0(\theta) = [T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta_0}) \geq t_{obs} | \theta],$$

são as chamadas funções potência generalizadas, da variável de teste generalizada $T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$ e de $T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta_0})$, respetivamente. Enquanto $\pi_0(\theta)$ pode ser usada para comparar regiões extremas alternativas, $\pi(\theta)$ é útil para realizar proposições gerais sem referência a uma hipótese específica. À semelhança do que foi referido no Capítulo 3, a otimalidade dos testes mais potentes, em testes de nível fixo, encontra-se incorporada na definição de variáveis de teste baseadas em estatísticas suficientes e na definição de

valores de prova *p* generalizados. Deste modo, e como já mencionado anteriormente, a função $\pi_0(\theta)$ não desempenha um papel fundamental em testes de significância.

Independentemente do tipo de teste (unilateral direito ou esquerdo), o valor de prova *p* generalizado, é a probabilidade exata de uma região extrema, que tenha o valor observado na sua fronteira.

Segundo momento da distribuição Gaussiana (continuação)

Considere-se novamente o exemplo que se utilizou de forma a ilustrar a definição de variável de teste generalizada e para o qual se definiu

$$T = \left(\bar{x} - s\frac{Z}{\sqrt{U}}\right)^2 + \frac{n\,s^2}{U} - \theta,$$

onde

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \cap N(0,1)\ e\ U = \frac{n\,S^2}{\sigma^2} \cap \chi^2_{n-1}.$$

Uma vez que se verificou que T é estocasticamente crescente em θ e o valor observado de T é zero, o valor de prova *p* generalizado para o referido teste é

$$PV_G = P[T \leq t_{obs}|\theta = \theta_0] = P\left[\left(\bar{x} - s\frac{Z}{\sqrt{U}}\right)^2 + \frac{n\,s^2}{U} \leq \theta_0\right].$$

À semelhança do que foi apresentado no Capítulo 3, suponha-se que se pretende testar

$$H_0\colon \theta = \theta_0\ vs\ H_A\colon \theta \neq \theta_0,$$

onde θ_0 é um valor particular especificado do parâmetro de interesse. Usa-se uma extensão da Definição 4.2.3 quando exista uma função de T que tenda a ter valores mais elevados para maiores discrepâncias entre θ e θ_0.

Seja $T = T(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$ uma variável de teste generalizada para testar θ e seja $C_t(\boldsymbol{\zeta})$ um subconjunto de τ, onde τ denota o espaço amostral dos valores possíveis que T pode tomar. Define-se o valor de prova generalizado com base em $C_t(\boldsymbol{\zeta})$ para testar H_0 se tiver a seguinte propriedade:

Propriedade 4. Dados quaisquer t e δ fixos, a probabilidade $P[T \in C_t(\boldsymbol{\zeta})]$ é uma função não decrescente de (i) $\theta - \theta_0$ quando $\theta \geq \theta_0$ e (ii) $\theta_0 - \theta$ quando $\theta \leq \theta_0$, isto é, existe uma função de T que aumenta estocasticamente em $|\theta - \theta_0|$.

Como referido no Capítulo 3, por vezes esta propriedade pode ser considerada muito restritiva sendo possível utilizar uma propriedade paralela à propriedade 3' da Definição 3.1.4, em vez da propriedade 4 supra mencionada.

Se existir um conjunto $C_t(\boldsymbol{\zeta})$ com a propriedade 4, então esse conjunto é considerado uma região extrema para testar H_0. O valor de prova p generalizado para testar H_0 vs H_A baseado na região extrema $C_{t_{obs}}$ é definido como

$$PV_G = P\left[T \in C_{t_{obs}}(\boldsymbol{\zeta}) \,|\theta = \theta_0\right],$$

onde $t_{obs} = T(\boldsymbol{x}; \boldsymbol{x}, \boldsymbol{\zeta})$ é o valor observado de T. A função potência generalizada é

$$\pi(\boldsymbol{x}; \theta) = P\left[T \in C_{t_{obs}}(\boldsymbol{\zeta}) \,|\theta\right].$$

Seguem-se algumas aplicações relevantes do conceito de valor de prova generalizado apresentado nesta secção.

4.2.1. Um teste exato de homogeneidade

Como referido no Capítulo 2, considera-se importante analisar o tratamento da comparação de efeitos médios, no caso de aquisição de dados com variâncias distintas de grupo para grupo, de modo a que a inferência realizada no âmbito da meta-análise, faça sentido. Sendo na sequência desta questão, que surge uma aplicação importantíssima dos valores de prova *p* generalizados, uma vez que através da utilização dos valores de prova *p* generalizados, é possível obter um teste exato para a comparação de efeitos médios perante uma situação de heterocedasticidade.

Um teste baseado no valor de prova generalizado pode ser obtido para comparar diversos efeitos médios com base em amostras independentes. Em particular, quando se comparam os efeitos médios de duas populações, encontramo-nos perante o problema de Behrens-Fisher, que é o de testar e obter um intervalo de confiança, para a diferença entre as médias de duas populações Gaussianas, numa situação de heterocedasticidade, com base em amostras independentes.

De forma a ilustrar o potencial da utilização do valor de prova generalizado em problemas que envolvam parâmetros perturbadores, Tsui e Weerahandi (1989), consideram o problema de Behrens-Fisher, que pode ser formulado da seguinte forma.

Considere-se duas amostras aleatórias independentes $\boldsymbol{X_1} = \left(X_{11}, \dots, X_{1n_1}\right)$ e $\boldsymbol{X_2} = \left(X_{21}, \dots, X_{2n_2}\right)$, com $X_1 \cap N(\mu_1, \sigma_1)$ e $X_2 \cap N(\mu_2, \sigma_2)$.

Pretende-se testar

$$H_0: \mu_1 - \mu_2 \leq 0 \; vs \; H_A: \mu_1 - \mu_2 > 0$$

com base nas estatísticas suficientes: $\bar{X}_1, \bar{X}_2, S_1^2$ e S_2^2, que são estimadores de μ_1, μ_2, σ_1^2 e σ_2^2, respetivamente.

Tem-se

$$\bar{X}_1 \cap N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right), \bar{X}_2 \cap N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right), \frac{n_1 S_1^2}{\sigma_1^2} \cap \chi^2_{n_1} \; e \; \frac{n_2 S_2^2}{\sigma_2^2} \cap \chi^2_{n_2}.$$

Considere-se deste modo que as variáveis aleatórias $\bar{X}_1, \bar{X}_2, S_1^2$ e S_2^2 são independentes e sejam $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2, \boldsymbol{x_1}$ e $\boldsymbol{x_2}$ os valores observados de $\bar{X}_1, \bar{X}_2, S_1^2, S_2^2, \boldsymbol{X_1}$ e $\boldsymbol{X_2}$, respetivamente.

Neste problema, o parâmetro de interesse é $\theta = \mu_1 - \mu_2$ e $\boldsymbol{\eta} = (\sigma_1^2, \sigma_2^2)$ é o vetor dos parâmetros perturbadores. S_1^2 e S_2^2 são duas variâncias amostrais, estimadores centrados de σ_1^2 e σ_2^2, respetivamente.

Defina-se $T(\boldsymbol{X_1}, \boldsymbol{X_2}; \boldsymbol{x_1}, \boldsymbol{x_2}, \theta, \boldsymbol{\eta})$ do seguinte modo

$$T(\boldsymbol{X_1}, \boldsymbol{X_2}; \boldsymbol{x_1}, \boldsymbol{x_2}, \theta, \boldsymbol{\eta}) = (\bar{X}_1 - \bar{X}_2)\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{-1/2}\left(\frac{\sigma_1^2 s_1^2}{n_1 S_1^2} + \frac{\sigma_2^2 s_2^2}{n_2 S_2^2}\right)^{1/2}.$$

O valor observado de T é $t_{obs} = \bar{x}_1 - \bar{x}_2$. Para $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ observados, $T(\boldsymbol{X_1}, \boldsymbol{X_2}; \boldsymbol{x_1}, \boldsymbol{x_2}, \theta, \boldsymbol{\eta})$ não depende de $\boldsymbol{\eta}$.

Tem-se que $Z \cap N(0,1)$ e considere-se U_1 e U_2,

$$U_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \cap \chi^2_{n_1 - 1}$$

$$U_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \cap \chi^2_{n_2 - 1}$$

em que Z, U_1 e U_2 são independentes. Desta forma, uma potencial variável de teste generalizada para θ pode ser

$$T = T(\boldsymbol{X_1}, \boldsymbol{X_2}; \boldsymbol{x_1}, \boldsymbol{x_2}, \theta, \boldsymbol{\eta}) = Z\left(\frac{s_1^2(n_1 - 1)}{U_1 n_1} + \frac{s_2^2(n_2 - 1)}{U_2 n_2}\right)^{1/2},$$

tendo-se que T é estocasticamente crescente em θ (Lehmann, 1986).

Conclui-se que T é efetivamente uma variável de teste generalizada, dado que depende não só de $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ mas também de $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ observados, verificando deste modo as propriedades enunciadas na Definição 4.2.1 (variável de teste generalizada).

Assim, o valor de prova *p* generalizado pode ser definido por

$$PV_G = P[T \geq \bar{x}_1 - \bar{x}_2 | \mu_1 = \mu_2] = P\left[Z\left(\frac{s_1^2(n_1-1)}{U_1 n_1} + \frac{s_2^2(n_2-1)}{U_2 n_2}\right)^{1/2} \geq \bar{x}_1 - \bar{x}_2\right],$$

onde a hipótese de igualdade de médias é rejeitada, se o valor de prova *p* generalizado for baixo.

Considerando um teste bilateral, o valor de prova *p* generalizado é dado pela expressão

$$PV_G = P\left[Z^2\left(\frac{s_1^2(n_1-1)}{U_1 n_1} + \frac{s_2^2(n_2-1)}{U_2 n_2}\right) \geq (\bar{x}_1 - \bar{x}_2)^2\right].$$

Este problema pode ser generalizado a uma situação de mais de duas populações (Hartung *et al.*, 2008).

Tsui e Weerahandi (1989), compararam os resultados obtidos com a utilização do valor de prova *p* generalizado, aos obtidos pela solução Bayesiana do problema de Behrens-Fisher (Johnson e Weerahandi, 1988), concluindo que a solução para este problema, do ponto de vista frequêncista e Bayesiano, é igual. Salientam que o valor de prova *p* generalizado, como função da amostra observada, obtido desta forma, não tem necessariamente uma distribuição Uniforme (Kempthorne e Folks, 1971). Deste modo, este resultado deve ser utilizado apenas no âmbito dos testes de hipóteses.

4.2.2. Valores *p* generalizados quando a alternativa à uniformidade é uma mistura de $Beta(1,2)$ e Uniforme

Como mencionado no Capítulo 3, os valores de prova *p* reportados em estudos independentes, podem ser usados para combinar informação. Se se recorrer a esta abordagem de combinação de informação, os testes de uniformidade são cruciais uma vez que os valores de prova *p* observados formam uma amostra proveniente de uma população Uniforme padrão. Existem porém duas preocupações fundamentais em meta-análise: o número reduzido de valores de prova reportados (que em última análise afetam a potência dos testes combinados) e a possível existência do viés de publicação, que pode originar amostras enviesadas.

Tendo em mente estas preocupações, referiu-se no Capítulo 3 as investigações de Gomes *et al.* (2009), Brilhante *et al.* (2010a) e Brilhante *et al.* (2010b), que estudaram os efeitos do aumento computacional de amostras em testes de uniformidade,

considerando a família das variáveis aleatórias X_m, com função densidade de probabilidade,

$$f_m(x) = \left(mx + 1 - \frac{m}{2}\right) I_{(0,1)}(x), m \in [-2,0],$$

como alternativa à uniformidade, isto é, para $m \in [-2,0]$, uma mistura de $Beta(1,2)$ e Uniforme padrão.

Usando a família de densidades acima, Brilhante (2013) explora o conceito de valor de prova generalizado para testar

$$H_0: m = 0\ (uniformidade)\ \ vs\ \ H_1: m < 0.$$

Repare-se que o objetivo é testar uniformidade contra uma densidade mais provável de gerar valores próximos de zero (consequência da eventual presença do viés de publicação).

Apesar de não existirem parâmetros perturbadores envolvidos no problema, o conceito de valor de prova generalizado pode ser aplicado. Idealmente deve-se basear qualquer estatística de teste em estatísticas suficientes, no entanto esta autora, não tendo obtido pelo teorema da factorização, uma solução não trivial na obtenção de uma estatística suficiente para m (Teorema A.1, Apêndice A), aplica o teorema da transformação uniformizante (Teorema A.4, Apêndice A) e utiliza o método proposto por Iyer e Patterson (2002) de forma a construir valores de prova *p* generalizados para testar a uniformidade. Deste modo, definiu a variável de teste generalizada

$$T = T(\boldsymbol{X}; \boldsymbol{x}, m) = m - \frac{V + 2\sum_{i=1}^{n} \ln x_i}{n\,(1 - \bar{x})},$$

onde V é uma quantidade pivotal invertível e

$$V = -2\sum_{i=1}^{n} \ln[F_m(X_i)] \sim \chi^2_{2n}.$$

Deste modo, o valor de prova generalizado é definido como,

$$PV_G = P[T \leq 0 | m = 0] = 1 - P\left[V \leq -2\sum_{i=1}^{n} \ln x_i\right].$$

Sendo a potência deste teste, dada pela expressão

$$\pi(\boldsymbol{x}; m) = P[T(\boldsymbol{X}; \boldsymbol{x}, m) \leq 0 | m] = 1 - P\left[V \leq mn(1 - \bar{x}) - 2\sum_{i=1}^{n} \ln x_i\right].$$

Observe-se que T é uma generalização do método de Fisher na situação específica de testar $H_0: m = 0$ (resultado expectável dada a forma como os testes foram obtidos),

sendo testes igualmente potentes para $m = 0$. Salienta-se que apesar de não existir um melhor método de combinação, estudos de simulação demonstram que o método de Fisher é o que tem apresentado melhor performance na maioria das situações.

Deste modo, o método proposto por Brilhante (2013) destaca-se, não só pelas razões acima mencionadas, mas também porque apresenta uma ligeira vantagem em relação ao método de Fisher, no sentido em que é permitido testar a não uniformidade: $H_0: m = m_0, m_0 \neq 0$ se desejável.

Capítulo 5

Estudo de simulação

O viés de publicação é consequência da tendência que existe em publicar maioritariamente, estudos em que se tenham obtido resultados estatisticamente significativos, sendo o nível de significância usualmente considerado $\alpha = 0{,}05$. Como resultado deste tipo de "censura", estudos com valores de prova *p* superiores ou iguais a $0{,}05$ tendem a não ser publicados, logo, não considerados nas sínteses meta-analíticas. Como consequência deste fenómeno, tem-se o enviesamento das amostras de valores de prova *p* a serem analisados nos métodos de combinação em sínteses meta-analíticas, dando origem a resultados que podem ser questionáveis.

No Capítulo 3 descreveram-se alguns métodos de forma a identificar e minimizar, quando presente, os efeitos do viés de publicação. Dado que é objeto mais detalhado deste estudo a combinação de valores de prova *p* independentes, considera-se o método *file-drawer* o mais adequado para estudar os efeitos do viés de publicação no que diz respeito à combinação de valores de prova *p*, aplicando os métodos descritos na secção 3.2.

Deste modo realizou-se um estudo por simulação, tendo por objetivo uma melhor compreensão do possível impacto do viés de publicação. A abordagem por via da simulação permite ultrapassar algumas questões levantadas pelos críticos do método *file-drawer* (Borenstein *et al.*, 2009). Nomeadamente, no que diz respeito aos pressupostos realizados anteriormente, valor médio dos valores de prova desconhecidos igual a 0 ou igualdade dos valores de prova desconhecidos, pressupostos estes utilizados no cálculo dos estimadores do número de estudos desconhecidos (não significativos) a incluir na amostra, de forma a mudar uma decisão de rejeição da hipótese nula global para não rejeição.

Repare-se que ao ser tomada uma decisão de rejeição, considerando uma amostra de 4 valores de prova *p* observados, se 1 estudo não publicado (não significativo) for suficiente para inverter essa decisão, esta pode ser considerada pouco sólida ou até certo ponto, pouco credível. No entanto, se forem necessários 8 estudos não publicados de forma a inverter a decisão de rejeição, pode-se considerar que esta decisão pode ser tomada com maior convicção, dado que seria necessário obter um número de valores de

prova *p* (não significativos) muito superior à dimensão da amostra observada. Salienta-se que fica ao critério do investigador, quando confrontado com um determinado número de estudos não publicados a incluir na análise, de forma a inverter a decisão de rejeição, se esse número é muito expressivo ou não, atendendo à natureza do estudo este pode chegar à conclusão de que é muito pouco provável (por variadas razões), existirem mais k_0 estudos não publicados (e não significativos) sobre o assunto em análise.

Com o método *file-drawer* (estudos na gaveta), pretende-se estimar qual o número médio de estudos estatisticamente não significativos ($\geq 0{,}05$), logo desconhecidos, que a serem incluídos na amostra inicial, inverteriam a decisão de rejeição para não rejeição da hipótese nula global. Partindo de amostras iniciais que levem à rejeição da hipótese nula global, considerou-se importante avaliar a relevância dessas amostras conterem ou não, apenas valores de prova *p* significativos ($\leq 0{,}05$).

O facto de serem incluídos em sínteses meta-analíticas, preferencialmente valores de prova *p* inferiores a $0{,}05$, não exclui que valores superiores a $0{,}05$ também sejam considerados. Tendo em vista esta questão, as simulações realizadas visam duas situações: amostras iniciais compostas apenas por valores inferiores a $0{,}05$ e amostras iniciais onde se podem encontrar valores superiores a $0{,}05$, mas onde se obtenha à partida, um resultado de rejeição, quando aplicados os métodos de combinação de valores de prova *p* considerados nesta simulação: método de Fisher, Stouffer, Logit e média geométrica.

Realizaram-se simulações para dimensões de amostras iniciais (k) de $4,\ 5,\ 6,\ 10,\ 20,\ 30,\ 40, 50$ e 100 observações, realizando-se 5000 réplicas para cada esquema de simulação. Identificou-se k_0 como o número de observações (valores *p*) superiores ou iguais a $0{,}05$ a incluir na amostra inicial, de modo a inverter a decisão de rejeição para não rejeição, sendo pretendido estimar a média e desvio padrão amostral de k_0.

Considerou-se relevante o cálculo do coeficiente de variação para cada um dos m métodos aplicados e respetiva dimensão da amostra inicial k, para efeitos de comparação entre métodos e esquemas de simulação, bem como interpretação da estimativa da média,

$$CV_{m,k} = \frac{\sqrt{S_{k_0}^{m,k}}}{\bar{k}_0^{m,k}}.$$

Foi considerado igualmente importante, calcular um rácio (r) entre a média das observações a acrescentar à amostra inicial e a dimensão da mesma (k), para cada um dos m métodos,

$$r_{m,k} = \frac{\bar{k}_0^{m,k}}{k}.$$

5.1. Esquemas de simulação

Consideraram-se três esquemas de simulação: um com a amostra inicial composta exclusivamente por valores inferiores a $0{,}05$ e outros dois com amostras iniciais que à partida obtenham uma decisão de rejeição da hipótese nula global, mas que possam conter valores superiores a $0{,}05$.

É possível resumir as principais caraterísticas dos esquemas considerados na Tabela 5.1.1.

Tabela 5.1.1. Principais características dos esquemas de simulação realizados.

	A amostra inicial contém apenas observações $< 0{,}05$	**A amostra inicial é igual para todos os métodos**	**As observações: p_{k+k_0}* são iguais para todos os métodos**
Esquema A	Sim	Sim	Sim
Esquema B	Não	Sim	Sim
Esquema C	Não	Não	Não

* p_{k+k_0} - observações não significativas ($\geq 0{,}05$) acrescentadas à amostra inicial.

De forma a possibilitar comparações entre métodos, considerou-se importante que as amostras iniciais e observações a acrescentar fossem iguais.

No entanto, dado que se estavam a criar condições muito específicas na obtenção das amostras iniciais (rejeição da hipótese nula global para todos os métodos), considerou-se o esquema C, que realiza a simulação com amostras iniciais e p_{k+k_0} valores acrescentados, diferentes de método para método.

Descrevem-se em seguida os esquemas implementados neste estudo de simulação.

Esquema A: *Amostra inicial composta exclusivamente por observações inferiores a* $0{,}05$

Geração da amostra inicial

1) Gera-se uma amostra $\boldsymbol{P} = (p_1, \dots, p_k)$, proveniente de uma população $U(0; 0{,}05)$;
2) Aplicam-se os quatro métodos de combinação de valores de prova p a $\boldsymbol{P}$. Considera-se $\boldsymbol{P}$ a amostra inicial, caso se obtenha uma situação de rejeição da hipótese nula global H_0^* para todos os métodos de combinação considerados. Caso contrário, a amostra obtida não é considerada e repete-se 1);

Inversão da decisão

3) Acrescenta-se a $\boldsymbol{P}$ uma observação p_{k+1}, proveniente de uma população $U(0{,}05; 1)$ e aplicam-se os quatro métodos de combinação a $(p_1, \dots, p_k, p_{k+1},)$, $k_0 = k_0 + 1$;

   Observação: Caso um (ou mais) dos métodos de combinação altere a decisão de rejeição para não rejeição pára-se o processo para esse(s) método(s), continuando o processo para os restantes.

4) Repete-se 3) até inverter a decisão de rejeição para não rejeição, em todos os testes considerados;

Repete-se o esquema 5000 vezes, calcula-se a média e desvio padrão amostral dos valores observados de n_0, para amostras iniciais de dimensão $k = 4, 5, 6, 10, 20, 30, 40, 50, 100$.

Esquema B: *Amostra inicial pode conter observações superiores a 0,05 e os métodos são aplicados em conjunto*

Geração da amostra inicial

1) Gera-se uma amostra $\boldsymbol{P} = (p_1, \dots, p_k)$, proveniente de uma população $U(0{,}1)$;

2) Aplicam-se os quatro métodos de combinação de valores de prova p a $\boldsymbol{P}$. Considera-se $\boldsymbol{P}$ a amostra inicial, caso seja obtido o resultado de rejeição da hipótese nula global em todos os métodos de combinação considerados. Caso contrário, a amostra obtida não é considerada e repete-se 1);

Inversão da decisão

3) Acrescenta-se à amostra inicial: $(p_1, \dots, p_k)$ uma observação p_{k+1}, proveniente de uma população $U(0{,}05; 1)$ e aplicam-se os quatro métodos de combinação a $(p_1, \dots, p_k, p_{k+1},)$, $k_0 = k_0 + 1$;

   Observação: Caso um (ou mais) dos métodos de combinação altere a decisão de rejeição para não rejeição pára-se o processo para esse(s) método(s), continuando o processo para os restantes.

4) Repete-se 3) até inverter a decisão de rejeição para não rejeição, em todos os testes considerados.

Repete-se o esquema 5000 vezes, calcula-se a média e desvio padrão amostral dos valores observados de n_0, para amostras iniciais de dimensão $k = 4, 5, 6, 10, 20, 30, 40, 50, 100$.

Esquema C: *Amostra inicial pode conter observações superiores a 0,05 e cada método é aplicado individualmente*

Geração da amostra inicial

1) Gera-se uma amostra $\boldsymbol{P} = (p_1, \dots, p_k)$, proveniente de uma população $U(0{,}1)$;

2) Aplica-se um dos métodos de combinação de valores de prova p a $\boldsymbol{P}$, considerando-se $\boldsymbol{P}$ a amostra inicial, caso seja obtido o resultado de rejeição da hipótese nula global nesse método considerado, caso contrário a amostra obtida não é considerada e repete-se 1);

Inversão da decisão

3) Acrescenta-se à amostra inicial: $(p_1, \dots, p_k)$ uma observação p_{k+1}, proveniente de uma população $U(0{,}05; 1)$ e aplica-se o método de combinação a $(p_1, \dots, p_k, p_{k+1},)$, $k_0 = k_0 + 1$;

   Observação: caso se altere a decisão de rejeição para não rejeição, o processo termina para o método aplicado.

4) Repete-se 3) até inverter a decisão de rejeição para não rejeição, no método considerado;
5) Realizam-se os passos 1) a 5) para os restantes métodos.

Repete-se o esquema 5000 vezes, calcula-se a média e desvio padrão amostral dos valores observados de n_0, para amostras iniciais de dimensão $k = 4, 5, 6, 10, 20, 30, 40, 50, 100$.

5.2. Resultados e algumas considerações

Apresentam-se seguidamente os gráficos com os resultados obtidos pelas simulações. Para informação mais detalhada sobre os resultados, convida-se o leitor a consultar o Apêndice B.

Gráfico 5.2.1. $\bar{k}_0$ - Esquema A

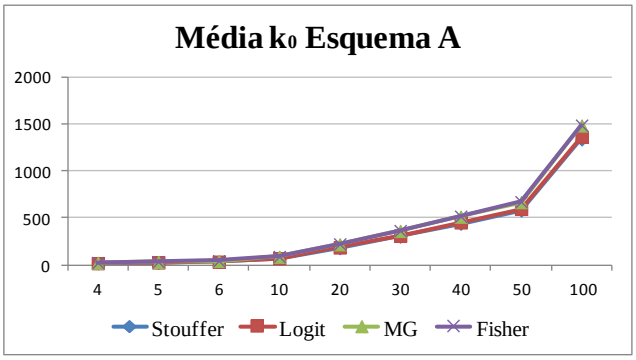


Gráfico 5.2.2. Rácio - Esquema A

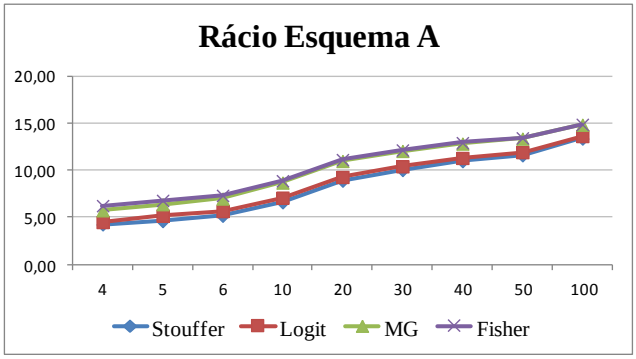


Gráfico 5.2.3. $\bar{k}_0$ - Esquema B

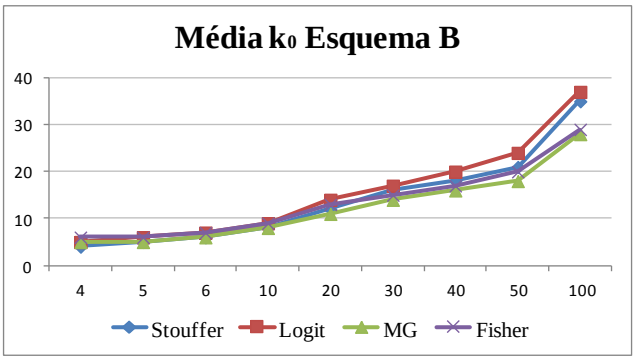


Gráfico 5.2.4. Rácio - Esquema B

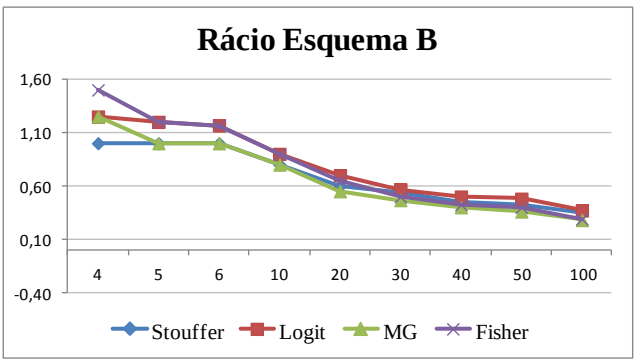


Gráfico 5.2.5. $\bar{k}_0$ - Esquema C

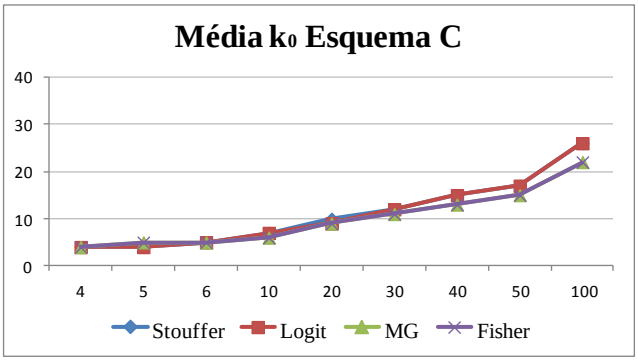


Gráfico 5.2.6. Rácio - Esquema C

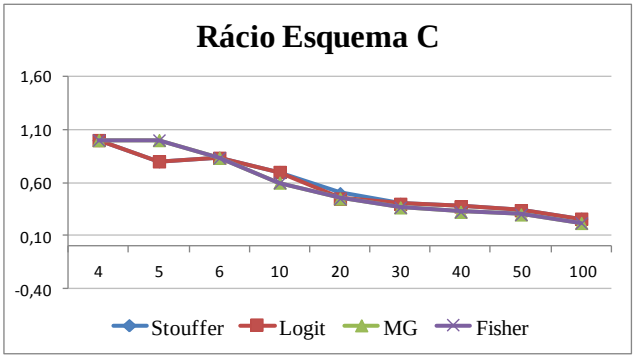


Gráfico 5.2.7. Desvio padrão - Esquema A

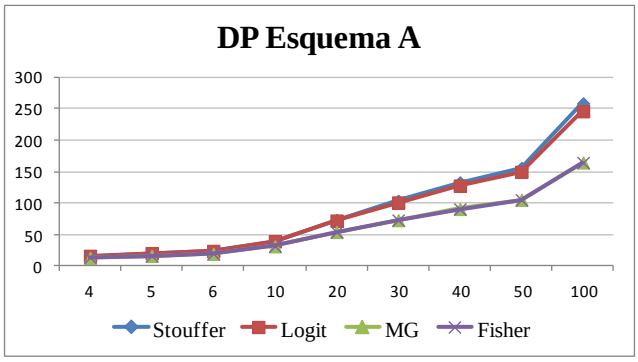


Gráfico 5.2.8. Coeficiente de Variação - Esquema A

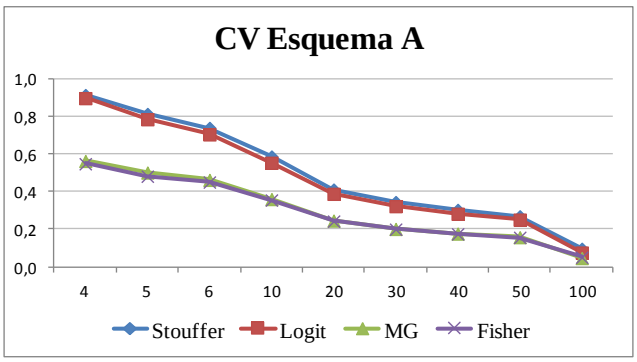


Gráfico 5.2.9. Desvio padrão - Esquema B

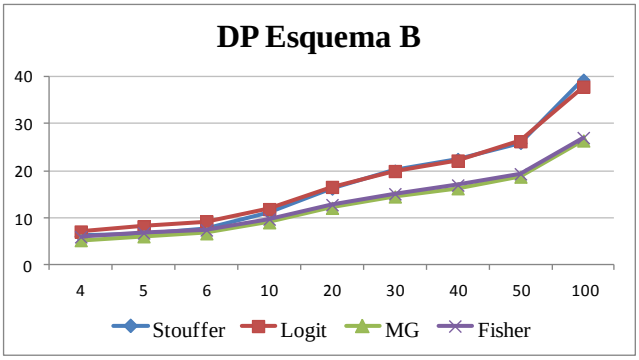


Gráfico 5.2.10. Coeficiente de Variação - Esquema B

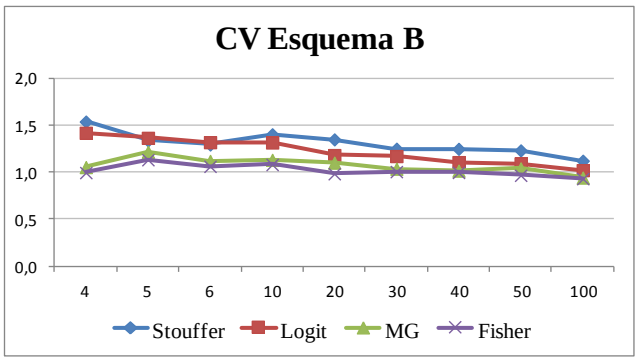


Gráfico 5.2.11. Desvio padrão - Esquema C

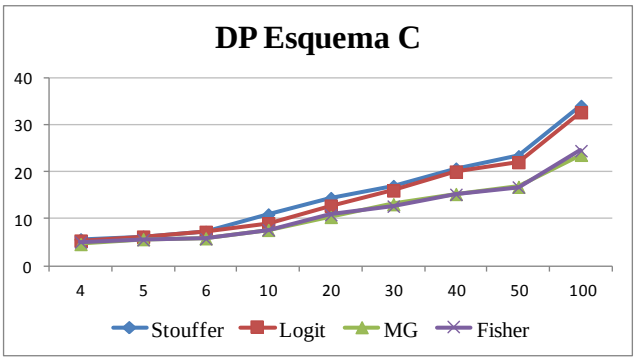
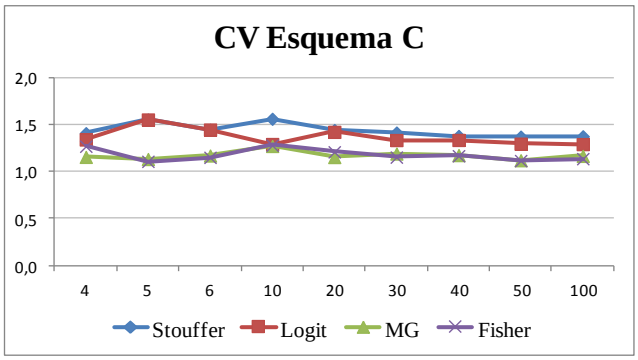


Gráfico 5.2.12. Coeficiente de Variação - Esquema C



Considerando os três esquemas de simulação realizados

O número médio de observações a acrescentar a uma amostra inicial, de forma a alterar a decisão de rejeição da hipótese nula global para uma decisão de não rejeição, é substancialmente maior quando se consideram amostras iniciais compostas exclusivamente por valores inferiores a 0,05 do que quando se consideram amostras iniciais que possam conter valores superiores a 0,05. Salientando-se que se observou uma menor variabilidade com amostras exclusivamente compostas por valores inferiores a 0,05, quando comparada a amostras iniciais que possam conter valores superiores a 0,05 (Esquemas B e C).

O número médio ($\bar{k}_0$) de observações (valores p não significativos) a acrescentar à amostra inicial, de forma a inverter a decisão de rejeição da hipótese nula global (H_0^*), para uma decisão de não rejeição, aumenta à medida que aumenta a dimensão da amostra inicial, independentemente da amostra inicial ser composta exclusivamente por valores de prova p inferiores a $0{,}05$ ou não. Esta constatação indicia que amostras de pequena dimensão (situação mais habitual em meta-análise), serão sempre mais problemáticas.

Os resultados obtidos, no que diz respeito às estimativas do número médio de observações a acrescentar de modo a inverter a decisão, apresentam uma grande variabilidade, que diminui à medida que a dimensão da amostra aumenta, independentemente da amostra inicial ser composta exclusivamente por valores de prova p inferiores a $0{,}05$ ou não. Observa-se que o Esquema B apresenta uma menor variabilidade comparativamente aos restantes esquemas. Considerando as estimativas calculadas, o método de Fisher é o que globalmente apresenta uma menor variabilidade, sendo o método de Stouffer o que apresenta a maior dentro dos quatro métodos utilizados.

É importante referir que o método de Tippett apresenta uma variabilidade consideravelmente superior a qualquer um dos métodos utilizados, o que é expectável. Realizou-se uma simulação adicional considerando o Esquema C onde se incluiu este método, tendo-se obtido resultados para o coeficiente de variação extremamente elevados comparativamente aos obtidos nos restantes métodos.

Identificam-se dois grupos distintos (independentemente do esquema utilizado)

1) Os métodos Stouffer e Logit apresentam resultados muito semelhantes no que diz respeito ao número médio de observações a acrescentar de modo a inverter a decisão, sendo observável, através do coeficiente de variação, uma variabilidade ligeiramente inferior na estimativa desse valor médio, aquando da aplicação do método Logit. Este resultado faz algum sentido dado que ambos os métodos se baseiam na aproximação pela Gaussiana, sendo que a t de Student converge para a Gaussiana quando $n \longrightarrow \infty$;

2) Os métodos Fisher e média geométrica apresentam resultados semelhantes no que diz respeito ao número médio de observações a acrescentar, de modo a inverter a decisão, sendo observável, através do coeficiente de variação, uma variabilidade ligeiramente inferior, na estimativa desse valor médio, aquando da aplicação do método de Fisher. Este resultado também faz algum sentido na medida em que quando se aplica o ln à média geométrica, a diferença entre este método e o método de Fisher reside na constante multiplicativa: $1/k$.

***Considerando amostras iniciais compostas exclusivamente por observações inferiores a* $\mathbf{0,05}$**

À medida que a dimensão da amostra inicial aumenta:

- Aumenta de forma acentuada, o número médio (e rácio) de valores de prova *p* (não significativos), necessários para mudar a decisão de rejeição para não rejeição de H_0^*;
- Diminui de forma acentuada, o coeficiente de variação (variabilidade das observações), independentemente do método aplicado.

Ordenando os métodos por ordem crescente do número médio de observações (valores *p* não significativos) a acrescentar (e rácio) de modo a inverter a decisão de rejeição para não rejeição, obteve-se os resultados apresentados na Tabela 5.1.2.

Analisando os resultados obtidos, o método de Fisher será provavelmente o método mais robusto ao enviesamento na publicação, considerando sínteses meta-analíticas compostas exclusivamente por resultados significativos (valores de prova *p* inferiores a 0,05).

Tabela 5.1.2. Esquema A – métodos ordenados por ordem crescente de $\bar{k}_0$.

Esquema A
1) Stouffer**
2) Logit
3) Média geométrica
4) Fisher*

* menor variabilidade; ** maior variabilidade.

Considerando amostras iniciais que possam conter observações superiores a 0,05

Apesar de se observarem semelhanças entre os resultados obtidos nos Esquemas 3 e 4 quando comparados com os resultados do Esquema 2, observou-se que número médio de observações a acrescentar à amostra inicial é superior, no Esquema 3, sendo a variabilidade dos dados inferior neste esquema, independentemente do método de combinação de valores *p*, aplicado;

À medida que a dimensão da amostra inicial aumenta:

- Aumenta o número médio de valores de prova *p* adicionais, necessários para mudar a decisão de rejeição para não rejeição de H_0^*;
- Diminui o rácio de valores de prova *p* adicionais em relação à dimensão da amostra inicial, necessários para inverter a decisão de rejeição para não rejeição de H_0^*;

- Diminui o coeficiente de variação (variabilidade das observações), independentemente do método aplicado;

- Observou-se um número (e percentagem) muito elevado de observações superiores a 0,05 nas amostras iniciais. Sendo esse valor mais baixo no Esquema 3 do que no Esquema 4 (Tabelas B.5 e B.6, Apêndice B).

Ordenando os métodos por ordem crescente do número médio de observações a acrescentar (e rácio) de modo a inverter a decisão de rejeição para não rejeição, obteve-se dois resultados distintos apresentados na Tabela 5.1.3.

Tabela 5.1.3. Esquema B e C – métodos ordenados por ordem crescente de $\bar{k}_0$.

Amostras iniciais de menor dimensão ($n \leq 20$ – Esquema B e $n \leq 6$ - Esquema C)	**Amostras iniciais de maior dimensão ($n > 20$ – Esquema B e $n > 6$ - Esquema C)**
1) Stouffer**	1) Média geométrica
2) Logit	2) Fisher*
3) Média geométrica*	3) Stouffer**
4) Fisher	4) Logit

* menor variabilidade; ** maior variabilidade.

Salienta-se que ao analisar as amostras iniciais geradas tanto no Esquema B como no Esquema C, observou-se uma grande proporção de valores superiores a 0,05 (Tabelas B.5 e B.6, Apêndice B). Neste sentido, será que as amostras geradas com valores de prova significativos e não significativos serão representativas de uma situação típica em sínteses meta-analíticas onde o viés de publicação esteja presente? Provavelmente não. Esta constatação explica até certo ponto que os resultados observados nestes esquemas sejam tão discrepantes dos obtidos através do Esquema A. Numa situação de amostras de dimensão tendencialmente mais baixa (Tabela 5.1.3), onde se verifica que a proporção de valores significativos ($\geq 0{,}05$) não é tão elevada comparativamente a amostras iniciais de maior dimensão (Tabelas B.5 e B.6, Apêndice B), os resultados são análogos aos obtidos no Esquema A. Consequentemente, considera-se que o método de Fisher será provavelmente o método mais robusto ao enviesamento na publicação, considerando sínteses meta-analíticas compostas por resultados significativos e não significativos.

5.3. Conclusões

Amostras iniciais compostas exclusivamente por valores de prova p significativos

Considerando sínteses meta-analíticas com base em testes de combinação de valores de prova *p*, compostas exclusivamente por valores *p* significativos, são necessários valores muito elevados de estudos não publicados, a serem considerados nessa análise, de forma a inverter a decisão da hipótese nula global de rejeição, para não rejeição.

Sendo que, quanto maior for o número de estudos incluídos na análise, maior será o número médio de estudos não significativos, a incluir para inverter a decisão de rejeição da hipótese nula global, isto é, quanto maior o número de estudos significativos a serem utilizados, mais forte e consolidada é a decisão de rejeição da hipótese nula global utilizando os métodos de combinação de valores *p* considerados (Stouffer, Logit, média geométrica e Fisher).

Amostras iniciais compostas por valores de prova p significativos e não significativos

Considerando sínteses meta-analíticas, com base em testes de combinação de valores de prova *p*, compostos por valores *p* significativos e não significativos, são necessários valores relativamente elevados de estudos não publicados a serem considerados na análise, de forma a inverter a decisão da hipótese nula global de rejeição para não rejeição. Sendo que, quanto maior for o número de estudos incluídos na análise, maior será o número médio de estudos não significativos a incluir para inverter a decisão de rejeição da hipótese nula global.

Não obstante da existência de aumento no número médio de valores de prova *p* significativos, a incluir na análise, à medida que a dimensão das amostras aumenta, existe diminuição da proporção (rácio) de estudos não publicados, relativamente à dimensão inicial da amostra. Esta situação ocorre, dado que quanto maior a dimensão da amostra, maior é o número (e proporção) de valores *p* não significativos, gerados pela simulação nas amostras iniciais. Isto é, quanto maior é a dimensão da amostra, maior é a proporção de valores de prova *p* não significativos, presentes na análise e quanto maior este valor, menos credível se torna a decisão de rejeição da hipótese nula global, sendo cada vez menor em termos proporcionais, o número de valores *p* a acrescentar de modo a inverter esta decisão.

Atendendo aos resultados obtidos, considera-se existir fortes indícios de que as sínteses meta-analíticas, com base exclusivamente em estudos significativos, ou não, necessitam de um número relativamente elevado de estudos não publicados, de forma a inverter uma decisão de rejeição da hipótese nula global para uma decisão de não

rejeição. Sendo que este número de estudos não publicados a incluir, aumenta com o aumento da dimensão das amostras iniciais, sendo tanto menor, quanto maior for o número de estudos significativos incluídos na análise síntese.

Dentro dos métodos analisados, considera-se que o método de Fisher é o método mais robusto perante uma síntese meta-analítica onde esteja presente o enviesamento na publicação.

Salienta-se a importância da publicação de estudos com resultados não significativos, bem como, da realização de sínteses meta-analíticas, independentemente do número de estudos contraditórios ou com muitos resultados não significativos.

Bibliografia

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

- Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, **49**, 559-575.

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

- Brilhante, M. F. (2013). Generalized *p*-values and random *p*-values when the alternative to uniformity is a mixture of a Beta(1,2) and Uniform. In Oliveira, P. *et al*. (eds), *Recent Developments in Modeling and Applications in Statistics*, Springer, Heidelberg, pp 159-167. DOI:10.1007/978-3-642-32419-2.

- Brilhante, M. F., Mendonça, S., Pestana, D., and Sequeira, F. (2010a). Using products and powers of products to test uniformity. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces*, 509-514.

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

- Casella, G., and Berger, R. L. (2002). *Statistical inference*. 2nd ed., Duxbury Press, Belmont, California.

- Chalmers, I. (2005). The scandalous failure of scientists to cumulate scientifically. *Clinical Trials*, **2**, 229-231.

- Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of Royal Statistical Society (Supplement)*, **4**, 102-118.

- Collins. R., Gray, R., Godwin, J., and Peto, R. (1987). Avoidances of large bias and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. *Statistics in Medicine*, **6**, 245-250.

- Collins, R., Yusuf, S., and Peto, R. (1985). Overview of randomized trials of diuretic pregnancy. *British Medical Journal*, **290**, 17-23.
- Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
- Dempster, A., and Schatzoff, M. (1965). Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association*, **60**, 420-436.
- Deng, L. Y., and George, E. O. (1992). Some characterizations of the uniform distribution with applications to random number generation. *Annals of the Institute of Statistical Mathematics*, **44**, 379-385.
- Dickersin, K. (2005) Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In Rothstein H. R., Sutton A. J., Borenstein M. (eds), *Publication Bias in Meta-Analysis Prevention, Assessment and Adjustments*, pp 356, John Wiley & Sons, West Sussex.
- Dickersin, K., and Min, Y. I. (1993). NIH clinical trials and publication bias. *Online Journal of Current Clinical Trials*, Doc. 50.
- Dickersin, K., Min, Y. I., and Meinert, C. L. (1992). Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *Journal of American Medical Association*, **267**, 374-378.
- Donahue, R. (1999). A note on information seldom reported via the *p*-value. *American Statistician*, **53**, 303-306.
- Duval, S., and Tweedie, R. (2000a). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, **95**, 89-98.
- Duval, S., and Tweedie, R. (2000b). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455-463.
- Easterbrook, P. J., Berlin, J. A., Gopalan, R., and Matthews, D. R. (1991). Publication bias in clinical research. *Lancet*, **337**, 867-872.
- 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&Boyd, London.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver&Boyd, Edinburgh.

- Fisher, R. A. (1995). *Statistical methods, experimental design and scientific inference (re-issue of Statistical methods for research workers, The design of experiments, and Statistical methods and scientific inference)*, Oxford University Press, Oxford.

- Fraga Alves, I., Gomes, M. I., e Sousa, L. (2007). *Fundamentos e Metodologias da Estatística*. Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa.

- Gamage, J., and Weerahandi, S. (1998). Size performance of some tests in one-way ANOVA. *Communication in Statistics – Simulation and Computation*, **27**, 625-640.

- George, E. O. (1977). *Combining independent one-sided and two-sided statistical tests – some theory and applications*. Doctoral dissertation, University of Rochester, Rochester.

- Gibbons, J. D., and Pratt, J. W. (1975). *P*-values: interpretation and methodology. *The American Statistician*. **29**, 20–25.

- Gilbert, R., Salanti, G., Harden, M., and See, S. (2005). Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International Journal of Epidemiology*, **34**, 874-887.

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

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

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

- Hartung, J. (1999). A note on combining dependent tests of significance. *Biometrical Journal*, **41**, 849-855.

- Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical meta-analysis with applications*. Wiley, New York.

- 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.

- Iyer, H. K., and Patterson, P. D. (2002). *A Recipe for Constructing Generalized Pivotal quantities and Generalized Confidence Intervals*. Technical Report 2002/10, Colorado State University - Department of Statistics, Colorado.

- Johnson, R. A., and Weerahandi, S. (1988). A Bayesian Solution to the Multivariate Behrens-Fisher Problem. *Journal of the American Statistical Association*, **83**, 145-149.

- Kempthorne, O., and Folks, L. (1971). *Probability, Statistics and Data Analysis*. Ames: Iowa State University Press.

- Kulinskaya, E., Morgenthaler, S., and Staudte, R. G. (2008). *Meta analysis - a guide to calibrating and combining statistical evidence*. Wiley, Chichester.

- Loughin, T. M. (2004). A systematic comparison of methods for combining *p*-values from independent tests. *Computational Statistics and Data Analysis*, **47**, 467-485.

- Lehmann, E. L. (1986). *Testing statistical hypotheses.* 2nd ed., John Willey, New York.

- Makambi, K. H. (2003). Weighted inverse chi-square method for correlated significance tests. *Journal of Applied Statistics*, **30**, 225-234.

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

- Marden, J. I. (1991). Sensitive and sturdy *p*-values. *The Annals of Statistics*, **19**, 918-934.

- Mosteller, F., and Bush, R. (1954). Selected quantitative techniques. In G. Lindsey (eds.), *Handbook of Social Psychology: Theory and Methods*, Vol. 1, Addison-Wesley, Cambridge, MA.

- Mudholkar, Govind., and George, E. O. (1978). *The logit statistic for combining probabilities - an overview*. Department of Statistics - Rochester University, New York.

- Murteira, B., Ribeiro, C. S., Silva, J. A., e Pimenta, C. (2010). *Introdução à estatística*. Escolar Editora, Lisboa.

- Orwin, R. G., and Boruch, R.F. (1983). RRT meets RDD: statistical strategies for assuring response privacy in telephone surveys. *Public Opinion Quarterly*, **46**, 560-571.

- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated systems of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 5th series, **1**, 157-175.

- Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, **3**, 1243-1246.

- Pearson, K. (1933). On a method of determining whether a sample of given size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**, 379-410.

- Pestana, D. (2011). Combining *p*-values. In M. Lovric (ed.), *International Encyclopedia of Statistical Science*, pp 1145-1147, Springer - Verlag, Berlin, Heidelberg.

- Pestana D., e Velosa S. (2010). *Introdução à probabilidade e à estatística - Volume I*. 4ª edição, Fundação Calouste Gulbenkian, Lisboa.

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

- Sackowitz, H., and Samuel-Cahn, E. (1999). *P* values as random variables, expected *p* values. *The American Statistician*, **53**, 326-331.

- Sequeira F. (2009). *Meta-Análise - Harmonização de Testes Usando os Valores de Prova.* Tese de Doutoramento em Estatística e Investigação Operacional (Especialidade de Probabilidades e Estatística), Faculdade de Ciências – Universidade de Lisboa, Lisboa.

- Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams, R. M. Jr. (1949). *The american soldier, vol. I: adjustment during army life*. Princeton University Press, Princeton NJ.

- Student (1908). The probable error of a mean. *Biometrika*, **6**, 1-25.

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

- Thompson, W. A., Jr. (1985). Optimal Significance Procedures for Simple Hypotheses. *Biometrika*, **72**, 230-232.

- 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 in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**, 602-607.

- Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, **88**, 899-905.
- Weerahandi, S. (2003). *Exact statistical methods for data analysis*. Springer, New York.
- Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, **38**, 330-336.
- Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, **48**, 156-158.

Apêndice A

Conceitos e resultados relevantes

Estatística suficiente

Uma estatística suficiente para um parâmetro θ é uma estatística que num certo sentido consegue capturar toda a informação sobre θ, que se encontra contida na amostra. Qualquer informação adicional presente na amostra, sem ser o valor da estatística suficiente, não contem mais informação sobre θ. Estas considerações levam a uma técnica de redução de dados conhecida como o Princípio da Suficiência.

Princípio da suficiência

Se $T(\boldsymbol{X})$ é uma estatística suficiente para θ, então qualquer inferência sobre θ deve depender da amostra $\boldsymbol{X}$ somente pelo valor de $T(\boldsymbol{X})$. Isto é, se $\boldsymbol{x}$ e $\boldsymbol{y}$ são dois pontos amostrais tais que $T(\boldsymbol{x}) = T(\boldsymbol{y})$, então a inferência sobre θ deve ser a mesma, quer $\boldsymbol{X} = \boldsymbol{x}$ ou $\boldsymbol{Y} = \boldsymbol{y}$, seja observado.

Definição A.1.

Uma estatística $T(\boldsymbol{X})$ é uma estatística suficiente para θ se a distribuição condicional da amostra $\boldsymbol{X}$, dado um valor de $T(\boldsymbol{X})$, não depende de θ, isto é,

$$P_\theta(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x})) = \frac{P_\theta\big(\boldsymbol{X} = \boldsymbol{x} \wedge T(\boldsymbol{X}) = T(\boldsymbol{x})\big)}{P_\theta\big(T(\boldsymbol{X}) = T(\boldsymbol{x})\big)} =$$

$$= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x})}{P_\theta\big(T(\boldsymbol{X}) = T(\boldsymbol{x})\big)}$$

não depende de θ.

Critério da factorização

Teorema A.1.

Seja $f(\boldsymbol{x}|\theta)$ a f.d.p. ou f.m.p. conjunta de uma amostra $\boldsymbol{X}$. A estatística $T(\boldsymbol{X})$ é suficiente para θ se e só se existirem funções $g(t|\theta)$ e $h(\boldsymbol{x})$, tais que para todos os pontos amostrais $\boldsymbol{x}$,

$$f(\boldsymbol{x}|\theta) = g(\boldsymbol{T}(\boldsymbol{x})|\theta)h(\boldsymbol{x}),$$

onde a função g depende do vetor aleatório observável $\boldsymbol{x}$ apenas pelo valor de $T(\boldsymbol{x})$ e a função h pode depender de $\boldsymbol{x}$ mas não de θ.

Estatística suficiente mínima

Definição A.2.

Uma estatística suficiente $T(\boldsymbol{X})$ é designada estatística suficiente mínima se, para qualquer outra estatística suficiente $T'(\boldsymbol{X})$, $T(\boldsymbol{X})$ é função de $T'(\boldsymbol{X})$.

Critério de estatística suficiente mínima de Lehman - Scheffé

Teorema A.2.

Seja $f(\boldsymbol{x}|\theta)$ a f.d.p. ou f.m.p. conjunta de uma amostra $\boldsymbol{X}$. Suponha-se que existe uma função $T(\boldsymbol{x})$ tal que para quaisquer pontos amostrais $\boldsymbol{x}$ e $\boldsymbol{y}$, a razão $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ é constante como função de θ se e só se $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Então $T(\boldsymbol{X})$ é uma estatística suficiente mínima para θ.

Obs.: uma estatística suficiente mínima não é única. Qualquer transformação bijetiva dessa estatística suficiente mínima ainda é mínima.

Família exponencial

Definição A.3.

Uma família de f.d.p. ou f.m.p. é designada por família exponencial se puder ser expressa como

$$f(\boldsymbol{x}) = h(\boldsymbol{x})c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(\boldsymbol{x})\right)$$

onde $\boldsymbol{\theta}$ é o vetor dos parâmetros de interesse, com funções reais tais que $h(\boldsymbol{x}) \geq 0$, $t_i(\boldsymbol{x})$ não dependentes de $\boldsymbol{\theta}$, $c(\boldsymbol{\theta}) \geq 0$ e $w_i(\boldsymbol{\theta})$ não dependentes de $\boldsymbol{x}$.

Suficiência na família exponencial

Teorema A.3.

Seja $\boldsymbol{X} = (X_1, \dots, X_k)$ uma amostra aleatória proveniente de um modelo com f.d.p. ou f.m.p. $f(\boldsymbol{x}|\boldsymbol{\theta})$ pertencente à família exponencial multiparamétrica,

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x})c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(\boldsymbol{x})\right),$$

com $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d), d \leq k$. Então

$$T(\boldsymbol{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \dots, \sum_{i=1}^{n} t_k(X_i)\right)$$

é uma estatística suficiente para $\boldsymbol{\theta}$.

Estatística suficiente completa

Definição A.4.

Suponha-se que $T(\boldsymbol{X})$ é uma estatística suficiente para θ. A estatística T e a família de distribuições de T parametrizada por θ são ditas completas se para uma dada estatística $h(T)$,

$$E_\theta[h(T)] = 0 \; para\ todo\ o\ \theta \in \Theta \Rightarrow P_\theta[h(T) = 0] = 1 \; , para\ todo\ o\ \theta \in \Theta.$$

Obs.: uma estatística suficiente completa é necessariamente mínima, no entanto o recíproco não é necessariamente válido.

Estimador centrado

Definição A.5.

Um estimador diz-se centrado para $\tau(\theta)$ se satisfaz $E_\theta[T] = \tau(\theta)$.

Estimador centrado de variância uniformemente mínima

Definição A.6.

Um estimador T^* é o melhor estimador centrado de $\tau(\theta)$ se satisfizer $E_\theta[T^*] = \tau(\theta)$ para todo o θ e para qualquer outro estimador T com $E_\theta[T] = \tau(\theta)$, se tiver $Var_\theta[T^*] \leq Var_\theta[T]$ para todo o θ. T^* é também chamado estimador centrado de variância uniformemente mínima (UMVUE) de $\tau(\theta)$.

Teste de nível α

Definição A.7.

Considere-se o teste

$$H_0: \theta \in \Theta_0 \; vs \; H_A: \theta \in \Theta_A = \Theta_0^c.$$

Para $0 \leq \alpha \leq 1$, um teste com função potência $\pi(\theta)$ é um teste de nível α se

$$Sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha.$$

Testes uniformemente mais potentes

Definição A.8.

Seja C_α, a classe de testes para testar ao nível α,

$$H_0: \theta \in \Theta_0 \; vs \; H_A: \theta \in \Theta_0^c \, .$$

Um teste da classe C_α, com função potência $\pi(\theta)$ é dito uniformemente mais potente (UMP) de nível α se

$$\pi(\theta) \geq \pi'(\theta),$$

para todo o $\theta \in \Theta_0^c$ com $\pi'(\theta)$ a função potência para qualquer teste de C_α.

Teorema da transformação uniformizante

Teorema A.4.

Seja X uma variável aleatória (v.a.) absolutamente contínua, com função distribuição $F_X(x)$ crescente. Defina-se a v.a. Y, $Y = F_X(x)$. Assim, Y tem distribuição Uniforme em (0,1), isto é: $P(Y \leq y) = y, \; 0 < y < 1$.

Teorema do limite central

Teorema A.5.

Seja $\{X_k\}_{k\geq 1}$ uma sucessão de variáveis aleatórias independentes e identicamente distribuídas, com variância finita $Var[X_k] = \sigma^2$, e denote-se $E[X_k] = \mu$, $S_n = \sum_{i=1}^{k} X_k$.

Quando $n \to \infty$, a função distribuição de S_n, convenientemente normalizada, converge em distribuição para uma função de distribuição Gaussiana.

Como $E[S_n] = n\mu$ e $Var[S_n] = n\sigma^2$, escrevemos

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \cap Gaussiana(0,1).$$

Método da média Geométrica – cálculo do quantil

Demonstração A.1.

Construção de um teste de nível α de forma a combinar valores de prova *p*, utilizando a média geométrica:

Sob H_0^*: $P_i \cap U(0,1)$. Seja $Y = -2ln(P_i)$ então: $Y_{|H_0^*} \cap exp\,(2)$. Considere-se $Z = -2\sum_{i=1}^{n} ln(P_i)$, então: $Z_{|H_0^*} \cap Gama(n,2) \equiv \chi^2_{2n}$.

Pretende-se:

$$P\left[\left(\prod_{i=1}^{n} p_i\right)^{1/n} < C_\propto | H_0^* verdadeira\right] = \propto \Leftrightarrow$$

$$\Leftrightarrow P\left[-2ln\left(\prod_{i=1}^{n} p_i\right)^{1/n} > -2\,ln\,C_\propto\,| H_0^* verdadeira\right] = \propto \Leftrightarrow$$

$$\Leftrightarrow P\left[-\frac{2}{n} ln\left(\prod_{i=1}^{n} p_i\right) < -\,2\,ln\,C_\propto\,| H_0^* verdadeira\right] = 1-\propto \Leftrightarrow$$

$$\Leftrightarrow P\left[-2\,ln\left(\prod_{i=1}^{n} p_i\right) < -\,2n\,ln\,C_\propto\,| H_0^* verdadeira\right] = 1-\propto \Leftrightarrow$$

$$\Leftrightarrow P\left[-2\sum_{i=1}^{n} ln(\,p_i) < -\,2n\,ln\,C_\propto\,| H_0^* verdadeira\right] = 1-\propto \Leftrightarrow$$

$$\Leftrightarrow P[Z < -2n\,ln\,C_\propto\,| H_0^* verdadeira] = 1-\alpha$$

Em que $Z = -2\sum_{i=1}^{n} ln(p_i)$, onde $Z_{|H_0^*} \cap \chi^2_{2n}$.

Cálculo de C_α:

$$-2n\, ln\, C_\propto = \chi^2_{2n;1-\alpha} \Leftrightarrow$$

$$\Leftrightarrow C_\alpha = exp\left\{-\frac{\chi^2_{2n;1-\alpha}}{2n}\right\}$$

Apêndice B

Resultados – simulação

Tabela B.1. Média: $\bar{k}_0$.

Esquema A		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	17	23	31	66	178	301	439	580	1338
	Logit	18	26	34	71	186	313	453	597	1365
	MG	23	32	42	87	220	362	514	669	1486
	Fisher	25	34	44	89	223	365	518	673	1489
Esquema B		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	4	5	6	8	12	16	18	21	35
	Logit	5	6	7	9	14	17	20	24	37
	MG	5	5	6	8	11	14	16	18	28
	Fisher	6	6	7	9	13	15	17	20	29
Esquema C		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	4	4	5	7	10	12	15	17	26
	Logit	4	4	5	7	9	12	15	17	26
	MG	4	5	5	6	9	11	13	15	22
	Fisher	4	5	5	6	9	11	13	15	22

Tabela B.2. Racio: $\bar{k}_0/k$.

Esquema A		Dimensão da amostra								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	4,25	4,60	5,17	6,60	8,90	10,03	10,98	11,60	13,38
	Logit	4,50	5,20	5,67	7,10	9,30	10,43	11,33	11,94	13,65
	MG	5,75	6,40	7,00	8,70	11,00	12,07	12,85	13,38	14,86
	Fisher	6,25	6,80	7,33	8,90	11,15	12,17	12,95	13,46	14,89
Esquema B		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	1,00	1,00	1,00	0,80	0,60	0,53	0,45	0,42	0,35
	Logit	1,25	1,20	1,17	0,90	0,70	0,57	0,50	0,48	0,37
	MG	1,25	1,00	1,00	0,80	0,55	0,47	0,40	0,36	0,28
	Fisher	1,50	1,20	1,17	0,90	0,65	0,50	0,43	0,40	0,29
Esquema C		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	1,00	0,80	0,83	0,70	0,50	0,40	0,38	0,34	0,26
	Logit	1,00	0,80	0,83	0,70	0,45	0,40	0,38	0,34	0,26
	MG	1,00	1,00	0,83	0,60	0,45	0,37	0,33	0,30	0,22
	Fisher	1,00	1,00	0,83	0,60	0,45	0,37	0,33	0,30	0,22

Tabela B.3. Desvio Padrão Amostral.

Esquema A		Dimensão da amostra								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	15,54	18,71	22,92	38,75	73,18	103,75	133,00	155,95	258,78
	Logit	16,16	20,42	23,96	39,31	72,22	100,96	127,86	149,85	246,25
	MG	12,96	16,05	19,43	31,36	54,11	72,90	90,79	105,37	164,77
	Fisher	13,71	16,39	19,84	31,61	54,13	72,94	90,61	105,24	164,54
Esquema B		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	6,16	6,70	7,78	11,24	16,21	20,03	22,50	25,91	39,32
	Logit	7,11	8,20	9,25	11,86	16,57	19,94	22,17	26,26	37,84
	MG	5,28	6,09	6,73	9,06	12,13	14,54	16,28	18,75	26,42
	Fisher	5,99	6,81	7,44	9,75	12,81	15,12	16,97	19,36	27,04
Esquema C		**Dimensão da amostra**								
		4	**5**	**6**	**10**	**20**	**30**	**40**	**50**	**100**
Método de combinação	**Stouffer**	5,64	6,22	7,23	10,94	14,43	17,00	20,70	23,34	34,05
	Logit	5,37	6,22	7,24	9,06	12,81	16,04	20,06	22,06	32,68
	MG	4,63	5,67	5,86	7,64	10,40	13,12	15,29	16,82	23,62
	Fisher	5,08	5,55	5,76	7,69	10,90	12,70	15,19	16,77	24,51

Tabela B.4. Coeficiente de Variação.

Esquema A		Dimensão da amostra								
		4	5	6	10	20	30	40	50	100
Método de combinação	Stouffer	0,91	0,81	0,74	0,59	0,41	0,34	0,30	0,27	0,10
	Logit	0,90	0,79	0,70	0,55	0,39	0,32	0,28	0,25	0,08
	MG	0,56	0,50	0,46	0,36	0,25	0,20	0,18	0,16	0,05
	Fisher	0,55	0,48	0,45	0,36	0,24	0,20	0,17	0,16	0,05
Esquema B		Dimensão da amostra								
		4	5	6	10	20	30	40	50	100
Método de combinação	Stouffer	1,54	1,34	1,30	1,41	1,35	1,25	1,25	1,23	1,12
	Logit	1,42	1,37	1,32	1,32	1,18	1,17	1,11	1,09	1,02
	MG	1,06	1,22	1,12	1,13	1,10	1,04	1,02	1,04	0,94
	Fisher	1,00	1,13	1,06	1,08	0,99	1,01	1,00	0,97	0,93
Esquema C		Dimensão da amostra								
		4	5	6	10	20	30	40	50	100
Método de combinação	Stouffer	1,41	1,55	1,45	1,56	1,44	1,42	1,38	1,37	1,38
	Logit	1,34	1,56	1,45	1,29	1,42	1,34	1,34	1,30	1,29
	MG	1,16	1,13	1,17	1,27	1,16	1,19	1,18	1,12	1,17
	Fisher	1,27	1,11	1,15	1,28	1,21	1,15	1,17	1,12	1,14

Tabela B.5. Valores $\geq 0{,}05$ observados nas amostras iniciais – Esquema B.

Esquema B	Dimensão da amostra								
	4	5	6	10	20	30	40	50	100
Média	2,8	3,7	4,6	8,1	17,3	26,5	35,7	45,0	91,7
%	69,83%	73,51%	76,19%	81,50%	86,44%	88,30%	89,34%	90,01%	91,66%

Tabela B.6. Valores $\geq 0{,}05$ observados nas amostras iniciais – Esquema C.

Esquema C		Dimensão da amostra								
		4	5	6	10	20	30	40	50	100
Método de combinação	Stouffer	3,0	3,9	4,8	8,5	17,7	27,0	36,4	45,7	92,6
	Logit	2,9	3,8	4,8	8,4	17,6	26,9	36,2	45,5	92,4
	MG	2,8	3,7	4,6	8,1	17,3	26,4	35,6	44,9	91,5
	Fisher	2,8	3,7	4,6	8,1	17,2	26,4	35,6	44,9	91,5
	Média Global*	2,9	3,8	4,7	8,3	17,5	26,8	36,1	45,4	92,2
	%	73,0%	76,4%	78,7%	83,3%	87,7%	89,2%	90,1%	90,8%	92,2%

*As amostras iniciais são diferentes entre os diversos métodos. De modo a facilitar a análise (e comparação) destes resultados, calculou-se a média global dos quatro métodos aplicados.

Código – R

Esquema A

```
testLogit <- function(pvalues){
	k <- length(pvalues)
	G=-sum(log(pvalues/(1-pvalues)))*(k*pi^2*(5*k+2)/(3*(5*k+4)))^(-1/2)
	if(G>qt(1-alfa,5*k+4))
		return(1)
	else return(0)
}
testStouffer <- function(pvalues){ ##Rejeitar
	k <- length(pvalues)
	(stouffer <- sum(qnorm(pvalues)) / sqrt(k))
	(pv.stouffer  <- pnorm(stouffer))
	if(pv.stouffer < alfa)
		return(1)
	else return(0)
}
testFisher <- function(pvalues){
	k <- length(pvalues)
	(fisher <- sum(-2 * log(pvalues)))
	(pv.fisher  <- 1 - pchisq(fisher, 2*k))
	if(pv.fisher < alfa)
		return(1)
	else return(0)
}
testMediaGeometrica <- function(pvalues){
	k <- length(pvalues)
```

```
	mg <- exp((1/k)*sum(log(pvalues)))
	vcritico <- exp(-(qchisq(1-alfa,2*k))/(2*k))
	if(mg < vcritico)
		return(1)
	else return(0)
}


###INICIO
###Definicao da simulacao
alfa=0.05
nruns=5000
maxdim=100

##Dimensoes aceitaveis
nnames <- c()
for(p in 1:maxdim){
	if((p==4)||(p>4 && p<=6)||(p==10)||(p<=50 && (p%%10==0))||(p==50)||(p>=100 &&
p%%100==0))
		nnames <-c(nnames,p)
}


###AUXILIAR PARA ANALIZAR RESULTADOS
result <-
array(0,c(4,length(nnames),2),list(c('Stouffer','Fisher','MediaGeometrica','Logit'),nnames,c('Media','DP')))

for(z in 1:length(nnames)){
	n=nnames[z]
	p=1
	n0sStouffer=c()
```

```
n0sFisher=c()
n0sMediaGeometrica=c()
n0sLogit=c()

while(p<=nruns){
	amostraGeral=c()
	n0Stouffer=0
	n0Fisher=0
	n0MediaGeometrica=0
	n0Logit=0
	u1=c()
	repeat{
		u1=runif(n,0,alfa)
if(testStouffer(u1)&&testFisher(u1)&&testMediaGeometrica(u1)&&testLogit(u1))
			break
	}

	########################
	#### TESTE STOUFFER ###
	########################
	y=u1
	l=1
	while(testStouffer(y)){
		if(l<length(amostraGeral))
			n0=amostraGeral[l]
		else {
			n0=runif(1,alfa,1)
			amostraGeral=c(amostraGeral,n0)
		}
		l=l+1
```

```
        y=c(y,n0)
        n0Stouffer=n0Stouffer+1
}
######################
#### TESTE FISHER #####
######################
y=u1
l=1
while(testFisher(y)){
        if(l<length(amostraGeral))
                n0=amostraGeral[l]
        else {
                n0=runif(1,alfa,1)
                amostraGeral=c(amostraGeral,n0)
        }
        l=l+1
        y=c(y,n0)
        n0Fisher=n0Fisher+1
}
################################
#### TESTE MEDIA GEOMETRICA ##
################################
y=u1
l=1
while(testMediaGeometrica(y)){
        if(l<length(amostraGeral))
                n0=amostraGeral[l]
        else {
                n0=runif(1,alfa,1)
                amostraGeral=c(amostraGeral,n0)
```

```
                }
                l=l+1
                y=c(y,n0)
                n0MediaGeometrica=n0MediaGeometrica+1
        }
        ######################
        #### TESTE LOGIT #####
        ######################
        y=u1
        l=1
        while(testLogit(y)){
                if(l<length(amostraGeral))
                        n0=amostraGeral[l]
                else {
                        n0=runif(1,alfa,1)
                        amostraGeral=c(amostraGeral,n0)
                }
                l=l+1
                y=c(y,n0)
                n0Logit=n0Logit+1
        }

        n0sStouffer=c(n0sStouffer,n0Stouffer);
        n0sFisher=c(n0sFisher,n0Fisher);
        n0sMediaGeometrica =c(n0sMediaGeometrica, n0MediaGeometrica);
        n0sLogit=c(n0sLogit,n0Logit);
        p=p+1
}

result['Stouffer',z,'Media']=round(mean(n0sStouffer),0)
```

```
        result['Stouffer',z,'DP']=sd(n0sStouffer)
        result['Fisher',z,'Media']=round(mean(n0sFisher),0)
        result['Fisher',z,'DP']=sd(n0sFisher)
        result['MediaGeometrica',z,'Media']=round(mean(n0sMediaGeometrica),0)
        result['MediaGeometrica',z,'DP']=sd(n0sMediaGeometrica)
        result['Logit',z,'Media']=round(mean(n0sLogit),0)
        result['Logit',z,'DP']=sd(n0sLogit)

}
```

Esquema B

```
testLogit <- function(pvalues){
        k <- length(pvalues)
        G=-sum(log(pvalues/(1-pvalues)))*(k*pi^2*(5*k+2)/(3*(5*k+4)))^(-1/2)
        if(G>qt(1-alfa,5*k+4))
                return(1)
        else return(0)
}
testStouffer <- function(pvalues){ ##Rejeitar
        k <- length(pvalues)
        (stouffer <- sum(qnorm(pvalues)) / sqrt(k))
        (pv.stouffer  <- pnorm(stouffer))
        if(pv.stouffer < alfa)
                return(1)
        else return(0)
}
testFisher <- function(pvalues){
        k <- length(pvalues)
```

```
	(fisher <- sum(-2 * log(pvalues)))
	(pv.fisher  <- 1 - pchisq(fisher, 2*k))
	if(pv.fisher < alfa)
		return(1)
	else return(0)
}
testMediaGeometrica <- function(pvalues){
	k <- length(pvalues)
	mg <- exp((1/k)*sum(log(pvalues)))
	vcritico <- exp(-(qchisq(1-alfa,2*k))/(2*k))
	if(mg < vcritico)
		return(1)
	else return(0)
}



###INICIO
###Definicao da simulacao
alfa=0.05
nruns=5000
maxdim=100


##Dimensoes aceitaveis
nnames <- c()
for(p in 1:maxdim){
	if((p==4)||(p>4 && p<=6)||(p==10)||(p<=50 && (p%%10==0))||(p==50)||(p>=100 &&
p%%100==0))
		nnames <-c(nnames,p)
}
```

```
###AUXILIAR PARA ANALIZAR RESULTADOS

result <-
array(0,c(4,length(nnames),2),list(c('Stouffer','Fisher','MediaGeometrica','Logit'),nnames,c('Media','DP')))

mediaPVSUP <- array(0,c(length(nnames)),list(nnames))


for(z in 1:length(nnames)){
        n=nnames[z]
        p=1
        n0sStouffer=c()
        n0sFisher=c()
        n0sMediaGeometrica=c()
        n0sLogit=c()
        mediaPVSUP[z]=0
        while(p<=nruns){
                amostraGeral=c()
                n0Stouffer=0
                n0Fisher=0
                n0MediaGeometrica=0
                n0Logit=0
                u1=c()
                repeat{
                        u1=runif(n)

        if(testStouffer(u1)&&testFisher(u1)&&testMediaGeometrica(u1)&&testLogit(u1))
                                break
                }
                mediaPVSUP[z]=mediaPVSUP[z]+length(u1[u1>=alfa])
                n1=length(u1)

                ########################
                #### TESTE STOUFFER ####
```

```
########################
y=u1
l=1


while(testStouffer(y)){
        if(l<length(amostraGeral))
                n0=amostraGeral[l]
        else {
                n0=runif(1,alfa,1)
                amostraGeral=c(amostraGeral,n0)
        }
        l=l+1
        y=c(y,n0)
        n0Stouffer=n0Stouffer+1
}
########################
#### TESTE FISHER #####
########################
y=u1
l=1
while(testFisher(y)){
        if(l<length(amostraGeral))
                n0=amostraGeral[l]
        else {
                n0=runif(1,alfa,1)
                amostraGeral=c(amostraGeral,n0)
        }
        l=l+1
        y=c(y,n0)
        n0Fisher=n0Fisher+1
```

```
}
################################
#### TESTE MEDIA GEOMETRICA ##
################################
y=u1
l=1
while(testMediaGeometrica(y)){
        if(l<length(amostraGeral))
                n0=amostraGeral[l]
        else {
                n0=runif(1,alfa,1)
                amostraGeral=c(amostraGeral,n0)
        }
        l=l+1
        y=c(y,n0)
        n0MediaGeometrica=n0MediaGeometrica+1
}
######################
#### TESTE LOGIT #####
######################
y=u1
l=1
while(testLogit(y)){
        if(l<length(amostraGeral))
                n0=amostraGeral[l]
        else {
                n0=runif(1,alfa,1)
                amostraGeral=c(amostraGeral,n0)
        }
        l=l+1
```

```
                y=c(y,n0)
                n0Logit=n0Logit+1
            }


            n0sStouffer=c(n0sStouffer,n0Stouffer);
            n0sFisher=c(n0sFisher,n0Fisher);
            n0sMediaGeometrica =c(n0sMediaGeometrica,n0MediaGeometrica);
            n0sLogit=c(n0sLogit,n0Logit);
            p=p+1
        }
        mediaPVSUP[z]=mediaPVSUP[z]/nruns
        result['Stouffer',z,'Media']=round(mean(n0sStouffer),0)
        result['Stouffer',z,'DP']=sd(n0sStouffer)
        result['Fisher',z,'Media']=round(mean(n0sFisher),0)
        result['Fisher',z,'DP']=sd(n0sFisher)
        result['Logit',z,'Media']=round(mean(n0sLogit),0)
        result['Logit',z,'DP']=sd(n0sLogit)
        result['MediaGeometrica',z,'Media']=round(mean(n0sMediaGeometrica),0)
        result['MediaGeometrica',z,'DP']=sd(n0sMediaGeometrica)

}
```

Esquema C

```
testLogit <- function(pvalues){
        k <- length(pvalues)
        G=-sum(log(pvalues/(1-pvalues)))*(k*pi^2*(5*k+2)/(3*(5*k+4)))^(-1/2)
        if(G>qt(1-alfa,5*k+4))
```

```
                return(1)
        else return(0)
}
testStouffer <- function(pvalues){ ##Rejeitar
        k <- length(pvalues)
        (stouffer <- sum(qnorm(pvalues)) / sqrt(k))
        (pv.stouffer  <- pnorm(stouffer))
        if(pv.stouffer < alfa)
                return(1)
        else return(0)
}
testFisher <- function(pvalues){
        k <- length(pvalues)
        (fisher <- sum(-2 * log(pvalues)))
        (pv.fisher  <- 1 - pchisq(fisher, 2*k))
        if(pv.fisher < alfa)
                return(1)
        else return(0)
}
testMediaGeometrica <- function(pvalues){
        k <- length(pvalues)
        mg <- exp((1/k)*sum(log(pvalues)))
        vcritico <- exp(-(qchisq(1-alfa,2*k))/(2*k))
        if(mg < vcritico)
                return(1)
        else return(0)
}
```

```
###INICIO

###Definicao da simulacao

alfa=0.05

nruns=5000

maxdim=100


##Dimensoes aceitaveis

nnames <- c()

for(p in 1:maxdim){

	if((p==4)||(p>4 && p<=6)||(p==10)||(p<=50 && (p%%10==0))||(p==50)||(p>=100 &&
p%%100==0))

		nnames <-c(nnames,p)

}



###AUXILIAR PARA ANALIZAR RESULTADOS

result <-
array(0,c(5,length(nnames),3),list(c('Stouffer','Fisher','MediaGeometrica','Logit','Tippett'),nnames,c('Medi
a','DP','MediaPVSUP')))


for(z in 1:length(nnames)){

	n=nnames[z]

	p=1

	n0sStouffer=c()

	n0sFisher=c()

	n0sMediaGeometrica=c()

	n0sLogit=c()

	n0sTippett=c()

	result['Stouffer',z,'MediaPVSUP']=0

	result['Fisher',z,'MediaPVSUP']=0

	result['MediaGeometrica',z,'MediaPVSUP']=0
```

```
result['Logit',z,'MediaPVSUP']=0
result['Tippett',z,'MediaPVSUP']=0

while(p<=nruns){
        n0Stouffer=0
        n0Fisher=0
        n0MediaGeometrica=0
        n0Logit=0
        n0Tippett=0

        ########################
        #### TESTE STOUFFER ####
        ########################

        u1=c()
        repeat{
                u1=runif(n)
                if(testStouffer(u1))
                        break
        }

result['Stouffer',z,'MediaPVSUP']=result['Stouffer',z,'MediaPVSUP']+length(u1[u1>=alfa])
        y=u1
        while(testStouffer(y)){
                y <- c(y,runif(1,alfa,1))
                n0Stouffer=n0Stouffer+1
        }
        ########################
        #### TESTE FISHER #####
        ########################
        u1=c()
```

```
repeat{
        u1=runif(n)
        if(testFisher(u1))
                break
}
result['Fisher',z,'MediaPVSUP']=result['Fisher',z,'MediaPVSUP']+length(u1[u1>=alfa])
y=u1
while(testFisher(y)){
        y <- c(y,runif(1,alfa,1))
        n0Fisher=n0Fisher+1
}
##################################
#### TESTE MEDIA GEOMETRICA #####
##################################
u1=c()
repeat{
        u1=runif(n)
        if(testMediaGeometrica(u1))
                break
}

    result['MediaGeometrica',z,'MediaPVSUP']=result['MediaGeometrica',z,'MediaPVSUP']+length(
u1[u1>=alfa])
y=u1
while(testMediaGeometrica(y)){
        y <- c(y,runif(1,alfa,1))
        n0MediaGeometrica = n0MediaGeometrica +1
}
```

```
######################
#### TESTE LOGIT #####
######################
u1=c()
repeat{
        u1=runif(n)
        if(testLogit(u1))
                break
}
result['Logit',z,'MediaPVSUP']=result['Logit',z,'MediaPVSUP']+length(u1[u1>=alfa])
y=u1
while(testLogit(y)){
        y <- c(y,runif(1,alfa,1))
        n0Logit = n0Logit +1
}
########################
#### TESTE TIPPETT #####
########################
u1=c()
repeat{
        u1=runif(n)
        if(testTippett(u1))
                break
}

result['Tippett',z,'MediaPVSUP']=result['Tippett',z,'MediaPVSUP']+length(u1[u1>=alfa])
        y=u1
        while(testTippett(y)){
                y <- c(y,runif(1,alfa,1))
                n0Tippett = n0Tippett +1
        }
```

```
            n0sStouffer=c(n0sStouffer,n0Stouffer);
            n0sFisher=c(n0sFisher,n0Fisher);
            n0sMediaGeometrica =c(n0sMediaGeometrica,n0MediaGeometrica);
            n0sLogit=c(n0sLogit,n0Logit);
            n0sTippett=c(n0sTippett,n0Tippett);
            p=p+1
        }

        result['Stouffer',z,'MediaPVSUP']=result['Stouffer',z,'MediaPVSUP']/nruns
        result['Fisher',z,'MediaPVSUP']=result['Fisher',z,'MediaPVSUP']/nruns
        result['MediaGeometrica',z,'MediaPVSUP']=result['MediaGeometrica',z,'MediaPVSUP']/nruns
        result['Logit',z,'MediaPVSUP']=result['Logit',z,'MediaPVSUP']/nruns
        result['Tippett',z,'MediaPVSUP']=result['Tippett',z,'MediaPVSUP']/nruns

        result['Stouffer',z,'Media']=round(mean(n0sStouffer),0)
        result['Stouffer',z,'DP']=sd(n0sStouffer)
        result['Fisher',z,'Media']=round(mean(n0sFisher),0)
        result['Fisher',z,'DP']=sd(n0sFisher)
        result['Logit',z,'Media']=round(mean(n0sLogit),0)
        result['Logit',z,'DP']=sd(n0sLogit)
        result['MediaGeometrica',z,'Media']=round(mean(n0sMediaGeometrica),0)
        result['MediaGeometrica',z,'DP']=sd(n0sMediaGeometrica)
        result['Tippett',z,'Media']=round(mean(n0sTippett),0)
        result['Tippett',z,'DP']=sd(n0sTippett)

}
```