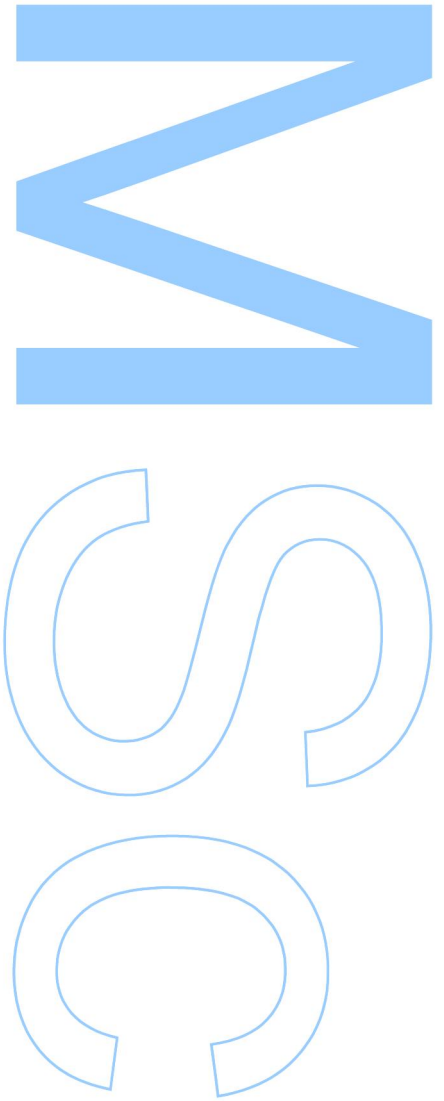
Estimação e controlo das taxas de erro em testes de hipóteses múltiplos – métodos de reamostragem

Carla Patrícia Rodrigues Pereira
Mestrado em Matemática
Departamento de Matemática
2020

Orientador
Ana Rita Pires Gaio, Professor Auxiliar
Faculdade de Ciências da Universidade do Porto

Coorientador
Margarida Maria Araújo Brito, Professor Associado
Faculdade de Ciências da Universidade do Porto



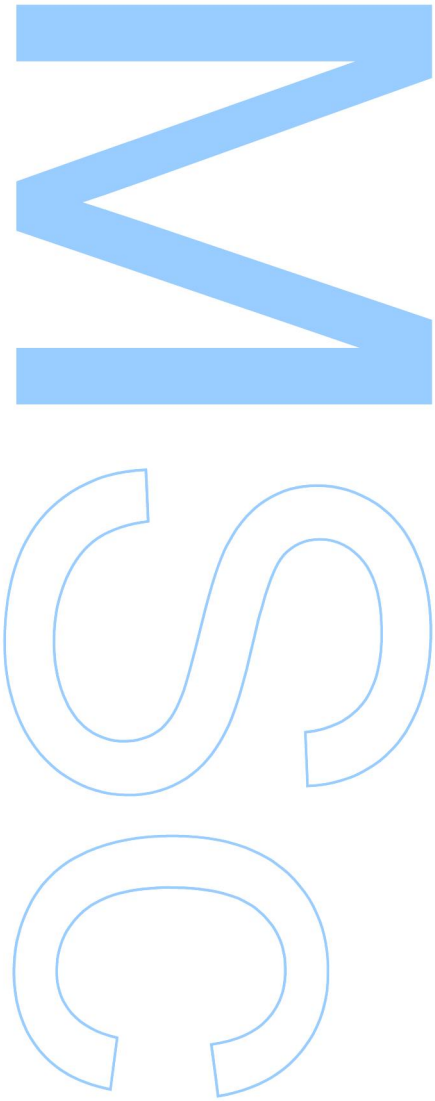
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



Agradecimentos

A realização desta dissertação de mestrado contou com o apoio e incentivo de diversas pessoas, sem as quais não se teria tornado realidade e aos quais estou profundamente grata.

À minha orientadora, Professora Ana Rita Pires Gaio e coorientadora Maria Margarida Araújo Brito agradeço pela dedicação, disponibilidade, paciência e apoio total. O incentivo constante permitiu que nunca desistisse mesmo nos momentos mais complicados.

À minha família, pilar da minha vida e apoio sempre presente em todas as minhas decisões pela crença constante em mim e no meu percurso. Por todas as palavras de coragem e orgulho que sempre expressaram neste projeto.

À impulsionadora da realização deste mestrado, a minha amiga Raquel Magano que acreditou sempre em mim e nas minhas capacidades, mesmo quando duvidei que seria capaz de o fazer. Um apoio sempre presente em todos os momentos com as palavras certas.

Aos meus colegas de mestrado Cláudio, Fernando e Jorge obrigada pela paciência, ajuda e força que me deram nestes dois anos.

Por fim, mas não menos importante, à minha amiga Paula Ferreira que permitiu e facilitou que pudesse trabalhar ao mesmo tempo que realizei o mestrado.

Resumo

Em testes de hipóteses múltiplos são testadas várias (por vezes milhares) hipóteses nulas simultaneamente. Se se realizar um teste de hipóteses de cada vez o erro global aumenta e facilmente poderá haver falsas rejeições. Este efeito indesejado é designado na literatura por efeito de multiplicidade. Nesta dissertação, é abordado o problema da estimação e controlo das taxas de erro em testes múltiplos. Os métodos clássicos de Bonferroni, Sidak e Holm são revisitados mas o foco são os métodos que Westfall e Young (1993) [11] propuseram que recorrem ao valor-p ajustado por reamostragem para controlar a taxa de erro de família (FWER). Um abordagem diferente apresentam Benjamini e Hochberg (1995) [1] centrada no controlo da taxa de erro de falsos significativos (FDR) em vez do FWER, no sentido de haver um ganho na potência do teste.

Palavras-chave: multiplicidade, estimação, taxa de erro de família, taxa de falsos significativos, controlo forte, valores-p ajustados, reamostragem, testes de permutações

Abstract

In multiple testing, several (sometimes thousands) of null hypotheses are tested simultaneously. If one hypothesis test is performed at a time, the overall error increases and there can easily be false rejections. This unwanted effect is referred to in the literature as the multiplicity effect. In this dissertation, the problem of estimating and controlling error rates in multiple tests is addressed. The classic methods of Bonferroni, Sidak and Holm are revisited but the focus is on the methods that Westfall and Young (1993) [11] proposed resampling based p-value adjustment to control the family-wise error rate (FWER). A different approach presents Benjamini and Hochberg (1995) [1] centered on controlling the false discovery rate (FDR) instead of FWER, in the sense that there is a gain in test power.

Keywords: multiplicity, estimation, family-wise error rate, false discovery rate, strong control, adjusted p-value, resampling, permutation tests

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Uma metodologia básica e fundamental em inferência estatística são os testes de hipóteses. Em geral, ao testar uma hipótese nula, dois tipos de erros podem ser cometidos: um falso positivo ou erro de tipo I, que surge ao rejeitar uma hipótese nula verdadeira, e um falso negativo, ou erro de tipo II, que surge quando não se rejeita a hipótese nula e esta é falsa. No caso em que várias hipóteses são testadas simultaneamente, e cada teste tem uma probabilidade de erro de tipo I especificada, a probabilidade de cometer um erro global aumenta, muitas vezes de forma acentuada, com o número de hipóteses. Isso pode levar a muitos falsos positivos e este fenómeno é geralmente chamado de problema da multiplicidade.

O objetivo desta dissertação é o estudo do problema da multiplicidade envolvendo diferentes métodos de análise e controlo das taxas de erro recorrendo a métodos de reamostragem. As principais referências bibliográficas que serviram de suporte ao trabalho realizado são o livro de Westfall e Young (1993) [11] e os artigos científicos de Youngchao Ge, Sandrine Dudoit e Terence Speed (2003) [4] para a estimação e controlo da taxa de erro de família (FWER) (a definir posteriormente) e o artigo de Benjamini e Hochberg (1995) [1] para o controlo de falsos significativos (FDR)(a definir posteriormente).

No capítulo 1 apresentam-se os conceitos básicos de um teste de hipóteses simples incluindo os vários tipos de erros associados a esse procedimento.

No capítulo 2 são apresentadas formalmente as definições de taxas de erro, valor-p bruto e ajustado, controlo forte e fraco das taxas de erro e uma breve comparação entre as diversas taxas de erro. Também são referidos os três tipos distintos de procedimentos de controlo do erro em teste múltiplos com base em valores-p ajustados: procedimentos de *single step*, *step down* e *stepwise*.

O capítulo 3 apresenta três dos métodos clássicos mais comuns em testes de hipóteses múltiplos: o método de Bonferroni, de Šidák e de Holm. Estes procedimendos devolvem valores-p ajustados que são depois diretamente comparados com o nível de significância α. O capítulo termina com a apresentação de um exemplo ilustrativo de aplicação de cada um dos métodos.

Como referido atrás, esta dissertação foca-se no controlo e estimação das taxas de erro em testes de hipóteses múltiplos usando métodos de reamostragem. O processo de reamostragem consiste num conjunto de métodos que se baseiam no cálculo de estimativas a partir de amostragens sucessivas da mesma amostra. Especificamente nos testes de permutações os dados são permutados repetidamente e a estatística de teste é calculada para cada uma das permutações resultantes. A grande vantagem dos testes de permutação sobre os testes clássicos é o facto de não ser necessário supor a aleatoriedade na recolha dos dados nem conhecer a distribuição dos mesmos. No capítulo 4 são referidas as orientações para a aplicação de teste de permutações em testes de hipóteses simples e múltiplos, com destaque para uma propriedade importante (*subset pivotality*), sob a qual é possível obter um controlo forte de um dado erro designado erro de família (FWER - *Family Wise Error Rate*), associado ao conjunto de testes considerados. A violação da primeira diretriz pode reduzir a potência de um teste. A segunda diretriz é importante quando a conclusão de um teste é ambígua. Não tem relação direta com a potência, mas melhora o nível de precisão de um teste.

No capítulo 5 apresentam-se os algoritmos de reamostragem minP e maxT de Westfall e Young (1993) [11] para o controlo do FWER. Os algoritmos são depois implementados computacionalmente em linguagen R, e ilustrados com um exemplo.

O controlo do FWER, que é a probabilidade de cometer pelo menos uma rejeição é geralmente exigido no sentido forte, ou seja, sob todas as configurações de subconjuntos de hipóteses nulas.

Em diversos problemas de multiplicidade o número de rejeições erradas deve ser tido em conta e não só a questão de se ter cometido pelo menos um erro. No entanto, ao mesmo tempo, a gravidade do prejuízo sofrido por rejeições erradas está inversamente relacionada com o número de hipóteses rejeitadas.

No capítulo 6 é apresentada uma abordagem diferente, segundo Benjamini e Hochberg (1995) [1] para testes múltiplos. Eles afirmam que, em muitas situações, o controlo do FWER pode adotar procedimentos excessivamente conservativos e tolerar alguns erros de tipo I, desde que esse número seja pequeno em comparação com o número de hipóteses rejeitadas. Essas considerações levam a uma abordagem menos conservativa que exige controlar o valor esperado da proporção de erros de tipo I de entre as hipóteses rejeitadas designada FDR - *False Discovery Rate*.

Capítulo 2

Testes de hipóteses simples

Os testes de hipóteses são dos procedimentos do âmbito da estatística inferencial, a par da estimação intervalar, mais usados em estudos estatísticos. O seu objetivo principal é analisar a compatibilidade das observações realizadas com uma hipótese formulada à priori sobre a população. Começa-se por formular uma hipótese sobre a população, traduzindo normalmente uma afirmação de "ausência de efeito" ou "ausência de diferença". Por oposição a esta, é formulada uma outra hipótese que se suspeita ser verdadeira. A primeira hipótese designa-se por *hipótese nula* e denota-se por H_0, enquanto que à segunda chama-se *hipótese alternativa* e denota-se por H_1. A hipótese alternativa indica, por exemplo, quais os valores do parâmetro que devemos considerar contra o valor especificado na hipótese nula. Ambas as hipóteses devem ser formuladas antes da recolha dos dados que se vão utilizar para efetuar o teste.

Um teste de hipóteses surge como um procedimento estatístico que nos permite medir, em termos de probabilidade, a evidência com que os dados se afastam da hipótese nula. Para avaliar essa evidência, define-se uma variável aleatória - designada por estatística de teste - que mede a compatibilidade entre a hipótese nula e as observações realizadas, e a partir da estatística de teste constroem-se regiões de rejeição que conduzem à rejeição da hipótese nula, também denominadas regiões críticas do teste. A probabilidade da estatística de teste tomar valores na região crítica quando H_0 é verda-

deira designa-se nível de significância e é comummente representada por α. O conceito de valor-p que é a probabilidade de se obter uma estatística de teste tão ou mais extrema que aquela observada na amostra, sob a hipótese nula pode também ser utilizada para concluir a rejeição ou não de H_0.

Sob H_0, a estatística de teste segue uma determinada distribuição (conhecida ou empírica). Para determinar a evidência contra H_0, avalia-se o valor da estatística de teste na amostra e compara-se esse valor com os valores-fronteira da região de rejeição do teste. Se o valor da estatística de teste cair na região de rejeição, rejeita-se H_0. Pela definição de valor-p, se este for inferior ao nível de significância, conclui-se que o correto é rejeitar a hipótese nula.

Exemplo 2.1. Suponha-se que se pretendem testar as seguintes hipóteses:

$$H_0 : \mu = 30 \qquad vs \qquad H_1 : \mu \neq 30$$

com um nível de significância 0.05. A partir de uma amostra aleatória $x_1, ..., x_n$ de uma variável aleatória X assumindo que X segue uma distribuição normal com variância (populacional) conhecida. Sob H_0, sabe-se que

$$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

A variável aleatória T é exatamente a estatística de teste nesta situação. A regra de decisão do teste é rejeitar H_0 se $|t_{obs}| \in R_c$, onde $R_c = \{t \in \mathbb{R} : |t| \geq Z_{1-\frac{\alpha}{2}}\}$ é a região de rejeição. Por exemplo, para $n = 36$, $\sigma = 6$ e $\overline{X} = 34$ tem-se $t_{observado} = 4 \in R_c$ logo rejeita-se H_0, a um nível de significância 0.05.

A escolha de um nível de significância α diz respeito a uma margem de erro tolerável e que sustenta a rejeição da hipótese nula. A sua escolha baseia-se nos custos de rejeitar uma hipótese verdadeira em comparação com não rejeitar uma hipótese falsa. Devido à dificuldade de quantificar esses custos e à subjetividade envolvida, geralmente é fixado num nível convencional, igual a 0.05. Os riscos envolvidos na rejeição incorreta

da hipótese nula são avaliado pela existência de dois tipos de erros:

I Rejeitar a hipótese H_0, quando ela é verdadeira.
II Não rejeitar a hipótese H_0, quando ela é falsa.

A Tabela seguinte resume os erros cometidos:

	Não rejeitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta	Erro de tipo I
H_0 falsa	Erro de tipo II	Decisão correta

Tabela 2.1: Tipos de erros

Se a hipótese H_0 for verdadeira e não rejeitada, ou falsa e rejeitada, a decisão estará correta. No entanto, se a hipótese H_0 for rejeitada sendo verdadeira, ou se não for rejeitada sendo falsa, a decisão estará errada. O primeiro destes erros é chamado de erro de tipo I e a probabilidade de cometê-lo é α a área introduzida atrás. O segundo é chamado de erro de tipo II e a probabilidade de cometê-lo é denotada por β. Assim tem-se,

$$\alpha = P(\text{Erro do tipo I}) = P(\text{rejeitar } H_0 \mid H_0 \text{ verdadeira})$$

$$\beta = P(\text{Erro do tipo II}) = P(\text{não rejeitar } H_0 \mid H_0 \text{ falsa})$$

A potência de um teste de hipóteses é a probabilidade $1-\beta$ de rejeitar a hipótese nula quando ela é falsa, ou seja,

$$1 - \beta = P\left(\text{rejeitar } H_0 \mid H_0 \text{ falso}\right)$$

Normalmente, escolhe-se um teste que apresenta a menor probabilidade de cometer um erro de tipo II, β, enquanto controla α, a um certo nível. Por outras palavras, maximiza-se a potência $1-\beta$ enquanto se mantém a probabilidade de erro de tipo I, a um nível desejado, tal como é descrito em Casella (2002) [2].

Capítulo 3

Testes de hipóteses múltiplos

Atualmente em diversas áreas, nomeadamente no campo da genómica existem grandes problemas de multiplicidade, uma vez que se pretende testar milhares de hipóteses simultaneamente. Por esta razão, os testes de hipóteses múltiplos são cada vez mais uma área de grande interesse.

Uma abordagem comum em testes de hipóteses múltiplos, usado para controlar o efeito de multiplicidade, consiste em recorrer a procedimentos clássicos de comparações múltiplas (MCPs - *Multiple Comparison Procedures*). Estes são procedimentos estatísticos que têm como finalidade o controlo da probabilidade de cometer um erro de tipo I.

Considerem-se *m* testes, numa situação em que se dispõem de *n* observações para cada teste. Pretendem-se testar m hipóteses nulas simultaneamente, $H_0^1, H_0^2, ..., H_0^m$.

Designe-se por $p_1, p_2, ..., p_m$ os para valores-p correspondentes.

A primeira ideia seria realizar um teste de hipótese para cada variável. Para um nível de significância α, a probabilidade de cometer um erro de tipo I na totalidade dos m testes pode crescer até aproximadamente $1-(1-\alpha)^m$. De facto consideram-se m testes estatísticos como acima e $H_0^c = H_0^1 \cap ... \cap H_0^m$ é a hipótese nula completa.

Admitindo que os m acontecimentos B_i (tomar a decisão correta ao rejeitar H_0^i) são independentes entre si, tem-se que:

$$\begin{aligned} P(rejeitar H_0 | H_0\, verdadeiro) &= 1 - P(nao \quad rejeitar H_0 | H_0 \quad verdadeiro) \\ &= 1 - P(nao \quad rejeitar H_0^1 \cap ... \cap nao \quad rejeitar H_0^m) \\ &= 1 - (1-\alpha)^m. \end{aligned}$$

Exemplo 3.1. Considere-se a análise de expressão diferencial de 1000 genes, em 2 condições distintas num chip (2 condições) e em que nenhum gene é diferencialmente expresso. Ao efetuar 1000 testes de hipóteses, com $\alpha = 0.01$ em cada teste, existe uma probabilidade máxima, ou igual no caso de independência, de 0.99996 de encontrar incorretamente pelo menos um valor-p inferior a α num dos testes individuais.

Para vários testes em simultâneo, um valor-p muito baixo num teste individual não é garantia de tomar uma decisão correta ao rejeitar H_0. Um dos métodos usados para contornar este problema é a distribuição mínima do valor-p em vez da distribuição individual do valor p.

Uma abordagem ao problema dos testes de hipóteses múltiplos consiste dos seguintes passos:

(i) calcular a estatística de teste T_i para cada teste i;

(ii) aplicar o procedimento de comparações múltiplas para determinar quais as hipóteses que se rejeitam, controlando o erro de tipo I.

A estatística de teste usada depende da experiência que se está a realizar e do tipo de variável de resposta. Para cada variável i, a hipótese nula H_0^i é testada baseada na estatística de teste T_i e a realização da mesma denota-se por t_i. Assume-se que a hipótese nula H_0^i é rejeitada para valores *suficientemente grandes* de $|T_i|$ (teste de hipóteses bilateral).

3.1 Tipos de erros em testes múltiplos

Considere-se o problema de testar simultaneamente *m* hipóteses nulas H_0^i, $i = 1, ..., m$. Numa interpretação frequencista, a situação pode ser resumida na tabela seguinte, retirada de Benjamini e Hochberg (1995) [1].

Number of	Number not rejected	Number rejected	
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	m_1
	$m - R$	R	m

Tabela 3.1: Número de erros e atitudes corretas cometidos num teste de *m* hipóteses nulas

Considere-se que as *m* hipóteses são previamente definidas, desconhecendo-se m_0 e m_1, que representam , respetivamente, as hipóteses nulas verdadeiras e falsas. Seja R o número de hipóteses nulas rejeitadas, V o número de hipóteses nulas verdadeiras rejeitadas de forma errada e S o número de hipóteses nulas falsas que são rejeitadas. Destas três variáveis aleatórias, apenas R pode ser observada.

Em geral, pretende-se minimizar o número V de falsos positivos (erros de tipo I) e o número T de falsos negativos (erros de tipo II). Para cada teste, pode cometer-se um erro de tipo I ou de tipo II e pretende-se analisar a forma como estes se combinam em taxas de erro globais.

Erro de tipo I

Ao efetuar um teste com uma única hipótese nula H_0, a probabilidade de cometer um erro de tipo I é controlada a um certo nível α. Este nível é escolhido através de um valor crítico c_α tal que $P(|T| \geq c_\alpha | H_0) \leq \alpha$ e a regra de decisão consiste em rejeitar H_0 quando $|T| \geq c_\alpha$. Existem diversas generalizações das taxas de erro tipo I para testes de hipóteses múltiplos que se irão apresentar a seguir.

Seja *m* o número de hipóteses nulas testadas simultaneamente, V o número de erros de tipo I (falsos positivos) e R o número total de hipóteses nulas rejeitadas. As taxas gerais de erro de tipo I são descritas a partir destes números, tal como em Youngchao Ge, Sandrine Dudoit e Terence Speed(2003) [4].

- *Per-Comparison Error Rate* (PCER)

  O PCER é definido como a razão entre o valor esperado de erros de tipo I e o número total de hipóteses, ou seja,

  $$PCER = \frac{E(V)}{m}.$$

- *Per-family Error Rate* (PFER)

  O PFER é apresentado como o valor esperado do número de erros de tipo I, isto é,

  $$PFER = E(V).$$

- *Taxa de Erro de Família* (FWER)

  O FWER é a probabilidade de rejeitar pelo menos uma falsa hipótese nula, ou seja, a probabilidade de cometer pelo menos um erro de tipo I.

  $$FWER = P(V \geq 1).$$

  O FWER é muito usado em comparações múltiplas.

- *Taxa de Falsos Significativos* (FDR)

  A definição mais usual é

  $$FDR = E(V/R),$$

  ou seja, o valor esperado da proporção entre o número de falsas rejeições e o número de hipóteses rejeitadas. No entanto, as diferentes formas de tratar o caso $R = 0$ levam a diferentes definições.

Tal como em Benjamini e Hochberg (1995) [1], considere-se a variável aleatória $Q = \frac{V}{V+S}$ que representa a proporção de hipóteses nulas rejeitadas de forma errada. Define-se $R = 0$ quando $V + S = 0$, uma vez que nenhuma falsa rejeição é cometida. Como Q é uma variável aleatória desconhecida então $q = \dfrac{v}{v+s}$ também o é, mesmo após a recolha e análise de dados. Assim, define-se o FDR, Q_e como o valor esperado de Q:

$$Q_e = E(Q) = E\left\{\frac{V}{V+S}\right\} = E\left(\frac{V}{R}\right)$$

ou de forma análoga,

$$FDR = E\left[\frac{V}{R}\mathbb{1}_{\{R>0\}}\right] = E\left[\frac{V}{R}|R>0\right]P(R>0).$$

O FDR tem duas propriedades importantes:

a) Se todas as hipóteses nulas são verdadeiras, o FDR é equivalente ao FWER: neste caso $s = 0$ e $v = r$ logo se $v = 0$ então $Q = 0$ e se $v > 0$ então $Q = 1$, o que leva a que $P(V \geq 1) = E(Q)$. Portanto, o controlo do FDR implica o controlo do FWER no sentido fraco.

b) Quando $m_0 < m$, o FDR é menor ou igual ao FWER: neste caso se $v > 0$ então $\frac{v}{r} < 1$ levando a que, aplicando o valor esperado em ambos os membros se obtenha $P(V \geq 1) \geq E(Q)$. Assim, qualquer procedimento que controla o FWER também controla o FDR. No entanto, se um procedimento controlar apenas o FDR, ele pode ser menos rigoroso e espera-se um aumento da potência. Em particular, quanto maior o número de hipóteses nulas não verdadeiras, maior tende a ser S e assim a diferença entre as taxas de erro.

- *Positive False discovery rate* (pFDR)

  É definido como o valor esperado condicionado da proporção de erros de tipo I entre as hipóteses nulas rejeitadas, dado que pelo menos uma hipótese nula é

rejeitada,

$$pFDR = E\left[\frac{V}{R}|R > 0\right]$$

Diz-se que um procedimento de testes múltiplos controla uma taxa de erro de tipo I, a um nível α, se essa taxa de erro é menor ou igual a α quando esse procedimento especificado é aplicado com o intuito de obter uma lista R de hipóteses rejeitadas. Por exemplo, o FWER é controlado a um nível α por um procedimento de testes múltiplos se $FWER \leq \alpha$; o mesmo se aplica aos restantes tipos de taxas de erro.

Erro de tipo II

Considerando os vários procedimentos de testes que controlam uma determinada taxa de erro de tipo I, a um nível aceitável α, procuram-se procedimentos que maximizem a potência do teste, ou seja, que minimizem a taxa de erro de tipo II. Assim como nas taxas de erro de tipo I, pode generalizar-se, para testes múltiplos, o conceito de potência. Segundo Shaffer (1995) [9], a potência pode definir-se como:

- A probabilidade de rejeitar pelo menos uma hipótese nula falsa, $P(S \geq 1) = P(T \leq m_1 - 1)$;
- A probabilidade média de rejeitar hipóteses nulas falsas, $\frac{E(S)}{m_1}$;
- A probabilidade de rejeitar todas as hipóteses nulas, $P(S = m_1) = P(T = 0)$.

De forma análoga ao FDR, segundo Sandrine Dudoit, Juliet Popper Shaffer e Jennifer C. Boldrick (2003) [3], também se pode definir a potência como

$$E\left(\frac{S}{R}|R > 0\right) P\left(R > 0\right) = P\left(R > 0\right) - FDR$$

3.2 Controlo fraco e controlo forte

O controlo das taxas de erro de tipo I pode ser feito no sentido fraco ou forte.

- **Controlo fraco**: Assumindo $H_0^c = \bigcap_{i=1}^{m} H_0^i$, a probabilidade de cometer pelo menos um erro de tipo I num teste individual é α, ou seja,

$$P(\#\{falsos \quad positivos\} > 1 | H_0^c) = \alpha$$

  O controlo fraco refere-se a controlar a taxa de ocorrências de erros de tipo I apenas quando todas as hipóteses nulas forem verdadeiras, isto é, sob a distribuição nula que satisfaça a hipótese nula completa H_0^c com $m_0 = m$.

- **Controlo forte**: Refere-se ao controlo da taxa de ocorrência do erro de tipo I para qualquer combinação de hipóteses nulas (verdadeiras ou falsas), ou seja, para qualquer subconjunto de hipóteses nulas verdadeiras.

Em geral, a hipótese nula completa H_0^c não é realista e o controlo fraco não é satisfatório. Na realidade, algumas hipóteses nulas podem ser verdadeiras e outras falsas, mas esse subconjunto é desconhecido. O controlo forte garante que o erro de tipo I é controlado sob a verdadeira e desconhecida distribuição dos dados.

Os conceitos de controlo forte e fraco aplicam-se a todas as taxas de erro mencionadas anteriormente (PCER, PFER, FWER e FDR). Neste trabalho considera-se que as probabilidades e os valores esperados são calculados para a combinação de hipóteses nulas verdadeiras e falsas correspondentes à verdadeira distribuição dos dados.

3.3 Comparação das taxas de erro de tipo I

Em geral, segundo Sandrine Dudoit, Juliet Shaffer e Jennifer Boldrick (2003) [3] para um dado procedimento de testes múltiplos, $PCER \leq FDR \leq FWER \leq PFER$. Como $0 \leq V \leq R \leq m$ e $R = 0$ implica $V = 0$, então

$$\frac{V}{m} \leq \frac{V}{R}\mathbb{1}_{\{R>0\}} \leq \mathbb{1}_{\{V>0\}} \leq V$$

Aplicando o valor esperado à expressão anterior prova-se o pretendido.

Para ilustrar as diferentes propriedades do erro de tipo I, suponha-se que cada hipótese H_0^i é testada individualmente, a um nível α_i, e a regra de decisão é baseada apenas nesse teste. Sob a hipótese nula completa, o PCER é simplesmente a média da α_i e o PFER é o somatório de α_i. Por outro lado, o FWER é uma função que não só envolve α_i como também a distribuição conjunta da estatística de teste T_i:

$$PCER = \frac{\alpha_1 + ... + \alpha_m}{m} \leq max(\alpha_1, ..., \alpha_m) \leq FWER \leq PFER = \alpha_1 + ... + \alpha_m$$

A abordagem clássica dos testes múltiplos exibe um controlo forte do FWER (procedimento de Bonferroni, apresentado mais à frente). A proposta de Benjamini e Hochberg (1995) ([1]) controla o FWER no sentido fraco (desde que $FDR = FWER$ sob a hipótese nula completa) e por outro lado pode ser menos conservativo do que FWER.

Tal como referido em segundo Sandrine Dudoit, Juliet Shaffer e Jennifer Boldrick (2003) [3], os procedimentos que controlam o PCER são geralmente menos conservativos do que aqueles que controlam o FDR ou FWER, mas tendem a ignorar completamente a multiplicidade do problema.

3.4 Valores-p

Valores-p brutos

Considere-se um teste com uma única hipótese nula H_0 a um nível α e regiões de rejeição encaixadas Γ_α tal que:

a) $\Gamma_{\alpha_1} \subseteq \Gamma_{\alpha_2}$, para $0 \leq \alpha_1 \leq \alpha_2 \leq 1$;

b) $P\left(T \in \Gamma_\alpha | H_0\right) \leq \alpha$, para $0 \leq \alpha \leq 1$.

Caso se pretenda usar a estatística $|T|$ para um teste bilateral, as regiões encaixadas $\Gamma_\alpha = [-\infty, -c_\alpha] \cup [c_\alpha, \infty]$ são tais que $P\left(T \in \Gamma_\alpha | H_0\right) = \alpha$.

Em vez de simplesmente afirmar a rejeição ou não da hipótese nula H_0, o valor-p está relacionado com o teste. O valor-p para uma estatística de teste observada $T = t$ é:

$$valor - p = \min_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} P\left(T \in \Gamma_\alpha | H_0\right)$$

O valor-p é a taxa mínima de erro de tipo I sobre todas as possíveis regiões de rejeição Γ_α que contêm o valor observado $T = t$. Para um teste bilateral, $valor - p = P(|T| \geq |t||H_0)$. Quanto menor o valor-p, mais forte a evidência contra a hipótese nula. Rejeitar H_0 quando $p \leq \alpha$ fornece um controlo da taxa de erro de tipo I ao nível α. Se se estender o conceito de valor-p para uma situação de testes múltiplos, surge a definição de valor-p ajustado.

Valores-p ajustados

O termo valor-p ajustado é usado na literatura, como se pode encontrar em Shafer e Olkin (1983), Heyse e Rom(1988), Westfall e Young (1989) e Dunnet e Tamhane (1992). Wright (1992) apresenta um resumo com vários exemplos e referências adicionais para o valor-p ajustado.

Seja t_i e $p_i = P(|T_i| \geq |t_i||H_0^i)$, respetivamente, a estatística de teste e o valor-p bruto para a hipótese H_0^i, com $i = 1, ..., m$. Tal como no caso de uma única hipótese, um procedimento de testes múltiplos pode ser definido em termos de valores críticos para a estatística de teste ou para valores-p dos testes individuais. Por exemplo, rejeitar H_0^i se $|t_i| \geq c_i$ ou se $p_i \leq \alpha_i$, onde os valores críticos c_i e α_i são escolhidos para controlar uma determinada taxa de erro de tipo I (FWER, PCER, PFER ou FDR), a um nível pré-especificado α. Alternativamente, o procedimento de testes múltiplos pode ser descrito em termos de valores-p ajustados.

Dado um teste com uma única hipótese H_0^i, o valor-p ajustado correspondente pode ser definido como o menor nível de significância para o qual ainda se rejeita H_0^i, dado um procedimento de teste simultâneo (STP-*Simultaneous Test Procedure*). Um STP é um procedimento pelo qual, a um determinado nível, cada uma das hipóteses H_0^i, com $i = 1, ..., m$ é rejeitada ou não. As respetivas variáveis aleatórias para valores-p não ajustados e ajustados denotam-se por P_i e $\tilde{P}_i$, respetivamente.

Definição 3.2. Dado um procedimento de comparações múltiplas (MCP), $\mathscr{M}$, o valor-p ajustado para o FWER $p_i^{\mathscr{M}}$ de uma hipótese H_0^i, designado por $\tilde{p_i}^{\mathscr{M}_0}$, é o menor nível de significância do teste $\mathscr{M}$ que rejeita H_i enquanto controla o FWER, dados os valores-p observados, ou seja,

$$\tilde{p_i}^{\mathscr{M}} = \inf\{\alpha : H_0^i \quad rejeitada \quad se \quad FWER = \alpha\}$$

Assim, a hipótese H_0^i é rejeitada, com FWER α, se $\tilde{p}_i \leq \alpha$.

Da mesma forma podem obter-se os valores-p ajustados de outros procedimentos para as diversas taxas de erro. Por exemplo, o valor-p ajustado para controlar o FDR, segundo Yekutieli e Benjamini ([12]) é:

Definição 3.3. Para qualquer MCP, $\mathscr{M}$, o valor-p ajustado para o FDR $p_i^{\mathscr{M}}$ de qualquer hipótese H_0^i é é o menor nível de significância do teste $\mathscr{M}$ que rejeita H_0^i enquanto controla o FDR, ou seja,

$$\tilde{p_i}^{\mathscr{M}} = \inf\{\alpha : H_0^i \quad rejeitada \quad se \quad FDR = \alpha\}$$

Os procedimentos de testes múltiplos são definidos de maneira mais conveniente em termos de valores-p ajustados e alguns destes são estimados pelos métodos de reamostragem, tal como apresenta Westfall and Young ([11]).

3.5 Procedimentos para controlo das taxas de erro

Existem três tipos de procedimentos para efetuar testes múltiplos com base numa ordenação dos valores-p brutos ou das estatísticas de teste: *single-step*, *step-down* e *step-up*. Os métodos de single step consistem em procedimentos simultâneos (STP) que ajustam todas as hipóteses, independentemente da ordem da estatística de teste ou dos valores-p brutos. Assim, para cada hipótese usa-se um valor crítico que é independente dos resultados dos testes das restantes hipóteses.

Para melhorar a potência, enquanto se preserva o erro de tipo I, usam-se procedimentos *stepwise*, nos quais a rejeição de uma hipótese, em particular, é baseada não apenas no número total de hipóteses mas também nos resultados dos testes das restantes hipóteses. Este método permite diferentes técnicas de ajuste para diferentes hipóteses, dependendo de como estão ordenadas as hipóteses. O mais natural é ordená-las pelo tamanho do respetivo valor-p, diga-se $p_1 \leq .... \leq p_m$.

No método *step-down*, as hipóteses que correspondem às estatísticas de teste mais significativas (isto é, menores valores-p não ajustados ou maiores valores absolutos da estatística de teste) são consideradas sucessivamente com outros testes que dependem dos resultados anteriores. Assim que um teste falha ao rejeitar uma hipótese nula, as restantes hipóteses não serão rejeitadas. Em contraste, no procedimento *step-up*, as hipóteses que correspondem às estatísticas de teste menos significativas são consideradas sucessivamente, novamente com outros testes dependentes dos anteriores. Desta forma, assim que uma hipótese é rejeitada, todas as hipóteses seguintes são rejeitadas. Em suma, os procedimentos *step-down* ordenam os valores-p brutos ou as estatísticas de teste começando nas mais significativas enquanto que o *step-up* começa com as menos significativas.

Embora os processos *single-step* sejam simples de implementar, tendem a ser conservativos para o controlo do FWER. A melhoria na potência, preservando o controlo forte do FWER, pode ser alcançada através de procedimentos *step-down*.

Definição 3.4. Sob a hipótese nula completa H_0^c e para estatísticas de teste independentes, os valores-p não ajustados ordenados $p_1 \leq p_2 \leq ... \leq p_m$ satisfazem:

$$P\left(p_i > \frac{\alpha_i}{m}, \forall i = 1, ..., m | H_0^c\right) \geq 1 - \alpha$$

com a igualdade a verificar-se no caso contínuo.

A desigualdade anterior é conhecida como *Desigualdade de Simes*. Nos casos em que as estatísticas de teste são dependentes, Simes mostrou que a probabilidade é maior do que $1 - \alpha$. No entanto, isso em geral não é válido para todas as distribuições conjuntas.

Capítulo 4

Métodos Clássicos

De forma a facilitar a sua utilização e a estabelecer uma comparação com o caso dos testes de hipóteses simples, os procedimentos de testes múltiplos devolvem valores-p ajustados para cada um dos testes individuais que devem depois ser diretamente comparados com α. Apresentam-se três métodos clássicos de testes múltiplos, muito comuns em aplicações.

Método de Bonferroni

O método *single-step* mais simples é o método de Bonferroni, que rejeita qualquer hipótese H_0^i com valor-p não ajustado menor ou igual a $\frac{\alpha}{m}$, onde *m* é o número total de testes e $\alpha = FWER$, previamente fixado. Os valores-p ajustados para o Bonferroni *single-step* são dados por:

$$\tilde{P}_i = min(mP_i, 1) \tag{4.1}$$

O controlo forte do FWER vem da desigualdade de Boole,

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i)$$

Assuma-se, sem perda de generalidade, que as hipóteses nulas são H_0^i, com $i = 1, ..., m_0$:

$$\begin{aligned} FWER = P(V \geq 1) &= P\left[\bigcup_{i=1}^{m_0}\{\tilde{P}_i \leq \alpha\}\right], \text{pela definição} \\ &\leq \sum_{i=1}^{m_0} P\left(\tilde{P}_i \leq \alpha\right), \text{desigualdade de Boole} \\ &\leq \sum_{i=1}^{m_0} P\left(P_i \leq \frac{\alpha}{m}\right), \quad \tilde{P}_i = mP_i \\ &= \frac{m_0\alpha}{m} \leq \alpha \text{ distribuição uniforme} \end{aligned} \tag{4.2}$$

Assumindo que todos os valores-p têm distribuição uniforme $[0, 1]$, sob a hipótese nula completa, tem-se $m_0 = m$ e o limite superior da ultima desigualdade é $m\frac{\alpha}{m} = \alpha$. Desde que a probabilidade de rejeitar pelo menos uma hipótese nula seja menor do que α, o método de Bonferroni é conservativo quando as distribuições marginais dos valores-p são uniformes. Para que tal se verifique tem de se considerar a seguinte propriedade:

Considere-se uma v.a. X absolutamente contínua, com função distribuição $F_X(x)$ crescente. Então $Y = F_X(x)$ tem distribuição Uniforme(0,1).

Segundo, Westfall e Young (2003), este método é extremamente conservativo ignorando a estrutura de correlação.

Exemplo 4.1. Considere-se um teste de hipóteses múltiplos com $m = 5$ e com os seguintes valores-p ordenados:

$$0.009 \qquad 0.011 \qquad 0.012 \qquad 0.134 \qquad 0.512$$

Seja H_0^1 a hipótese nula que corresponde ao valor-p 0.009, H_0^2 a hipótese nula que corresponde ao valor-p 0.011 e assim sucessivamente. Os valores-p ajustados, com $\alpha = 0.05$, usando o método de Bonferroni são calculados da seguinte forma:

- $\tilde{P}_1 = \min\{5(0.009), 1\} = 0.045$

- $\tilde{P}_2 = \min\{5(0.011), 1\} = 0.055$
- $\tilde{P}_3 = \min\{5(0.012), 1\} = 0.060$
- $\tilde{P}_4 = \max\{5(0.134), 1\} = 0.670$
- $\tilde{P}_5 = \max\{5(0.512), 1\} = 1$

Como apenas $\tilde{P}_1 < 0.05$ então apenas se rejeita a hipótese H_0^1.

Método de Šidák

O procedimento de Šidák está relacionado com o de Bonferroni. Quando os valores-p brutos são independentes e com distribuição uniforme $U(0,1)$, o FWER é mantido de forma exata, sob a hipótese nula. Os valores-p ajustados pelo método *single-step* de Šidák são dados por:

$$\tilde{P}_i = 1 - (1 - P_i)^m \tag{4.3}$$

Este procedimento apresenta um controlo forte, do FWER como se mostra a seguir.

$$\begin{aligned} FWER = P(V > 0) &= 1 - P\left(\bigcap_{i=1}^{m_0} \{\tilde{P}_i \geq \alpha\}\right), \text{pela definição de V} \\ &= 1 - \prod_{i=1}^{m_0} P\left(\tilde{P}_i \geq \alpha\right), \text{assumindo independência} \\ &= 1 - \prod_{i=1}^{m_0} P\left(P_i \geq 1 - (1-\alpha)^{\frac{1}{m}}\right), \text{pela definicao} \quad \tilde{P}_i \\ &= 1 - \{(1-\alpha)^{\frac{1}{m}}\}^{m_0}, \quad \text{assumindo} \quad P_i \sim U(0,1) \\ &= \alpha \end{aligned} \tag{4.4}$$

Exemplo 4.2. Recuperado o exemplo 4.1 e considerando $\alpha = 0.05$, os valores-p ajustados de Šidák são:

- $\tilde{P}_1 = 1 - (1 - 0.009)^5 = 0.0442$

- $\tilde{P_2} = 1 - (1 - 0.011)^5 = 0.0538$
- $\tilde{P_3} = 1 - (1 - 0.012)^5 = 0.0586$
- $\tilde{P_4} = 1 - (1 - 0.134)^5 = 0.5129$
- $\tilde{P_5} = 1 - (1 - 0.512) = 0.9723$

Como, de entre os valores-p ajustados, apenas $\tilde{P_1} < 0.05$ então, tal como para o método de Bonferroni, só se rejeita a hipótese H_0^1.

Os procedimentos de Bonferroni e Šidák estabelecem condições exigentes para a rejeição de pelo menos uma hipótese nula dado que, quando m é grande, $\frac{\alpha}{m}$ e $1-\left(\sqrt[m]{1-\alpha}\right)$ são muito pequenos. A correção nos erros de tipo I individuais implica uma perda de potência no teste múltiplo: para m grande, é difícil rejeitar as hipóteses nulas que são falsas. Em diversas situações, a estatística e consequentemente os valores-p não independentes, por exemplo na anova o que levou ao desenvolvimento de novos procedimentos, apresentados no capitulo 5.

Método de Holm

O procedimento de Holm (1979) [7] é a versão *step-down* do procedimento de Bonferroni e é menos conservativo e mais potente do que este, segundo Westfall e Young (1993) [11].

Sejam $p_1 \leq p_2 \leq ... \leq p_m$ os valores-p brutos ordenados e $H_0^1, ..., H_0^m$, as hipóteses nulas correspondentes. Para controlar o FWER, a um nível α, Holm em 1979 desenvolveu o seguinte procedimento:

- **Passo 1**:comparar p_1 com $\frac{\alpha}{m}$:
  - se $p_1 > \frac{\alpha}{m}$, não rejeitar nenhuma H_0^i e terminar o procedimento.
  - se $p_1 \leq \frac{\alpha}{m}$, então rejeitar H_0^1 e passar ao ponto 2.

- **Passo 2**: comparar p_2 com $\frac{\alpha}{m-1}$:
  - se $p_2 > \frac{\alpha}{m-1}$, não rejeitar nenhuma hipótese nula do conjunto $H_0^2, ..., H_0^m$ e terminar o procedimento.
  - se $p_2 \leq \frac{\alpha}{m-1}$, então rejeitar H_0^2 e passar para o passo 3;
- iterar o procedimento
- **Passo i**: comparar P_i com $\frac{\alpha}{m-i+1}$:
  - se $P_i > \frac{\alpha}{m-i+1}$, não rejeitar nenhuma hipótese nula do conjunto $H_0^1, ..., H_0^m$ e terminar o procedimento.
  - se $P_i \leq \frac{\alpha}{m-i+1}$, então rejeitar H_0^i e passar para o passo i+1.

Assim, os valores-p ajustados do método *step-down* Holm define-se como:

$$\tilde{P}_{r_i} = \max_{k=1,...,i} \{min((m-k+1)p_k, 1)\} \tag{4.5}$$

Note-se que, no passo i, o método de Holm compara P_i com $\frac{\alpha}{m-i+1}$, que é maior do que $\frac{\alpha}{m}$ do método de Bonferroni.

No procedimento, observe-se que tomar os máximos sucessivos das quantidades $min((m-k+1)p_k, 1)$ impõe a monotonicidade dos valores-p ajustados, isto é,

$\tilde{p}_{r1} \leq \tilde{p}_{r2} \leq .... \leq \tilde{p}_{rm}$ e pode rejeitar-se uma hipótese específica somente se todas as hipóteses com os menores valores-p ajustados forem rejeitados previamente.

Para mostrar que este método controla o FWER, considere-se que i hipóteses nulas são verdadeiras, com $0 < i < m$. Se $i = m$, o teste da primeira etapa resulta num erro de tipo I com probabilidade menor ou igual a α. Se $i = m - 1$, um erro pode ocorrer na primeira etapa mas certamente irá ocorrer na segunda e a probabilidade de erro de tipo I é menor ou igual a α. Tal verifica-se porque há $i - 1$ hipóteses verdadeiras e nenhuma pode ser rejeitada a menos que pelo menos uma tenha associado um valor-p $\leq \frac{\alpha}{i-1}$. De forma similar, qualquer que seja o valor de i, um erro de tipo I pode ocorrer numa etapa inicial, mas certamente ocorrerá se houver uma rejeição na etapa $m-i+1$, caso em que

a probabilidade de erro de tipo I é menor ou igual a α. Assim, o $FWER \leq \alpha$ para todas as configurações possíveis de hipóteses nulas verdadeiras e falsas.

Considere-se, o exemplo seguinte retirado de Westfall e Young (2003), que aplica o procedimento de Holm.

Exemplo 4.3. Considere-se o exemplo 4.1 e aplique-se o procedimento de Holm com $\alpha = 0.05$. Então:

- Rejeita-se H_0^1 porque $0.009 \leq \dfrac{0.05}{5}$ e continua-se;
- Rejeita-se H_0^2 porque $0.011 \leq \dfrac{0.05}{4}$ e continua-se;
- Rejeita-se H_0^3 porque $0.012 \leq \dfrac{0.05}{3}$ e continua-se;
- Como $0.134 > \frac{0.05}{2}$, o procedimento pára e não se rejeita H_0^4 e H_0^5.

Os valores-p ajustados são calculados da seguinte forma:

- $\tilde{P_1} = 5(0.009) = 0.045$
- $\tilde{P_2} = \max\{0.045, 4(0.011)\} = 0.045$
- $\tilde{P_3} = \max\{0.045, 3(0.012)\} = 0.045$
- $\tilde{P_4} = \max\{0.045, 2(0.134)\} = 0.268$
- $\tilde{P_5} = \max\{0.268, 1(0.512)\} = 0.512$

De forma semelhante, os valores-p ajustados para o método *step-down* de Sidák são:

$$\tilde{p}_{r_i} = \max_{k=1,\ldots,i} \{1 - (1 - p_{rk})^{m-k+1}\} \tag{4.6}$$

Os métodos clássicos referidos nesta secção usam desigualdades entre probabilidades e não têm em consideração as estruturas de correlação nem as caraterísticas da distribuição dos valores P_i. Em contraste, os métodos de reamostragem (próximo capítulo) incluem estruturas de correlação e caraterísticas das distribuições, o que permite valores-p ajustados menores.

Capítulo 5

Reamostragem

Na grande parte dos estudos estatísticos o objetivo é conseguir inferir resultados para a população. Assim, a inferência estatística é realizada utilizando os mais diversos métodos. Um deles é a reamostragem, que consiste num conjunto de métodos que se baseiam no cálculo de estimativas a partir de amostragens repetidas dentro da mesma amostra (única). Existem diversos tipos de reamostragem: testes de permutações, validação cruzada, Jackknife e bootstrap.

A reamostragem calcula uma distribuição estatística empírica, criando várias amostras a partir da amostra original. A computação passa a ser de extrema importância para a estatística pois é através desta que é estimado um valor estatístico para cada amostra.

Os métodos semelhantes ao método de Bonferroni usam desigualdades de probabilidades, ignorando as estruturas de correlação e outras caraterísticas das distribuições. Ao incluir a estrutura de correlação, a potência dos testes de reamostragem é, em geral, mais elevada do que a dos métodos que não têm em consideração a mesma.

5.1 Orientações para testes de hipóteses simples

O bootstrap não paramétrico é uma forma bastante versátil para análise de dados. Hall e Wilson (1991) [5] apresentam duas diretrizes para o booststrap não paramétrico para o caso univariado mas que podem ser ampliadas para o caso de hipóteses múltiplas. A primeira recomenda que a reamostragem seja feita de forma que reflita a hipótese nula, mesmo quando a verdadeira hipótese está distante da nula, ou seja, mesmo que os dados não pertençam a uma população que satisfaça H_0, a reamostragem deve ser feita de forma a que verifique H_0. Uma segunda diretriz afirma que o bootstrap deve aplicar métodos já reconhecidos como tendo boas características na construção do intervalo de confiança do problema.

A violação da primeira diretriz pode reduzir seriamente a potência de um teste. A segunda diretriz tem alguma importância quando a conclusão de um teste é ambígua e apesar de não ter influência direta na potência, melhora a precisão do nível de um teste.

Primeira diretriz

Suponha-se, por uma questão de simplificação, que a hipótese nula é $H_0 : \theta = \theta_0$, que está a ser testada contra a hipótese alternativa bilateral $H_1 : \theta \neq \theta_0$.

Seja $\hat{\theta}$, uma função da amostra $X_1, ..., X_n$, que denota um estimador desconhecido de θ e $\hat{\theta}^*$ o valor de $\hat{\theta}$ calculado para uma nova amostra$X_1^*, .., X_n^*$, retirada da amostra com reposição. Um teste de H_0 contra H_1 geralmente é baseado na diferença $\hat{\theta} - \theta_0$, cuja distribuição sob Ho, se pretende estimar. Note-se que é muito importante estimar esta distribuição sob H_0, mesmo quando a amostra é retirada de uma população que não satisfaz H_0. Assim, a primeira diretriz refere que se deve reamostrar $\hat{\theta}^* - \hat{\theta}$ e não $\hat{\theta}^* - \theta_0$.

Para entender a importância desta orientação observe-se que o teste rejeitará H_0 se $|\hat{\theta} - \theta_0|$ for "muito grande". Se θ_0 estiver muito distante do verdadeiro valor de θ_0 então a diferença $|\hat{\theta} - \theta_0|$ não será muito grande em comparação com a distribuição bootstrap não paramétrica de $|\hat{\theta} - \theta_0|$. Uma comparação mais significativa será com a distribuição

de bootstrap, $|\hat{\theta}^* - \hat{\theta}|$. De facto, se o verdadeiro valor de θ é θ_1, então a potência do teste de bootstrap aumenta para 1 quando $|\theta_1 - \theta_0|$ aumenta, desde que o teste se baseie na reamostragem de $|\hat{\theta}^* - \hat{\theta}|$. No entanto a potência diminuí para, no máximo, o nível de significância (à medida que $|\theta_1 - \theta_0|$ aumenta) se o teste é baseado na reamostragem $|\hat{\theta} - \theta_0|$. Assim, a primeira diretriz do teste de reamostragem tem o efeito de aumentar a potência.

Segunda diretriz

A segunda diretriz, que se apresenta de seguida, reduz o erro no nível de significância. O erro de nível é definido como a diferença entre o nível de significância real de um teste de bootstrap e o nível nominal, por exemplo 5%. Seja $\hat{\sigma}$ um desvio de $\hat{\theta}$, com a propriedade de que a distribuição de $\dfrac{\hat{\theta} - \theta_0}{\hat{\sigma}}$ está praticamente livre de parâmetros desconhecidos quando H_0 é verdadeiro. Por exemplo, pode ser aproximadamente normal $N(0, 1)$. Seja $\hat{\sigma}^*$ o valor de $\hat{\sigma}$ calculado para uma nova amostra em vez da amostra original. Assim, a distribuição bootstrap de $\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$ estima a distribuição de $\dfrac{\hat{\theta} - \theta_0}{\hat{\sigma}}$ sob a hipótese nula.

A segunda diretriz afirma que o teste deve ter por base a distribuição Bootstrap de $\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$ em vez da distribuição bootstrap de $\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}}$ ou de $\hat{\theta}^* - \hat{\theta}$. Para um teste com um nível de 5% primeiro calcula-se um número $\hat{t}$ tal que:

$$P^* \left(\frac{|\hat{\theta}^* - \hat{\theta}|}{\hat{\sigma}^*} > \hat{t} \right) = 0.05$$

onde P^* denota a medida de probabilidade sob a distribuição bootstrap e rejeita-se H_0 se $\dfrac{|\hat{\theta} - \theta_0|}{\hat{\sigma}} > \hat{t}$. Note-se que a primeira orientação foi incorporada aqui, ao usar-se $\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$ em vez de $\dfrac{\hat{\theta}^* - \theta_0}{\hat{\sigma}^*}$.

A base que suporta a segunda orientação é o facto da distribuição bootstrap de $T^* = \dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$ ser uma melhor aproximação para a distribuição de $T^* = \dfrac{\hat{\theta} - \theta_0}{\hat{\sigma}}$ sob H_0, logo a distribuição bootstrap de $S^* = \hat{\theta}^* - \hat{\theta}$ é a distribuição de $S = \hat{\theta} - \theta_0$ sob Ho. Tal pode

ser explicado pelo facto de a distribuição assintótica de *S* e S^* dependerem da escala desconhecida e existem diferenças significativas entre os fatores de escala apropriados nos casos de S e S^*, tal como se apresenta em Hall (1988).

A técnica de dividir por $\hat{\sigma}^*$ é conhecido como "pivotagem bootstrap"porque produz uma estatística $\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$ que é assintoticamente pivotal, isto é, qualquer distribuição assintótica não depende de quantidades desconhecidas. Essa técnica pode ser usada sempre que se tenha uma boa estimativa da variância (baixa variância).

5.2 Orientações para testes de hipóteses múltiplos

As diretrizes de Hall e Wilson's podem ser estendidas para testes de hipóteses múltiplos. Neste caso podem construir-se estatísticas de teste e métodos de reamostragem aplicando as diretrizes às distribuições marginais de todas as estatísticas de teste. Desta forma, a reamostragem reflete a hipótese nula completa e todas as estatísticas de teste são pivotáveis.

Aos requisitos da pivotagem no caso univariado, acrescenta-se o conceito de *subset pivotality* para que seja válido no caso multivariado. Esta propriedade representa a situação ideal para a distribuição conjunta da estatística do teste de hipóteses individual.

Definição 5.1. A distribuição dos valores-p brutos $(p_1, p_2, ..., p_m)$ satisfaz a propriedade de *subset pivotality* se, para todos os subconjuntos $L \subset \{1, 2, ..., m\}$ a distribuição do subvetor $\{p_i | i \in L\}$ é idêntica sob as restrições $\cap\{H_0^i | i \in L\}$ e H_0^c.

A condição definida anteriormente é importante por duas razões. A primeira é o facto de a reamostragem assumir a hipótese nula completa H_0^c em vez da hipótese nula parcial H_0^k e a segunda é que há controlo forte do FWER.

A propriedade introduzida pela definição 5.1 requer que a distribuição multivariada de qualquer subvetor de valores-p não seja afetado pela verdade ou falsidade das hipóteses que correspondem aos restantes valores-p não incluídos no subvetor. Nos casos típicos de análise paramétrica, a distribuição marginal de qualquer p_i é $U[0, 1]$ sempre que H_0^i

é verdadeira, portanto a condição 5.1 refere apenas a estrutura de dependência nesses casos.

5.3 Testes de Permutações

Os testes de permutações remontam a 1930 e foram apresentados pela primeira vez por Fisher (1935), Pitman (1937/1938) entre outros. Em 1935, Fisher usou as medições das alturas do milho adubado e fertilizado feitas por Charles Darwin para ilustrar a análise de medidas quantitativas usando um teste t. A hipótese nula considerada seria a amostra provir de uma população com distribuição normal de média zero. Do ponto de vista prático, as alturas das plantas de milho cruzadas e auto fertilizadas seriam as mesmas. No entanto, concluiu que o milho fertilizado cruzado é pouco mais alto que o milho auto-fertilizado.

O teste t pode ser usado na situação em as observações são extraídas de uma população com distribuição normal. Nas suas experiências, Fisher refletiu sobre até que ponto sua conclusão teria impacto se as diferenças observadas não fossem provenientes de uma distribuição normal. Neste sentido, aparecem os testes de permutações que permitem calcular valores-p ajustados sem ter uma distribuição predefinida dos valores observados.

Estes testes foram usados em amostras de tamanho reduzido e ainda, hoje em dia, apesar do atual poder computacional, o grande número de permutações associadas a amostras pequenas é ainda um desafio. Por esse motivo, na prática, normalmente apenas amostramos aleatoriamente um grande número de amostras do número total disponível. Apesar de ser algo simples de se fazer e, embora o teste não seja completamente exato, podemos usar o mínimo necessário, usando um número adequado de amostras aleatórias. Quando se usam todas as permutações possíveis, o teste é chamado de "teste de randomização". No entanto, alguns autores usam os termos "permutação"e "randomização"de forma equivalente neste contexto. E, às vezes, o termo "teste exato"é usado ambos casos.

Nos testes de permutação, não existe inicialmente uma distribuição pré-determinada para a estatística de teste e para tal é construída uma distribuição empírica/amostral através de várias reamostragens da amostra.

Considerem-se duas amostras aleatórias $X_{11}, X_{12}, ..., X_{1,n}$ de tamanho n da variável aleatória X_1 e $X_{21}, X_{22}, ..., X_{2,m}$ de tamanho m de uma variável aleatória X_2. Ao reunir os valores das duas amostras $x_{11}, x_{12}, ..., x_{1n}, x_{21}, x_{22}, ..., x_{2m}$ existem $\binom{n+m}{n}$ formas distintas de agrupar estes valores em dois grupos, um com tamanho n e outro de tamanho m. Cada um destes arranjos constitui uma permutação.

Após a formação de todas as permutações possíveis (ou um número razoável delas), para cada permutação calcula-se o valor de uma estatística de teste (por exemplo, diferença de médias, estatística do teste t, ...). Para um nível de significância α, rejeita-se H_0 se a estatística de teste associada à amostra original estiver entre os $100\alpha\,\%$ valores mais extremos dos valores obtidos das permutações. Isto é, o valor-p de permutação corresponde à percentagem de permutações com valor da estatística de teste tão ou mais extremo que o valor observado na amostra.

Capítulo 6

Algoritmos de reamostragem para controlo do FWER

Os valores-p podem ser obtidos através de reamostragem por métodos *single-step* e *step-down*. Um valor-p ajustado por *single-step* é a proporção de reamostrados para os quais o valor-p minimo é menor do que o valor-p observado. Relativamente ao método *step down* oferece uma melhoria da potência (valores-p ajustados menores) enquanto mantém o controlo forte do FWER.

Na situação em que a distribuição original é desconhecida, esta pode ser estimada por reamostragem ou por outros métodos de simulação.

Nesta secção apresentam-se os valores-p ajustados baseados em Ge, Dudoit e Speed(2003) [4] *single-step* e *step down* minP e maxT, tais como os respetivos algoritmos. No final são aplicados esses mesmo algoritmos num exemplo ilustrativo simulado no software *R*.

6.1 Procedimento minP

Os métodos clássicos referidos falham ao não incorporar a dependência e as caraterísticas da distribuição do valor-p. Em muitas situações, as estatísticas de teste e portanto, os valores-p não ajustados são dependentes. Tal acontece, por exemplo, em experimentos com microarrays de DNA, uma vez que os grupos de genes tendem a apresentar uma alta correlação. No caso em que se conhece a distribuição conjunta dos valores-p, Westfall e Young (1993) ([11]) propuseram valores-p ajustados para procedimentos de múltiplos testes que sejam menos conservativos e que tenham em conta a estrutura de dependência entre as estatísticas de teste. Os valores-p ajustados pelo método de *single-step* minP são dados por:

$$\tilde{p_i} = P\left(\min_{1\leq l\leq m} P_l \leq p_i | H_0^C\right) \tag{6.1}$$

onde H_0^c designa a hipótese nula completa e P_l a variável aleatória dos valores-p não ajustados da l-ésima hipótese.

O FWER é controlado por (6.1) quando é calculado sob a hipótese nula completa. Se a hipótese nula H_0^i é rejeitada quando $\tilde{p_i} \leq \alpha$, onde $\tilde{p_i}$ é obtido por (6.1) e então:

$$P\left(rejeitar \quad pelo \quad menos \quad um \quad H_0^i | H_0^c\right) = P\left(pelo \quad menos \quad um \quad \tilde{P_i} \leq \alpha | H_0^c\right)$$

onde $\tilde{P_i}$ é a variável aleatória que representa o valor-p ajustado.

Tendo em atenção a expressão (6.1), note-se que $\tilde{P_i} \leq \alpha$ se e só se $P_i \leq c_\alpha$, onde c_α denota o quantil da distribuição do valor-p mínimo,

$$c_\alpha = \max\left\{p \in S | P\left(\min_{1\leq j\leq m} P_j \leq p | H_0^c\right) \leq \alpha\right\}$$

com S é o espaço amostral dos valores-p mínimos. O quantil c_α é o maior valor observado da variável aleatória $\min_{1\leq j\leq m} P_j$, para o qual a probabilidade cumulativa é no máximo α.

Sob a condição de *subset pivotality*, verifica-se o controlo forte de FWER, consequência da desigualdade $\tilde{P}_i \leq \alpha$. Suponha-se que $K = \{i_1, ..., i_j\}$ é o conjunto de hipóteses H_0^i que são verdadeiras.

Note-se que: $P\left(Rejeitar \quad pelo \quad menos \quad um \quad H_0^i, i \in K \quad | \quad \bigcap_{i \in k} H_{0i}\right) =$

$$\begin{aligned}
&= P\left(\min_{i \in K} \tilde{P}_i \leq \alpha \quad | \quad \bigcap_{i \in k} H_{0i}\right) \\
&= P\left(\min_{i \in K} P_i \leq c_\alpha \quad | \quad \bigcap_{i \in k} H_{0i}\right) \\
&= P\left(\min_{i \in K} P_i \leq c_\alpha \quad | \quad H_0^c\right) \qquad (\textit{subset pivotality}\ ) \\
&\leq P\left(\min_{1 \leq i \leq k} P_i \leq c_\alpha | H_0^c\right) \\
&\leq \alpha
\end{aligned}$$

o que mostra que os valores-p ajustados *single-step* minP têm um controlo forte do FWER.

Os valores-p ajustados pelo método *step-down* minP de Westfall e Young (1993) [11] são definidos por:

$$\tilde{P}_{ri} = \max_{k=1,...,i} \left\{ P\left(\min_{l \in \{k,...,m\}} P_{rl} \leq p_{r_k} | H_0^c\right)\right\} \tag{6.2}$$

Seja $S = \{1, ..., k\}$ o conjunto de índices das hipóteses testadas H_0^i. Para um controlo forte do FWER é necessário que a probabilidade de rejeitar pelo menos um H_0^i não seja maior que α, independentemente do subconjunto de $K \subseteq S$ de hipóteses verdadeiras.

Denote-se por $K_0 = \{i_1, ..., i_j\}$ o conjunto de hipóteses H_0^i que são verdadeiras e suponha-se $K_0 \neq \emptyset$. (Se $K_0 = \emptyset$ não existem erros de tipo I). Seja x_K^α o α quantil de $\min_{t \in K} P_t | H_0^c$, então tem-se as seguintes relações:

$$\{V \geq 1\} = \{Rejeitar \quad pelo \quad menos \quad um \quad H_0^t, t \in K_0\} \subseteq \{\min_{t \in K_0} P_t \leq x_j^\alpha\}$$

onde $j \leq k - |K_0| + 1$ é definido por $\min_{t \in K_0} P_t = P_{r_j}$.

Então,

$$P\left(Existe \quad pelo \quad menos \quad um \quad erro \quad do \quad tipoI\right) =$$

$$\begin{aligned} &= P\left(Rejeitar \quad pelo \quad menos \quad umH_{0t}, t \in K_0 | \bigcap_{i \in K_0} H_0^i\right) \\ &\leq P\left(\min_{t \in K_0} P_t \leq x_j^\alpha | \bigcap_{i \in K_0} H_0^i\right) \\ &= P\left(\min_{t \in K_0} P_t \leq x_j^\alpha | H_0^c\right) \qquad (\textit{subset pivotality}\ ) \end{aligned}$$

Desde que $K_0 \in \overline{S_j}$, tem-se $x_j^\alpha \leq x_{K_0}^\alpha$ o que implica que

$$P\left(\min_{t \in K_0} P_t \leq x_j^\alpha | H_0^c\right) \leq P\left(\min_{t \in K_0} P_t \leq x_{K_0}^\alpha | H_0^c\right) \leq \alpha$$

o que demostra o controlo forte do FWER.

Note-se que, assumindo $P_i \sim U(0,1)$ e aplicando a desigualdade de Boole na quantidade 6.2 obtém-se os valores-p de Holm. Um procedimento baseado nos valores-p ajustados de *step-down* minP é, portanto, menos conservativo do que os procedimentos de Holm (Westfall e Young (1993) [11]).

Em situações que envolvem dependência entre os P_i, a distribuição de $\left(\min_{l \in \{k,\ldots,m\}} P_{rl} | H_0^c\right)$ é usualmente difícil de tratar. Por outro lado, nos casos em que os vetores $\mathbb{P}^*$ têm a mesma distribuição do vetor dos valores-p originais $\mathbb{P}$, podem ser facilmente simulados.

6.2 Procedimento maxT

Caso os valores-p ajustados (6.2) possam ser obtidos usando uma estatística de teste T_i então pode considerar-se procedimentos para valores-p ajustados pelo método *single-step* maxT:

$$\tilde{P}_i = P\left(\max_{1 \leq l \leq m} |T_l| \geq |t_i| | H_0^c\right) \tag{6.3}$$

O procedimento *single-step* maxT apresenta um controlo forte do FWER, assumindo *subset pivotality*.

Os valores ajustados pelo método *step-down* maxT são definidos por:

$$\tilde{p}_{r_i} = \max_{k=1,...,i} \left\{ P\left(\max_{l \in \{r_k,...,r_m\}} |T_l| \geq |t_{r_k}| | H_0^c \right) \right\} \tag{6.4}$$

onde $|t_{r_1}|, |t_{r_2}|, ..., |t_{r_m}|$ são as estatísticas de teste observadas e ordenadas.

Observações

Os procedimentos descritos têm alguns aspectos em comum (Ge, Dudoit e Speed (2003) [4]), tais como:

1. Se os valores-p não ajustados $P_1, ..., P_m$ são independentes e P_i tem uma distribuição $U[0,1]$, sob a hipótese H_i, então os valores-p ajustados *single-step* minP e os de Šidák coincidem.

2. Os procedimentos de cálculo dos valores-p ajustados maxT e minP fornecem um controlo fraco do FWER. O controlo forte do FWER verifica-se sob a suposição de *subset pivotality*, sem a qual, o ajuste da multiplicidade é mais complexo, pois é necessário considerar a distribuição das estatísticas de teste sob a hipótese nula parcial, ao invés da hipótese nula completa H_0^c.

3. Os valores-p ajustados maxT são mais fáceis de calcular do que os minP e ambos são iguais quando as estatísticas de teste T_i são identicamente distribuídas.

4. Quando os valores-p ajustados são estimados por permutação e um grande número de hipóteses são testadas, os procedimentos baseados nos valores-p minP tendem a ser mais sensíveis ao número de permutações e mais conservativos do que os baseado no maxT. Além disso,o procedimento minP requer mais cálculos do que o procedimento maxT , porque os valores-p não ajustados devem ser estimados antes de se considerar a distribuição dos seus mínimos sucessivos.

6.3 Descrição dos algoritmos

Em diversas situações, as distribuições marginal e conjunta das estatísticas de teste são desconhecidas. A reamostragem usando permutações ou *bootstrap* pode ser usada para estimar os valores-p brutos e ajustados evitando suposições sob a distribuição conjunta da estatística de teste.

A distribuição conjunta sob a hipótese nula da estatística de teste $T_1, ..., T_m$ pode ser estimada permutando as colunas da matriz de dados X, mantendo a sua estrutura de independência. Se o tamanho da amostra for elevado pode ser inviável considerar todas as permutações possíveis e como tal considera-se um subconjunto aleatório de B permutações (incluindo o observado).

Observe-se que, embora os métodos de permutação sejam apropriados para o tipos de hipóteses nulas consideradas neste trabalho, os procedimentos que recorrem a permutações não são indicados para outros tipos de hipóteses. Por exemplo, considere-se o caso simples de uma variável binária $Y \in \{1, 2\}$ e suponha-se que a hipótese nula H_0^i consiste nas distribuições condicionais de X_i dado $Y = 1$ e de X_i dado $Y = 2$ terem médias iguais, mas possivelmente variâncias diferentes. Sob a hipótese nula, impõe-se distribuições iguais nos dois grupos, o que é claramente mais forte do que simplesmente médias iguais. Como resultado, uma hipótese nula H_0^i pode ser rejeitada por outras razões que não a diferença entre as médias. Neste tipo de hipóteses, a reamostragem de bootstrap é mais apropriada pois preserva a estrutura de covariância presente nos dados originais. Para mais dados acerca deste assunto pode consultar-se Pollard e Van der Laan (2003) [8].

Apresenta-se, de seguida algoritmos de permutação para estimar valores-p ajustados e em anexo a sua implementação no software computacional *RStudio*.

Algoritmo de permutação para valores-p brutos

O algoritmo 1, seguindo (Ge, Dudoit e Speed(2003) [4]), descreve a forma de calcular os valores-p brutos usando permutações. Os valores-p ajustados dos procedimentos de Bonferroni, Sidák e Holm podem ser obtidos substituindo p_i por p_i^* nas equações 4.1, 4.3, 4.5 e 4.6.

Algoritmo 1 Permutações para os valores-p brutos

Para a permutação b , com b=1,...,B:

1. Permute as n colunas da matriz de dados X.
2. Calcule as estatísticas de teste $t_{1,b}, ..., t_{m,b}$ para cada hipótese.

Após a realização das B permutações, para um teste de hipóteses bilateral, o valor-p das permutações para a hipótese H_i é dado por:

$$p_i^* = \frac{\#\{b : |t_{i,b}| \geq |t_i|\}}{B} \quad , para \quad i = 1, ..., m$$

Nos testes de permutação, compara-se o valor da estatística observada com as saídas apresentadas seguindo o algoritmo computacional, que consiste em criar amostras através de permutações dos dados originais (maiores detalhes podem ser vistos, por exemplo, em Duczmal et al., 2003). Se estas estatísticas estimadas têm valores semelhantes à estatística observada nos dados, então não haverá justificação para rejeitar H_0, ou seja, os desvios da estatística observada podem ser atribuídos ao acaso.

Algoritmo de permutação para valores-p ajustados, pelo método *step-down* maxT

Para os valores-p ajustados pelo método maxT de Westfall and Young (1993) ([11]), sob a hipótese nula dos máximos sucessivos $\max_{l=i,...,m}$ da estatística de teste tem de ser estimada. O caso *single-step* é mais simples e para o implementar apenas é necessário a distribuição de $\max_{l=1,...,m} |T_{s_l}|$. O algoritmo computacional, retirado de Ge, Dudoit e Speed(2003) [4] apresentado de seguida é baseado no algoritmo 4.1 das páginas 116-117 de Westfall and Young (1993) [11].

Algoritmo 2: Algoritmo de permutação para os valores-p ajustados *step-down* maxT

Para os dados originais, ordenar de forma decrescente as estatísticas de teste observadas:

$$|t_{s_1}| \geq |t_{s_2}| \geq ... \geq |t_{s_m}|$$

Para a permutação b , com b=1,...,B:

1. Permute as n colunas da matriz de dados X.
2. Calcule as estatísticas de teste $t_{1,b}, ..., t_{m,b}$ para cada hipótese.
3. Calcula $u_{i,b} = \max_{l=i,...,m} t_{s_l,b}$ (ver 6.4), os máximos sucessivos das estatísticas de teste:

$$u_{m,b} = |t_{s_m,b}|$$

$$u_{i,b} = \max(u_{i+1,b}, |t_{s_i,b}|) \quad , para \quad i = m-1, ..., 1$$

Os passos anteriores são repetidos B vezes e os valores-p ajustados são estimados por:

$$\tilde{p}^*_{s_i} = \frac{\#\{b : |u_{i,b}| \geq |t_{s_i}|\}}{B} \quad , para \quad i = 1, ..., m$$

com as restrições de monotocidade impostas através de:

$\tilde{p}^*_{s_1} \leftarrow \tilde{p}^*_{s_1}$ e $\tilde{p}^*_{s_i} \leftarrow \max\left(\tilde{p}^*_{s_{i-1}}, \tilde{p}^*_{s_i}\right)$ para i=2,...,m.

Algoritmo tradicional de permutação dupla, para valores-p ajustados pelo método *step down* minP

Os valores-p ajustados *step down* e *single step* minP de Westfall e Young (1993) ([11]) são em geral difíceis de calcular computacionalmente sob a hipótese nula de $P_1, P_2, ..., P_m$. O algoritmo computacional, de dupla permutação tradicional retirado de Ge, Dudoit e Speed(2003) [4] que se apresenta de seguida é baseado no algoritmo 2.8 das paginas 66-67 de Westfall and Young (1993) [11].

Quando os valores-p brutos são desconhecidos, a reamostragem adicional da etapa 2 para estimar esses valores-p pode ser computacionalmente inviável. O algoritmo 3 é chamado de algoritmo de dupla permutação porque efetua dois procedimentos de reamostragem.

Para ultrapassar o problema computacional pode recorrer-se a procedimentos com base em valores-p ajustados maxT, que podem ser estimados a partir de uma única permutação usando o algoritmo 2. No entanto, se as estatísticas de teste não forem identicamente distribuídas sob as hipóteses consideradas, os valores-p ajustados maxT podem ser diferentes dos valores-p ajustados minP e podem dar pesos diferentes para hipóteses diferentes. Por exemplo, se a estatística de teste T_i para uma hipótese particular H_0^i tem uma distribuição de cauda pesada, tenderá tenderá a apresentar valores superiores do que as restantes estatísticas de teste e, portanto, H_0^i tenderá a ter um valor-p ajustado menor do que as restantes hipóteses. Nesses casos é melhor calcular o minP em vez dos valores-p ajustados maxT.

Algoritmo 3: Algoritmo de dupla permutação para os valores-p ajustados *step-down* minP

Para os dados originais, usar o algoritmo 1 para calcular os valores-p brutos $p_1^*, p_2^*, ..., p_m^*$ e ordená-los de forma crescente:

$$p_{r_1}^* \leq p_{r_2}^* \leq ... \leq p_{r_m}^*$$

Para a permutação b , com b=1,...,B:

1. Permute as n colunas da matriz de dados X.
2. Calcule os valores-p brutos $p_{1,b}, ..., p_{m,b}$ para cada hipótese hipótese, usando os dados permutados.
3. Calcule $q_{i,b} = \min_{l=i,...,m} p_{r_l,b}$ (ver 6.2), os mínimos sucessivos dos valores-p brutos..

$$q_{m,b} = p_{r_m,b}$$

$$q_{i,b} = \min\left(q_{i+1,b}, p_{r_i,b}\right) \quad , para \quad i = m-1, ..., 1$$

Os passos anteriores são repetidos B vezes e os valores-p ajustados são estimados por:

$$\tilde{p}_{r_i}^* = \frac{\#\{b : q_{i,b} \leq p_{r_i}^*\}}{B} \quad , para \quad i = 1, ..., m$$

com as restrições de monotocidade impostas através de:

$\tilde{p}_{r_1}^* \leftarrow \tilde{p}_{r_1}^*$ e $\tilde{p}_{r_i}^* \leftarrow \max\left(\tilde{p}_{r_{i-1}}^*, \tilde{p}_{r_i}^*\right)$ para i=2,...,m.

6.4 Exemplo Ilustrativo

Considere-se uma simulação no software *RStudio* em que se pretende testar 20 hipóteses nulas H_0^i, com $i = 1, ...20$, numa amostra de 7 observações. Estes dados podem ser apresentados numa matriz com 20 linhas e 7 colunas. As primeiras quatro colunas são uma amostra proveniente de uma população com distribuição normal $N(0, 1)$ e as restantes três colunas uma amostra de uma população com distribuição normal $N(3, 1.5)$, tal como mostra o *boxplot* seguinte.

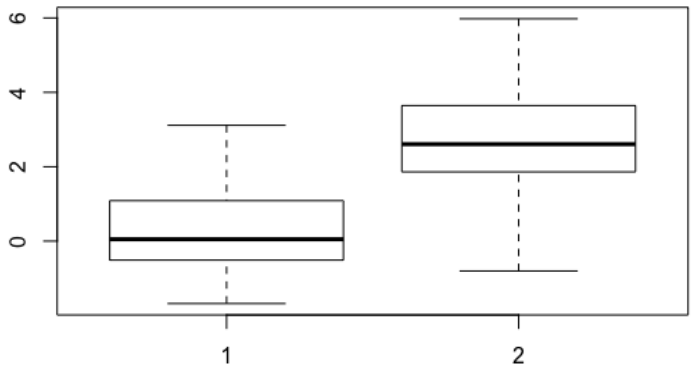


Figura 6.1: Histograma das distribuições dos dados

Nestes testes de hipóteses múltiplos pretende-se testar se as médias de ambas as amostras são iguais, com $\alpha = 0.05$. Para tal, aplicam-se os algoritmos 1,2 e 3 e analisam-se as diferenças entres eles, considerando $\binom{7}{4} = 35$ permutações das colunas.

Os valores p brutos usando a permutação obtidos pelo algoritmo 1 são:

0.02857143	0.14285714	0.17142857	0.28571429	0.05714286
0.02857143	0.25714286	0.02857143	0.05714286	0.34285714
0.02857143	0.08571429	0.02857143	0.02857143	0.11428571
0.02857143	0.14285714	0.20000000	0.02857143	0.08571429

Como apenas 8 valores p brutos são menores que 0.05 então rejeitam-se 8 hipóteses nulas.

Nas mesmas condições anteriormente referidas, aplica-se o algoritmo 2 em que são calculados os valores-p ajustados, por permutação. Os resultados obtidos foram os seguintes:

0.17142857	0.08571429	0.08571429	0.08571429	0.08571429
0.08571429	0.08571429	0.08571429	0.08571429	0.08571429
0.08571429	0.08571429	0.08571429	0.08571429	0.08571429
0.05714286	0.05714286	0.05714286	0.02857143	0.02857143

Tal como para o algoritmo 1, rejeitam-se 2 hipóteses nulas.

Através do algoritmo 3 obtém-se os valores-p ajustados seguintes:

0.08571429	0.08571429	0.02857143	0.00000000	0.00000000
0.00000000	0.00000000	0.00000000	0.05714286	0.05714286
0.17142857	0.17142857	0.20000000	0.22857143	0.22857143
0.31428571	0.37142857	0.51428571	0.37142857	0.31428571

Tendo em atenção os valores-p ajustados anteriores, usando o algoritmo 3 são rejeitadas 6 hipóteses nulas.

Capítulo 7

Procedimentos de controlo do FDR

Nesta secção apresenta-se um ponto de vista diferente sobre a forma de contornar o problema da multiplicidade. Em diversos problemas de multiplicidade deve ter-se em conta o número de rejeições erradas em vez de apenas a questão de se ter cometido um erro.No entanto, ao mesmo tempo, a gravidade do prejuízo sofrido por rejeições erradas está inversamente relacionada com o número de hipóteses rejeitadas.

Benjamini e Hochberg ([1]) propuseram uma abordagem diferente para testes múltiplos, afirmando que em muitas situações, o controlo do FWER pode adotar procedimentos excessivamente conservativos e é preparado para tolerar alguns erros do tipo I,desde que esse número seja pequeno em comparação com o número de hipóteses rejeitadas. Essas considerações levam a uma abordagem menos conservativa, que exige controlar o valor esperado da proporção de erros do tipo I de entre as hipóteses rejeitadas (FDR).

FDR com hipóteses nulas independentes

Benjamini e Hochberg (1995) ([1]) apresentaram um procedimento *step-up* para um controlo forte do FDR para estatísticas de teste independentes, embora a suposição da independência sob a hipótese alternativa não seja necessária. Considere-se o FDR definido como $FDR = E\left(\frac{V}{R}\mathbb{1}_{\{R>0\}}\right)$.

Sejam H_0^1, H_0^2,..., H_0^m as hipóteses a testar e p_1, p_2, ..., p_m os correspondentes valores-p brutos. Usando a mesma notação da secção 2 considere os valores-p brutos ordenados, p_{r_1}, p_{r_2}, ..., p_{r_m} e denote-se H_0^i a hipótese nula correspondente a p_{r_i}. O procedimento do tipo Bonferroni para testes múltiplos para controlar o FDR a um nível q^* define-se do seguinte modo:

Seja *k* o maior *i* para o qual

$$p_{r_i} \leq \frac{i}{m}q^* \tag{7.1}$$

então rejeita-se todos os H_0^i, i=1,...,k.

Note que se não existir k então não se rejeita nenhuma hipótese.

Tomando a mesma linha dos valores-p ajustados FWER e aplicando a definição introduzida por Benjamini e Hochberg (1995) [1] podem reformular-se os valores-p ajustados para o FDR como:

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m}\left\{\min\left(\frac{m}{k}p_{r_k}, 1\right)\right\} \tag{7.2}$$

Teorema 7.1. Para estatísticas de teste independentes e para qualquer configuração de hipóteses falsas, o procedimento anterior controla o FDR, a um nível q^*.

PROVA. O teorema segue do lema seguinte, cuja prova pode ser consultada no apêndice do artigo Benjamini e Hochberg (1995) [1]. □

Lema 7.2. Para quaisquer, $0 \leq m_0 \leq m$, valores-p independentes correspondentes às hipóteses nulas verdadeiras e para quaisquer $m_1 = m - m_0$ valores-p correspondentes às hipóteses nulas falsas, o procedimento de hipóteses múltiplos definido por (7.1) satisfaz

a seguinte desigualdade:

$$E\left(Q|P_{m_0+1}=p_1,...,P_m=p_{m_1}\right) \leq \frac{m_0}{m}q^* \tag{7.3}$$

Suponha-se que $m_1 = m - m_0$ são as hipóteses nulas falsas. Portanto a distribuição conjunta de $p_1^{''}$, ...,$p_{m_1}^{''}$ que corresponde a essas hipóteses nulas falsas, integrando a desigualdade (7.3) é:

$$E(Q) \leq \frac{m_0}{m}q^* \leq q^*$$

e o FDR é controlado.

Observação 7.3. Note-se que a independência das estatísticas de teste correspondentes às hipóteses nulas falsas não é necessário para a prova do teorema.

Este procedimento foi citado por Simes (1986) ([10]) como uma extensão exploratória do seu procedimento para a rejeição da hipótese nula completa, se para algum i, $p_{r_i} \leq \frac{i}{m}\alpha$. Enquanto que, Simes (1986) mostrou que esse procedimento controla o FWER sob a hipótese nula completa, Hommel(1988) mostrou que o procedimento estendido para inferências sobre hipóteses individuais não tem um controlo forte do FWER: para algumas configurações de hipóteses nulas falsas, a probabilidade de uma rejeição errada é maior do que α.

Hochberg(1988) ([6]) sugeriu uma forma diferente para utilizar o procedimento de Simes de modo a ter um controlo forte do FWER. Esse procedimento é o seguinte:

Seja *k* o maior *i* para o qual

$$p_{r_i} \leq \frac{i}{m+1-i}\alpha \tag{7.4}$$

então rejeita-se todos os H_0^i, i=1,...,k.

Observe-se a relação entre os procedimentos de Hochberg's e do controlo do FDR quando q^* é igual a α. Ambos são procedimentos *step-down*, que começam por comparar p_{r_m} com α e, caso seja menor, todas as hipóteses nulas são rejeitadas, como se fosse tomada uma abordagem PCER. Se $p_{r_m} > \alpha$, prossegue-se para valores-p menores até

que um satisfaça a condição. Caso os procedimentos não terminem antes, terminam comparando p_{r_1} com $\frac{\alpha}{m}$, como numa comparação de Bonferroni pura.

Apesar das semelhanças nos procedimentos, enquanto que neste procedimento a sequência de p_i's é comparada com $\{1 - \frac{(i-1)}{m}\}\alpha$, no procedimento de Hochberg é comparada com $\{\frac{1}{m+1-i}\}\alpha$. A série de constantes linearmente decrescentes do método de controlo do FDR é sempre maior do que as constantes hiperbolicamente decrescentes de Hochberg e a razão extrema é tão grande quanto $\frac{4m}{(m+1)^2}$ com $i = \frac{m+1}{2}$. Isto mostra que o procedimento sugerido rejeita da mesma forma, pelo menos tantas hipóteses como o método de Hochberg e portanto tem maior potência do que outros métodos de controlo do FWER, tal como em Holm's(1979) ([7]).

Veja-se um exemplo, retirado de Benjamini e Hochberg (1995) [1].

Exemplo 7.4. Considere-se um estudo, com 421 pessoas que sofreram enfarto do miocárdio. Pretende-se saber qual o efeito de dois medicamentos na prevenção da mortalidade. O estudo apresenta quatro famílias de hipóteses, das quais apenas se considera a que contém 14 comparações, entre elas, uma sobre a mortalidade. Os valores p_{r_i} ordenados para as 15 comparações são dados por:

0.0001	0.0004	0.0019	0.0095	0.0201	0.0278	0.0298	0.0344
0.0459	0.3240	0.4262	0.5719	0.6528	0.7590	1.000	

Com recurso à abordagem de Bonferroni para o controlo do FWER, usando $\frac{0.05}{15} = 0.0033$ rejeitam-se três hipóteses correspondentes aos valores-p mais pequenos. Estas hipóteses não incluem a comparação de mortalidade. Usando o procedimento de Hochberg's rejeitam-se as mesmas três hipóteses. Assim, a afirmação sobre uma redução significativa na mortalidade é injustificada do ponto de vista clássico.

Usando o procedimento de Benjamimi e Hochberg para controlo do FDR, com $q^* = 0.05$ e comparando sequencialmente cada p_{r_i} com $\frac{0.05i}{15}$. O primeiro valor-p que satisfaz essa limitação é p_{r_4}:

$$p_{r_4} = 0.0095 \leq \frac{4}{15}0.05 = 0.013$$

Assim, rejeitam-se quatro hipóteses cujos valores-p são menores ou iguais a 0.013. Portanto com uma confiança apropriada, pode-se afirmar que há uma redução da mortalidade, da qual não tínhamos evidências suficientemente fortes anteriormente.

O procedimento descrito acima pode ser visto como um procedimento de maximização *post hoc*, como sugere o teorema apresentado em Benjamini e Hochberg (1995) [1].

Teorema 7.5. O procedimento de controlo do FDR referido em (7.1) é a solução do seguinte problema de maximização com restrição:

Escolher α que maximize o número de rejeições nesse nível, $r(\alpha)$, sujeito à restrição $\frac{\alpha m}{r(\alpha)} \leq q^*$.

A prova pode ser consulta em Benjamini e Hochberg (1995) [1].

Capítulo 8

Conclusão

A pergunta que todos os investigadores fazem quando pretendem realizar um estudo de hipóteses múltiplas é qual o melhor método. Nesta área não há respostas absolutas. Há vários métodos diferentes, com diferentes especificidades, e uns funcionarão melhor num contexto do que noutro, mas no entanto não se sabe como porque não se sabe quantas hipóteses nulas H_0 são verdadeiras. Apesar de todos os estudos de simulação já realizados não há orientações conhecidas para quando utilizar um ou outro.

Os procedimentos baseados em reamostragem poderão ser úteis para tamanhos amostrais pequenos, por exemplo, em contraposição com os paramétricos, que assumem normalidade. Estes métodos são os mais recentes e exigem um poder computacional maior do que os paramétricos, o que não é essencialmente um problema, hoje em dia.

A abordagem mais utilizada continua a ser a do controlo do FWER, embora comece cada vez mais a ganhar forma uma outra perspectiva que é controlo do FDR.

O problema de estimação e controlo das taxas de erro em testes de hipóteses múltiplos é um problema atual e que ganha cada vez mais importância, na medida em que a sua utilização é bastante relevante em estudos nomeadamente na área da genonica.

Anexos

Algoritmo de permutação para valores-p brutos

```
library(gtools)
set.seed(102)
xa <- matrix(data = rnorm(80, mean = 0, sd = 1), nrow = 20, ncol = 4)
xb <- matrix(data = rnorm(60, mean = 3, sd = 1.5), nrow = 20, ncol = 3)
x <- cbind(xa, xb)
m<-nrow(x)
ncomb <-choose(7,4)
valorp <- as.numeric()
est <- matrix(1000,nrow=m,ncol=ncomb)
for (i in 1:m){
  #Calcular a estatistica observada original da linha i
  obs[i] <- as.numeric(t.test(x[i,1:4],x[i,5:7])$statistic)
  #Selecionar a linha i da matriz x
  x1 <- x[i,]
  #Matriz de todas as combinacoes 7 - 4  da linha i
  p <- combinations(n=7,r=4,v=x1,repeats.allowed=F)
  #Para cada combinacao b, calcular a estatistica de teste da linha i
  for (b in 1:ncomb){
    #Selecionar cada linha i da matriz x e calcular a estatistica de teste
    est[i,b] <- as.numeric(t.test(p[b,1:4],x1[!x1 %in% p[b,]])$statistic)
  }# end b
}# end i
for (i in 1:m) {
  #Calculo do valor p de cada linha i
  valorp[i] <- sum(abs(est[i,])>=abs(obs[i]))/(ncomb)
}
```

Listing 8.1: Algoritmo de permutação para valores-p brutos

Algoritmo de permutação para valores-p ajustados *step-down* maxT

```
library(gtools)
set.seed(102)
xa <- matrix(data = rnorm(80, mean = 0, sd = 1), nrow = 20, ncol = 4)
xb <- matrix(data = rnorm(60, mean = 3, sd = 1.5), nrow = 20, ncol = 3)
x <- cbind(xa, xb)
#Ordenar de forma decrescente as estatisticas de teste observadas (ts1,...,
    tsm)
m <- nrow(x)
ncomb <-choose(7,4)
#Calcular a estatistica observada para cada linha i
for (i in 1:m){
  est.obs[i] <- as.numeric(t.test(x[i,1:4],x[i,5:7])$statistic)
}
#Ordenar o modulo da estatistica observada da linha i
est.obs.ord <- sort(abs(est.obs),decreasing=TRUE)
#Para cada permutacao, calcular as estatisticas de teste (t1b,...,tm,b)
valorps <- as.numeric()
est <- matrix(1000,nrow=m,ncol=ncomb)
u<- matrix(0,ncol=ncomb,nrow=m)
#Para cada permutacao b, calcular o maximo sucessivo das estatisticas de
    teste (u)
for (b in 1:ncomb) {
  for (i in 1:m) {
    #Para cada linha i da matriz x
    x1 <- x[i,]
    #Matriz de todas as combinacoes 7 - 4  da linha i
    p <- combinations(n=7,r=4,v=x1,repeats.allowed=F)
    #Estatistica de teste da linha i para cada permutacao b
    est[i,b] <-as.numeric(t.test(p[b,1:4],x1[!x1 %in% p[b,]])$statistic)
  }#end i
  #Calcular o maximo sucessivo das estatisticas de teste, para cada
      permutacao b
  u[,b] <- sort(abs(est[,b]),decreasing=TRUE)
}# end b

```

```
  #Calculo dos valores p ajustados para cada i
  for (i in 1:m) {
    valorps[i] <- sum(u[i,]>=abs(est.obs.ord))/ncomb
  }
```

Listing 8.2: Algoritmo de permutação para valores-p ajustados *step-down* maxT

Algoritmo de permutação para valores-p ajustados *step-down* minP

```
library(gtools)
set.seed(102)
xa <- matrix(data = rnorm(80, mean = 0, sd = 1), nrow = 20, ncol = 4)
xb <- matrix(data = rnorm(60, mean = 3, sd = 1.5), nrow = 20, ncol = 3)
x <- cbind(xa, xb)

m <- nrow(x)
ncomb <-choose(7,4)
valorp <- as.numeric()
obs <- as.numeric()
est <- matrix(1000,nrow=m, ncol=ncomb)
#Algoritmo da box 1 para calcular os valores p brutos
for (i in 1:m){
  #Calcular a estatistica observada original da linha i
  obs[i] <- as.numeric(t.test(x[i,1:4],x[i,5:7])$statistic)
  #Selecionar a linha i da matriz x
  x1 <- x[i,]
  #Matriz de todas as combinacoes 7 - 4  da linha i
  p <- combinations(n=7,r=4,v=x1,repeats.allowed=F)
  #Para cada combinacao b, calcular a estatistica de teste da linha i
  for (b in 1:ncomb){
    #Selecionar cada linha i da matriz x e calcular a estatistica de teste
    est[i,b] <- as.numeric(t.test(p[b,1:4],x1[!x1 %in% p[b,]])$statistic)
  }# end b
  valorp[i] <- sum(abs(est[i,])>=abs(obs[i]))/(ncomb)
}# end i

```

```
#Ordenar os valores p da linha i, por ordem crescente
valorp.ord <- sort(valorp)

#Para cada permutacao, calcular os valores p brutos (p1b,...,pm,b)
valorpr <- as.numeric()
ncomb <-choose(ncol(x),4)
valorp.aux <- matrix(2,ncol = ncomb,nrow=m)
est2 <- matrix(1000,nrow=m,ncol=ncomb)
est.obs.aux <- as.numeric()
q <- matrix(2,ncol = ncomb,nrow=m)
x.novo <- matrix(1000, nrow=nrow(x), ncol=ncol(x))
p74 <-combinations(n=7,r=4,repeats.allowed=F)
for (b in 1:ncomb) {
  #Passo1 box3: implementar a permutacao b
  # nas colunas do x
  for (i in 1:nrow(x)){
    xi <- x[i,]
    xi4b <- x[i,p74[b,]]
    xi3b <- xi[!xi %in% xi4b]
    x.novo[i,] <- c(xi4b,xi3b)
  }
  # para a permutacao b calcular os valores p para cada linha do x novo
  for (i in 1:m) {
    x1 <- x.novo[i,] # Para a linha i da x novo
    # estatistica observada da linha i
    est.obs.aux[i] <- t.test(x1[1:4],x1[4:6])$statistic
    #Matriz de todas as combinacoes 6 - 3  da linha i
    px1 <- combinations(n=7,r=4,v=x1,repeats.allowed=F)
    #para cada permutacao bb da linha i calcular estatistica de teste
    for (bb in 1:ncomb){
      est2[i,bb] <-as.numeric(t.test(px1[bb,1:4],x1[!x1 %in% px1[bb,]])$
          statistic)
    } #end bb
    #calculo do valor p da linha i
    valorp.aux[i,b] <- sum(abs(est2[i,])>=abs(est.obs.aux[i]))/(ncomb)
    #Ordenacao dos valores valorp.ord(=q)
```

```
63    }#end i
64    for (i in 1:m){
65      q[i,b] <- sort(valorp.aux[i:m,b])[1]
66    } # end i
67  } # end b
68
69  #Calculo dos valores p ajustados
70  for (i in 1:m) {
71    valorpr[i] <- sum(q[i,]<=valorp.ord[i])/ncomb
72  }
```

Listing 8.3: Algoritmo de permutação dupla, para valores-p ajustados *step down* minP

Bibliografia

[1] Yoav Benjamini e Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. Em: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[2] George Casella e Roger L Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002.

[3] Sandrine Dudoit, Juliet Popper Shaffer, Jennifer C Boldrick et al. “Multiple hypothesis testing in microarray experiments”. Em: *Statistical Science* 18.1 (2003), pp. 71–103.

[4] Youngchao Ge, Sandrine Dudoit e Terence P Speed. “Resampling-based multiple testing for microarray data analysis”. Em: *Test* 12.1 (2003), pp. 1–77.

[5] Peter Hall e Susan R Wilson. “Two guidelines for bootstrap hypothesis testing”. Em: *Biometrics* (1991), pp. 757–762.

[6] Yosef Hochberg. “A sharper Bonferroni procedure for multiple tests of significance”. Em: *Biometrika* 75.4 (1988), pp. 800–802.

[7] Sture Holm. “A simple sequentially rejective multiple test procedure”. Em: *Scandinavian journal of statistics* (1979), pp. 65–70.

[8] Katherine S Pollard e Mark J van der Laan. “Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data”. Em: (bepress (2003).

[9] Juliet Popper Shaffer. “Multiple hypothesis testing”. Em: *Annual review of psychology* 46.1 (1995), pp. 561–584.

[10] R John Simes. “An improved Bonferroni procedure for multiple tests of significance”. Em: *Biometrika* 73.3 (1986), pp. 751–754.

[11] Peter H Westfall, S Stanley Young et al. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. John Wiley & Sons, 1993.

[12] Daniel Yekutieli e Yoav Benjamini. “Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics”. Em: *Journal of Statistical Planning and Inference* 82.1-2 (1999), pp. 171–196.