U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Controlo da taxa de falsas descobertas em testes de hipóteses múltiplas – um estudo de simulação

Rita Marília Duarte Nogueira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Engenharia Matemática

2021

MSc

2.º
CICLO

FCUP
2021

U.PORTO

Controlo da taxa de falsas descobertas em testes de hipóteses múltiplas – um estudo de simulação

Rita Marília Duarte Nogueira

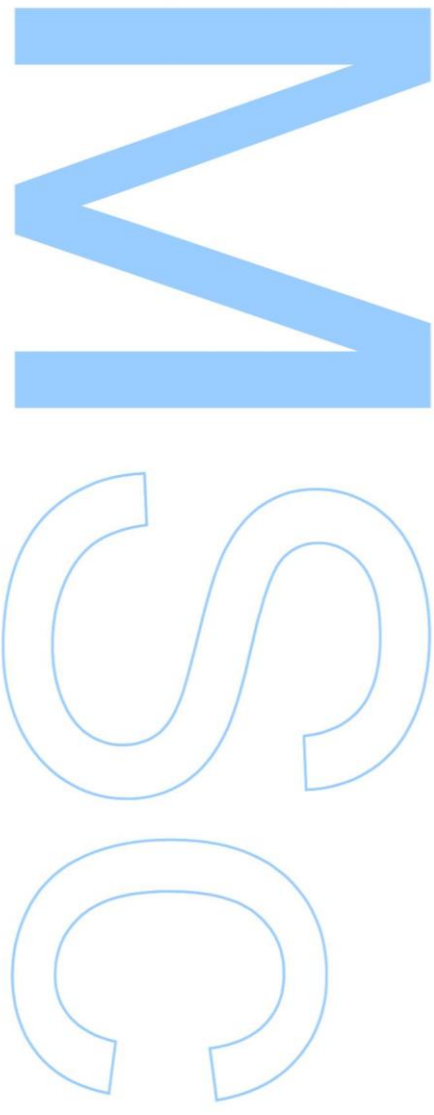
FC

Controlo da taxa de falsas descobertas em testes de hipóteses múltiplas – um estudo de simulação

Rita Marília Duarte Nogueira
Mestrado em Engenharia Matemática
Departamento de Matemática
2021

Orientador
Margarida Maria Araújo Brito, Professora Associada,
Faculdade de Ciências da Universidade do Porto

Coorientador
Ana Rita Pires Gaio, Professora Auxiliar,
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, _____/_____/_____

" We never fail when we try to do our duty, we always fail when we neglect to do it. "

Robert Baden-Powell

Agradecimentos

Esta dissertação é fruto de muito trabalho e dedicação, não sendo possível a sua realização sem o apoio e incentivo de muitas pessoas. Por isso, gostaria de agradecer-lhes.

Às minhas orientadoras, professora Margarida Brito e Rita Gaio, pela paciência e confiança que foram determinantes para o resultado final alcançado.

Aos meus amigos, pelos 434 dias de aventuras, tardes de trabalho e noitadas de jogos. Agradeço pela presença constante e por terem tornado esta viagem da vida tão bela!

Aos meus pais, Lurdes e Jorge, agradeço por estarem sempre presentes e por tudo o que fizeram por mim.

Por último, um agradecimento especial ao Tiago, pelo apoio incondicional, pelas palavras e piadas (MUITO secas) nos momentos certos e por todo o carinho dado.

Abstract

Master in Mathematical Engineering

False discovery rate control in multiple hypothesis tests - a simulation study

by Rita NOGUEIRA

A multiple hypothesis test consists of testing several (often thousands) null hypotheses simultaneously. This multiple inference has an associated statistical problem, because as the number of hypotheses increases, the global error increases and so does the probability of false rejections. There are several methods proposed to deal with this problem, but no acceptable solution for all situations.

In this dissertation, this problem is approached, along with some procedures to overcome it, such as the classical methods of Bonferroni and Holm, among others, with special focus on Benjamini and Hochberg's method. The latter will be compared with the method most used by statisticians - Bonferroni - in terms of errors and power. Furthermore, a Bayesian strategy is also introduced and studied.

The simulation study carried out shows that the Benjamini and Hochberg method is situated between the Bonferroni and the unadjusted procedures, in terms of the method's power and the number of errors made. Regarding the Bayesian approach, it is observed that this is influenced by the sample size and not by the number of tests performed.

The methods studied in the above simulation will be applied to a real database in the field of Genomics.

Resumo

Mestrado em Engenharia Matemática

Controlo da taxa de falsas descobertas em testes de hipóteses múltiplas – um estudo de simulação

por Rita NOGUEIRA

Um teste de hipóteses múltiplas consiste em testar várias (muitas vezes, milhares) hipóteses nulas em simultâneo. No entanto, esta inferência múltipla tem um problema estatístico associado, pois à medida que se aumenta o número de hipóteses a testar, o erro global aumenta bem como a probabilidade de haver falsas rejeições. Vários são os métodos que têm sido propostos para lidar com este problema, mas nenhuma solução será aceitável para todas as situações.

Nesta dissertação, este problema é abordado, juntamente com alguns procedimentos para o contornar, tais como os métodos clássicos de Bonferroni e de Holm, entre outros, incidindo em especial no método de Benjamini e Hochberg. Este último irá ser comparado com o método mais usado pelos estatísticos – o de Bonferroni - e pelo método não ajustado em termos de erros cometidos e potência. Para além disso, é ainda descrita uma abordagem bayesiana.

O estudo de simulação efetuado mostra que o método de Benjamini e Hochberg situa-se entre os procedimentos de Bonferroni e não ajustado, no que toca à potência do método e ao número de erros cometidos. Relativamente à abordagem bayesiana, observa-se que esta é influenciada pelo tamanho amostral e não pelo número de testes realizados.

Os métodos estudados na simulação referida serão aplicados a uma base de dados real da área da Genómica.

Conteúdo

Lista de Figuras

Lista de Tabelas

Glossário

BF	Bayes Factor
BH	Benjamini-Hochberg
FDR	False Discovery Rate
FWER	Family-Wise Error Rate

Capítulo 1

Introdução

Na década de 1920, Jerzy Neyman e Egon Pearson [1] deram um enorme contributo para a Estatística com o desenvolvimento da teoria dos testes de hipóteses. Esta teoria, juntamente com a teoria do valor-p desenvolvida por Ronald Fisher [2] na mesma época, tem estado constantemente presente na comunidade científica, uma vez que ambas permitem abordar uma certa hipótese, como a superioridade de um tratamento sobre outro ou a associação entre uma característica e um resultado, e defini-la, com um certo grau de certeza, como correta ou errada. A introdução destas teorias no raciocínio científico forneceu importantes ferramentas quantitativas para os investigadores, na medida em que permitem planear estudos, relatar descobertas, comparar resultados e até mesmo tomar decisões.

Em estatística, os testes de hipóteses múltiplas ocorrem quando se considera um conjunto de inferências estatísticas simultaneamente. Uma vez que cada inferência feita está sujeita a erros, visto que se está a tirar informação de uma população utilizando apenas dados de uma (pequena) amostra desta, quanto mais inferências são feitas, mais prováveis se tornam as inferências erradas. Este problema recebeu atenção crescente na década de 1950 com o trabalho de estatísticos como Tukey e Scheffé. Nas décadas seguintes, várias técnicas estatísticas foram desenvolvidas para resolver esse problema, normalmente exigindo um limite de significância mais rígido para comparações individuais, de modo a compensar o número de inferências feitas [3, 4].

Uma possível abordagem a este problema usa a chamada FWER (Family-Wise Error Rate) - probabilidade de cometer pelo menos uma falsa rejeição - e a mais recente abordagem usa a FDR (False Discovery Rate) - proporção esperada de falsas rejeições. O método

de Bonferroni [5] é o mais conhecido para o controlo da FWER. No caso da FDR, Benjamini e Hochberg [6] desenvolveram, em 1995, um método que permite identificar as hipóteses a serem rejeitadas mantendo a FDR abaixo de um certo limiar.

Nesta dissertação, estudam-se as abordagens mais relevantes dos últimos tempos no controlo do erro em testes de hipóteses múltiplas. A dissertação é apresentada da seguinte maneira:

No capítulo II apresenta-se uma breve introdução aos testes de hipóteses simples, expondo-se os seus princípios básicos em abordagens frequencistas e bayesianas.

No capítulo III a introdução feita no capítulo anterior é aprofundada para testes de hipóteses múltiplas, com a apresentação do problema inerente a este tipo de testes e de algumas soluções clássicas para o mesmo. Aborda-se também o fator de Bayes nesse contexto.

No capítulo IV é apresentada uma nova taxa, cujo objetivo é solucionar o problema supra-referido, juntamente com um método, designado por método de Benjamini e Hochberg, para a controlar. Posto isto, uma razão pela qual este método tem muito sucesso será exposta.

No capítulo V é apresentado um estudo de simulação que compara o método apresentado no capítulo anterior com o estudo bruto - sem ajuste do valor-p - e com o método clássico de Bonferroni no que toca à quantidade de erros de tipo I e II num teste de hipóteses múltiplos. Este estudo será feito em função do número de hipóteses total, do número de hipóteses verdadeiras e do tamanho de efeito. Para além disso, também é feita uma análise ao comportamento do valor-p bruto à medida que se aumenta o tamanho do efeito. Após isto, é aplicado o fator de Bayes comparando-o com o método bruto e com o método de BH. Por fim, irá ser feita uma aplicação dos métodos bruto e de BH, e ainda o fator de Bayes a um conjunto de dados real.

No capítulo VII são apresentadas as conclusões deste estudo, assim como algumas considerações sobre trabalho futuro.

Capítulo 2

Testes de Hipóteses Simples

A Inferência Estatística permite estudar o comportamento de populações que, normalmente, são muito grandes ou inacessíveis, a partir de amostras representativas. Os métodos de inferência permitem avaliar o quão perto uma estatística (obtida de amostras) está do parâmetro (que existe na população). Pode-se tomar decisões e fazer previsões sobre as populações, mesmo tendo dados relativos a poucos indivíduos dessa população [7].

Uma das metodologias basilares da Inferência são os testes de hipóteses, que usam probabilidades para quantificar o quão plausível é um determinado valor do parâmetro enquanto controlam a probabilidade de uma inferência incorreta. O principal objetivo é verificar se os dados amostrais apoiam certas afirmações sobre uma população. Essas afirmações são designadas de hipóteses.

Este processo de testagem necessita de duas hipóteses, uma definida como hipótese nula e a outra como hipótese alternativa. A hipótese nula é denotada por H_0 e a hipótese alternativa por H_1. Geralmente, a hipótese nula assume um valor específico para um determinado parâmetro, enquanto que a hipótese alternativa afirma que o parâmetro cai nalgum intervalo de valores alternativo. Existem outros testes que permitem, por exemplo, estudar a distribuição da população, testes estes que não serão considerados por não fazerem parte do âmbito desta dissertação.

Geralmente, um teste de hipóteses é usado para avaliar a diferença entre uma estimativa (amostral) e o respetivo parâmetro (populacional). Uma estatística de teste descreve o quão longe essa estimativa amostral está do valor do parâmetro referido na hipótese nula. Para interpretar o valor estatístico do teste, usa-se um resumo da probabilidade da evidência contra a hipótese nula. Essa probabilidade é chamada de valor-p, e corresponde

à probabilidade da estatística de teste ser igual ao valor observado ou a um valor ainda mais extremo. É calculado presumindo que a hipótese nula H_0 é verdadeira.

Quanto menor o valor-p, mais forte é a evidência contra H_0. Rejeitar H_0 significa que a probabilidade de obter uma amostra que evidencie que H_0 é verdadeira é muito reduzida, pelo que o teste é, num certo sentido, conclusivo e fornece informações suficientes para se rejeitar H_0. Caso contrário, não rejeitar H_0 significa que o teste é inconclusivo e que a amostra não contém evidência suficiente para se fazer essa rejeição [5]. Quando se rejeita H_0 diz-se que o teste é estatisticamente significativo.

Como a decisão para rejeitar ou não a hipótese nula é tomada de acordo com os elementos de uma amostra, e não de toda a população em causa, essa decisão está sujeita a erros. Com base nos resultados obtidos numa amostra, não é possível tomar decisões que sejam definitivamente corretas. Existem dois tipos de erros quando se realiza um teste de hipóteses:

- O erro de tipo I consiste em rejeitar H_0 quando esta é, de facto, verdadeira. A probabilidade de o cometer é denotada pela letra α;
- O erro de tipo II consiste em não rejeitar H_0 quando H_1 é verdadeira. A probabilidade de o cometer é denotada pela letra β.

Esquematizando, obtêm-se os resultados da Tabela 2.1.

	Não rejeitar H_0	Rejeitar H_0
H_0 Verdadeira	Decisão correta	Erro do tipo I
H_0 Falsa	Erro do tipo II	Decisão correta

TABELA 2.1: Erros cometidos num teste de hipóteses

Enquanto que α consiste num único valor, β é uma função, ou seja, para cada valor de H_1 existe um valor de β. Designa-se por potência do teste a probabilidade de rejeitar a hipótese nula quando a hipótese alternativa é verdadeira, ou seja, $1 - \beta$.

2.1 Descrição de um procedimento de um teste de hipóteses

Apresenta-se agora um exemplo elucidativo da aplicação de um teste de hipóteses. Suponha-se que os valores de glicemia em jejum seguem uma distribuição normal, quer para uma população diabética quer para uma população não diabética. Assuma-se que as variâncias para as duas populações são iguais. Deseja-se testar se, em média, os valores

de glicemia em jejum em pessoas diabéticas são diferentes dos valores de glicemia em jejum em pessoas não diabéticas.

Passo 1: Definição das hipóteses

Primeiramente, é necessário escolher a hipótese nula e a hipótese alternativa com base no problema em questão.

Atendendo ao exemplo apresentado e definindo μ_D (respetivamente, $\mu_{\overline{D}}$) como a variável aleatória que representa os valores de glicemia em jejum na população diabética (respetivamente, população não diabética), quer-se testar se as médias μ dos valores de glicemia são iguais na população diabética e na população não diabética. Assim, têm-se as seguintes hipóteses nula e alternativa:

$$H_0 : \mu_D = \mu_{\overline{D}}$$
$$H_1 : \mu_D \neq \mu_{\overline{D}}$$

Passo 2: Estatística de teste

De seguida, estabelece-se a estatística de teste que vai ser usada para testar a hipótese nula a partir da teoria estatística e das informações disponíveis na amostra. Como a variável de interesse segue uma distribuição normal em cada uma das populações, mas como não se sabe a variância destas distribuições, então a estatística de teste T é:

$$T = \frac{(\overline{X}_D - \overline{X}_{\overline{D}}) - (\mu_D - \mu_{\overline{D}})}{\sqrt{S^2 \times \left(\frac{1}{n_D} + \frac{1}{n_{\overline{D}}}\right)}} \tag{2.1}$$

onde $\overline{X}_D$ e $\overline{X}_{\overline{D}}$ são as médias amostrais dos valores de glicemia em jejum nas amostras da população diabética e da população não diabética, respetivamente, n_D e $n_{\overline{D}}$ correspondem ao tamanho das amostras diabética e não diabética, respetivamente, e S^2 é um estimador da variância comum, obtido através da média ponderada dos estimadores das variâncias de cada população:

$$S^2 = \frac{(n_D - 1)S_D^2 + (n_{\overline{D}} - 1)S_{\overline{D}}^2}{n_D + n_{\overline{D}} - 2} \tag{2.2}$$

onde S_D^2 e $S_{\overline{D}}^2$ são as variâncias amostrais em diabéticos e não diabéticos, respetivamente.

A estatística de teste T segue uma distribuição *t-Student* com $n_D + n_{\overline{D}} - 2$ graus de liberdade. Por esse motivo, o teste é designado por teste-t (para duas amostras com variâncias iguais mas desconhecidas).

Considere-se a amostra representada na Tabela 2.2.

Diabéticos	126	134	134	164	145	131	137	153	141	-
Não diabéticos	86	81	78	76	83	89	68	82	90	92

TABELA 2.2: Valores do nível de glicemia em jejum na amostra da população diabética e na amostra da população não diabética.

Obtém-se o seguinte valor para a estatística de teste:

$$t = \frac{58.06 - 0}{\sqrt{9.71 \times 0.46}} = 13.013$$

Passo 3: Valor-p

Para interpretar o valor da estatística de teste, usa-se um resumo da probabilidade da evidência na amostra contra a hipótese nula, H_0. Assumindo que H_0 é verdadeira, considera-se a distribuição da estatística de teste. Relativamente ao exemplo, a distribuição é uma *t-Student* com $n_D + n_{\overline{D}} - 2$ graus de liberdade. Se o valor da estatística de teste calculada pertencer à "cauda"da distribuição, está longe do que H_0 prevê. Se H_0 fosse realmente verdadeiro, tal valor seria incomum [7].

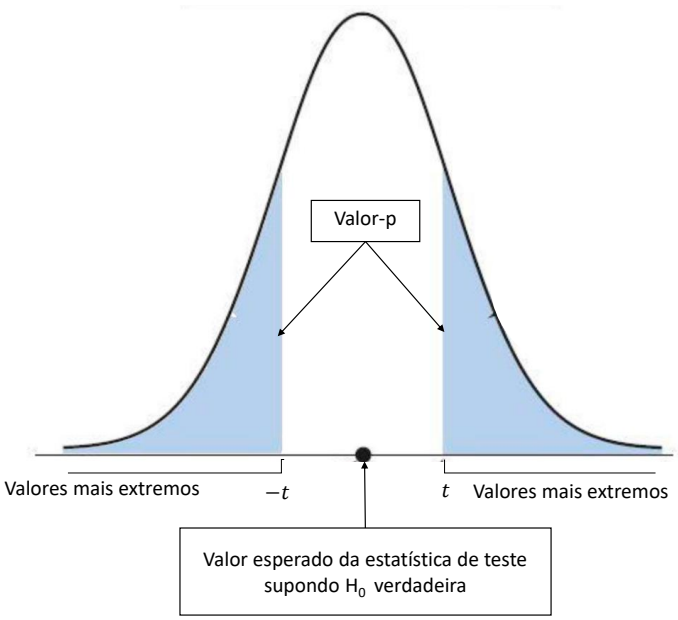


FIGURA 2.1: Esquema do procedimento para um teste de hipóteses bilateral. Adaptado de [7].

O "quão longe na cauda a estatística de teste cai"define a probabilidade de cauda do valor observado e valores ainda mais extremos (ou seja, ainda mais longe do que H_0 prevê) - ver Figura 2.1. Essa probabilidade é chamada de valor-p e representa a área sombreada da cauda da distribuição amostral. Quanto menor for o valor-p, mais forte será a evidência contra H_0.

O valor-p é então definido como a probabilidade de se obter uma estatística de teste igual ou mais extrema do que o valor obtido na amostra, assumindo que a hipótese nula é verdadeira.

Atendendo a que a hipótese alternativa é $H_1 : \mu_D \neq \mu_{\overline{D}}$, no cálculo do valor-p considera-se a área das duas caudas. Assim, o valor-p para o exemplo exposto é:

$$p = 2 \times P(T \leq -13.013) = 2.884 \times 10^{-10}$$

onde $T \sim t(17)$.

Passo 4: Conclusão/Decisão

Rejeita-se a hipótese nula se e só se o valor-p for menor do que o nível de significância α pré-estabelecido. Caso contrário, nenhuma conclusão pode ser retirada. Os valores de α mais comuns são 5% e 1%.

No exemplo acima, suponha-se que o nível de significância α foi fixado em 5%. Como $2.884 \times 10^{-10} < 0.05$ então rejeita-se H_0, e pode-se concluir que, com um nível de significância de 0.05, o valor médio de glicemia em jejum para a população diabética é diferente do valor médio de glicemia em jejum para a população não diabética.

Considere-se, agora, que o objetivo do problema é testar se o valor médio de glicemia é superior na população diabética. Então H_1 é $\mu_D > \mu_{\overline{D}}$ e o valor-p é $P(T > 13.013) = 1.442 \times 10^{-10}$. Nesta situação também se rejeita H_0 e conclui-se que o valor médio de glicemia em jejum na população diabética é superior ao valor médio de glicemia em jejum na população não diabética.

Caso contrário, se a hipótese alternativa for $H_1 : \mu_D < \mu_{\overline{D}}$ então está-se a averiguar se o valor médio de glicemia em jejum na população diabética é inferior ao respetivo valor na população não diabética. Obter-se-ia $p = P(T < 13.013) \approx 1$ e portanto, nenhuma conclusão poderia ser retirada.

2.2 Interpretação do valor-p

Os testes de hipóteses são apropriados quando se deseja quantificar a evidência contra H_0. Mas, e quando se quer quantificar a evidência a favor de H_0 - ou seja, contra H_1? Quando se realizam testes de hipóteses, não se consegue perceber se os resultados não significativos suportam H_0 ou se os resultados são simplesmente não informativos.

Desta forma, não se retira nenhuma conclusão caso o valor-p seja maior do que o nível de significância.

Quando se faz um teste de hipóteses para comparar duas situações, A e B, obtém-se o respetivo valor-p. Quando este valor está abaixo do nível de significância α, então é improvável que se encontrem diferenças tão ou mais extremas do que as encontradas na amostra, supondo que H_0 é verdadeiro, como já foi referido no capítulo 2. Para um tamanho da amostra fixo, quanto menor o valor-p, mais evidências se tem contra H_0.

A situação de ter um valor-p inferior ao nível de significância pode surgir devido a três motivos: variabilidade da amostra, ou seja, má sorte na obtenção da mesma; tamanho da amostra reduzido, pelo que não se consegue chegar à conclusão de que H_0 é falso; ou, realmente, à existência de evidências para aceitar H_0.

De facto, quando a amostra é pequena, qualquer explicação é plausível. À medida que o tamanho da amostra aumenta, seria intuitivo que, se valores-p baixos fornecem evidências contra H_0, então valores-p altos deviam fornecer evidências a favor de H_0. Portanto, seria de esperar que nas simulações em que o tamanho do efeito é realmente 0, os valores-p altos fossem mais frequentes, especialmente em tamanhos amostrais grandes. Porém, isso não acontece.

Na Figura 2.2 (a) estão representados vários histogramas que mostram a distribuição de valores-p obtidos a partir de 1.000 testes-t unicaudais de tamanho n de uma distribuição normal com média μ e desvio padrão $\sigma = 1$.

Observa-se que, quando há um efeito ($\mu > 0$), o aumento do tamanho amostral (para baixo, na figura) ou do tamanho do efeito (para a direita, na figura) leva a um deslocamento da distribuição dos valores-p para a esquerda, o que mostra que há mais evidências contra H_0. Neste caso, os valores-p < 0.05 estão corretos e, portanto, são apresentados a verde, e os valores-p > 0.05 são considerados erros e são apresentados a vermelho.

No entanto, um pouco contra-intuitivamente, o contrário não é verdadeiro. Na ausência de um efeito ($\mu = 0$) o aumento do tamanho da amostra não leva a um deslocamento para a direita dos valores-p, isto é, não aumenta os valores-p mais altos. Em vez disso, a distribuição é uniforme, com todos os valores-p igualmente prováveis. Neste caso, valores-p < 0.05 representam alarmes falsos e são mostrados a vermelho, e valores-p > 0.05 representam decisões corretas e são indicados a verde.

Assim, pode-se concluir que valores-p mais altos não são uma métrica confiável para mais evidências a favor de H_0, pois casos correspondentes a muita evidência a favor de H_0

(por exemplo, $\mu = 0$ com $n = 100$) não tornam os valores-p altos mais frequentes [8]. Em vez disso, qualquer valor-p permanece tão provável quanto qualquer outro. Portanto, valores-p significativos indicam evidências contra H_0, mas valores-p não significativos não permitem concluir que os dados suportam H_0.

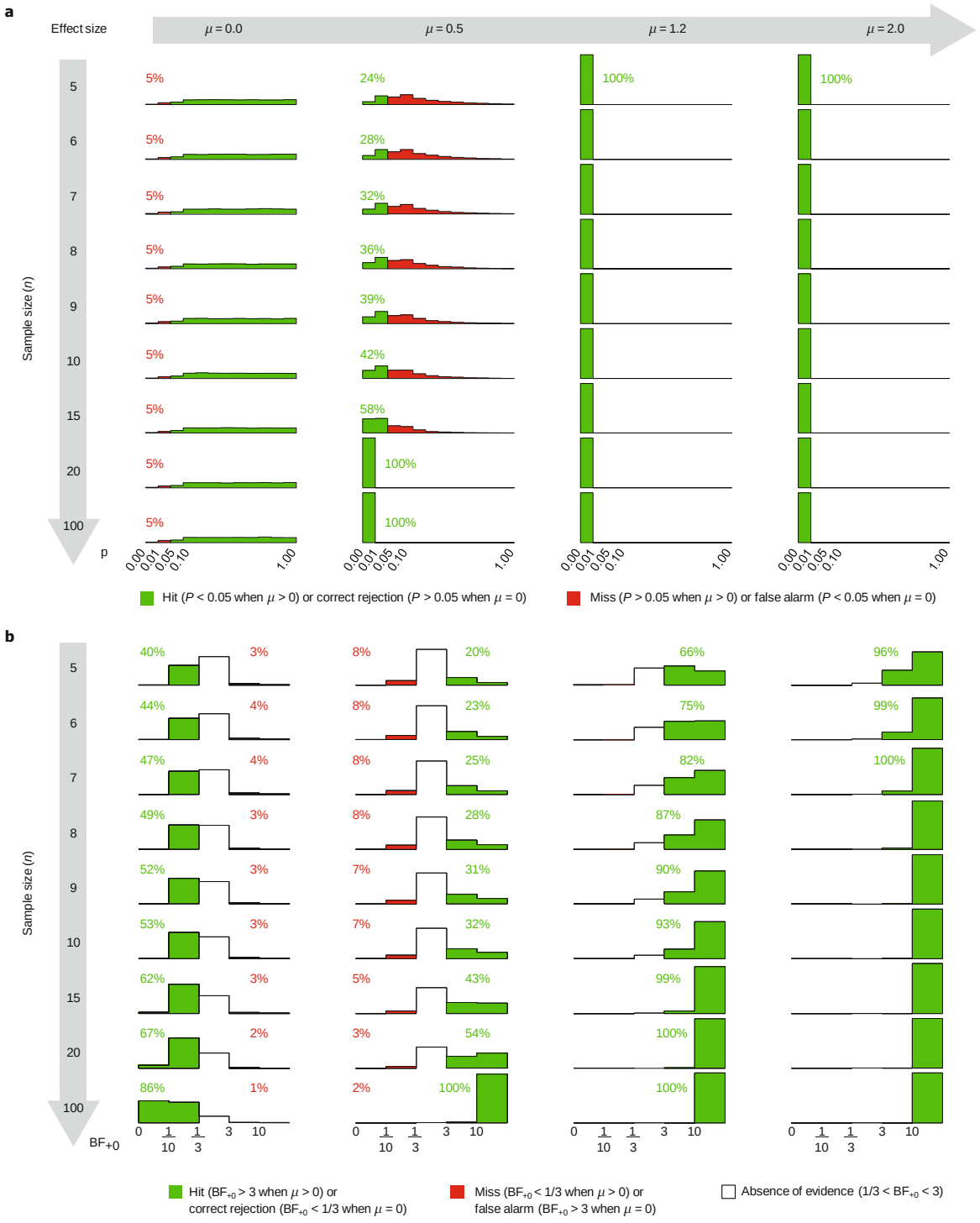


FIGURA 2.2: Valor-p e BF_{+0} de um teste-t em função do tamanho do efeito e do tamanho da amostra. Adaptado de [8].

2.3 Uma solução Bayesiana

Em contraste com os testes de hipóteses frequencistas, que se concentram exclusivamente na rejeição da hipótese nula, os testes de hipóteses bayesianos visam quantificar a plausibilidade relativa de H_0 contra H_1.

Considere-se um teste de hipóteses baseado num conjunto de observações D, que surgem de uma de duas hipóteses H_0 e H_1 de acordo com uma densidade de probabilidade $p(D|H_0)$ ou $p(D|H_1)$. Dadas as probabilidades *a priori* das hipóteses $p(H_0)$ e $p(H_1) = 1 - p(H_0)$, os dados são responsáveis pelas frequências relativas *a posteriori* das hipóteses $p(H_0|D)$ e $p(H_1|D) = 1 - p(H_0|D)$, sendo que estas novas probabilidades representam a evidência fornecida pelos dados para cada umas das hipóteses [9].

Pelo teorema de Bayes, tem-se:

$$p(H_k|D) = \frac{p(D|H_k)p(H_k)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)} \text{ para k} = 0, 1 \tag{2.3}$$

Assim, calculando o quociente entre $p(H_0|D)$ e $p(H_1|D)$, obtém-se:

$$\underbrace{\frac{p(H_0|D)}{p(H_1|D)}}_{\textit{odds a posteriori}} = \underbrace{\frac{p(D|H_0)}{p(D|H_1)}}_{\text{fator de Bayes}} \times \underbrace{\frac{p(H_0)}{p(H_1)}}_{\textit{odds a priori}} \tag{2.4}$$

onde o fator de Bayes foi definido como:

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} \tag{2.5}$$

isto é, a razão entre o *odds a posteriori* de H_0 e o *odds a priori* de H_0, independentemente do valor do *odds a priori*.

O fator de Bayes representa a evidência do desempenho preditivo relativo de uma hipótese contra a outra [8]. Dado que $BF_{01} = \frac{p(D|H_0)}{p(D|H_1)}$ e que $BF_{10} = \frac{p(D|H_1)}{p(D|H_0)}$, tem-se $BF_{10} = BF_{01}^{-1}$.

A interpretação do fator de Bayes é a seguinte: se $BF_{10} = 8$, então há 8 vezes mais evidência para H_1 do que para H_0, ou seja, os dados utilizados foram previstos 8 vezes melhor por H_1 do que por H_0.

Em 1961, Jeffreys [10] propôs valores de referência para orientar a interpretação da força da evidência dada pelo valor do fator de Bayes. Esses valores ainda permanecem em uso e são apresentados na Tabela 2.3.

BF_{10}	Interpretação
$< \frac{1}{10}$	Evidência forte a favor de H_0
$\frac{1}{10}$ - $\frac{1}{3}$	Evidência moderada a favor de H_0
$\frac{1}{3}$ - 3	Não há evidência suficiente para tirar uma conclusão
3 - 10	Evidência moderada a favor de H_1
> 10	Evidência forte a favor de H_1

TABELA 2.3: Graus de evidência propostos por H. Jeffreys.

Este sistema permite diferenciar entre evidências de ausência de efeito - fator de Bayes inferior a $1/3$ - e ausências de evidência - fator de Bayes entre $1/3$ e 3.

A Figura 2.2 (b) é análoga à Figura 2.2 (a) e pretende estudar a distribuição de BF_{+0}. Como tal, são geradas 1000 amostras de tamanho n de uma distribuição normal $N(\mu, 1)$ e é calculado o fator de Bayes para cada uma dessas amostras. Denote-se H_1 unilateral por H_+ para indicar a direção do efeito e o fator de Bayes correspondente $BF_{+0} = \frac{H_+}{H_0}$. A hipótese $H_0 : \mu = 0$ afirma que o efeito está ausente, enquanto que a hipótese alternativa, $H_+ : \mu > 0$, afirma que o efeito é positivo.

Pode-se observar que, quando o efeito está ausente ($\mu = 0$), o teste Bayesiano raramente chega à conclusão errada de que um efeito está presente, porque exibe uma barra vermelha - que representa os erros cometidos no teste - muito baixa. O mesmo acontece na abordagem frequencista - Figura 2.2 (a). No entanto, ao contrário desta, a abordagem bayesiana fornece evidências crescentes para a ausência de um efeito ($BF_{+0} < 1/3$ - barras verdes) com o aumento do tamanho da amostra.

No mesmo sentido, quando um efeito está presente ($\mu > 0$), a evidência de um efeito positivo ($BF_{+0} > 3$ - barras verdes) aumenta à medida que o tamanho da amostra e o tamanho do efeito aumentam. Na situação $\mu = 0.5$, as falhas ($BF_{+0} < 1/3$ - barras vermelhas) são raras, enquanto que, na situação $\mu = 1.2$ ou 2, as falhas são ausentes.

Também é possível notar que os casos inconclusivos - barra branca da Figura 2.2 (b) - que são os casos em que há ausência de evidência para qualquer uma das hipóteses, isto é, $1/3 < BF < 3$, tornam-se cada vez mais raros à medida que o tamanho da amostra aumenta.

Pode-se concluir que, ao contrário do valor-p, o BF_{10} tanto consegue quantificar evidências a favor de H_1 como a favor de H_0. Obviamente que essa quantificação se torna mais forte à medida que o tamanho da amostra aumenta. Contudo, quando o tamanho do efeito é

diferente de 0 mas muito próximo dele, o BF_{10} apresenta uma evidência a favor de H_1 menos frequente do que o procedimento frequencista, não sendo no entanto uma diferença considerável. Porém, conforme os tamanhos do efeito se tornam maiores, os dois métodos divergem mais. O valor-p torna-se mais significativo mesmo para tamanhos do efeito reduzidos, enquanto que o BF_{10} exige tamanhos do efeito maiores.

Capítulo 3

Testes de Hipóteses Múltiplas

Em muitas áreas de aplicação da estatística é necessária a realização de vários testes em simultâneo devido ao grande número de hipóteses existentes. como por exemplo comparar diversos fármacos para descobrir quais, possivelmente, produzem um resultado superior na cura de uma doença; comparar inúmeros genes e identificar quais são diferencialmente expressos em experiências de expressão genética; e comparar diferentes grupos demográficos em termos das suas atitudes em relação a questões de interesse. Este cenário é geralmente referido como um teste de hipóteses múltiplas.

A abordagem deste tipo de testes consiste em fazer cada comparação independentemente usando um teste de hipóteses adequado [11], ou seja, fazer um teste de hipóteses simples para cada par de hipóteses respeitantes à mesma situação (H_0 e H_1). Quando muitas hipóteses são testadas, e cada teste tem uma probabilidade de erro de Tipo I especificada, a probabilidade de que pelo menos um erro de Tipo I seja cometido na totalidade dos testes aumenta, frequentemente de forma acentuada, com o número de hipóteses [12].

Considerando α como o nível de significância em cada teste, isto é, a probabilidade de um erro de tipo I ocorrer num único teste, então a probabilidade de não se cometer um erro desse tipo num teste é $1-\alpha$. Consequentemente, se m testes independentes forem realizados, a probabilidade de não se cometerem erros de tipo I nos m testes é $(1-\alpha)^m$. Como $0 < \alpha < 1$, tem-se $(1-\alpha)^m < (1-\alpha)$. Assim, pode-se concluir que a probabilidade de não se cometer erros de tipo I em m testes ($m > 1$) é menor do que no caso de um único teste. Dessa maneira, a probabilidade mínima de se cometer pelo menos um destes erros nos m testes é $\tilde{\alpha} = 1-(1-\alpha)^m$, que é maior do que no caso de um teste. A probabilidade de se cometer pelo menos uma falsa rejeição na totalidade dos m testes de hipóteses é designada por taxa de erro da família (em inglês, Family-Wise Error Rate - FWER).

O fenómeno descrito é ilustrado na Figura 3.1, que mostra que, à medida que o número de testes m aumenta, a probabilidade de se cometer pelo menos um erro do tipo I também aumenta. Também é visível que, quando se testa por exemplo 100 hipóteses nulas, a probabilidade de ocorrer pelo menos um erro do tipo I é, aproximadamente, 1.

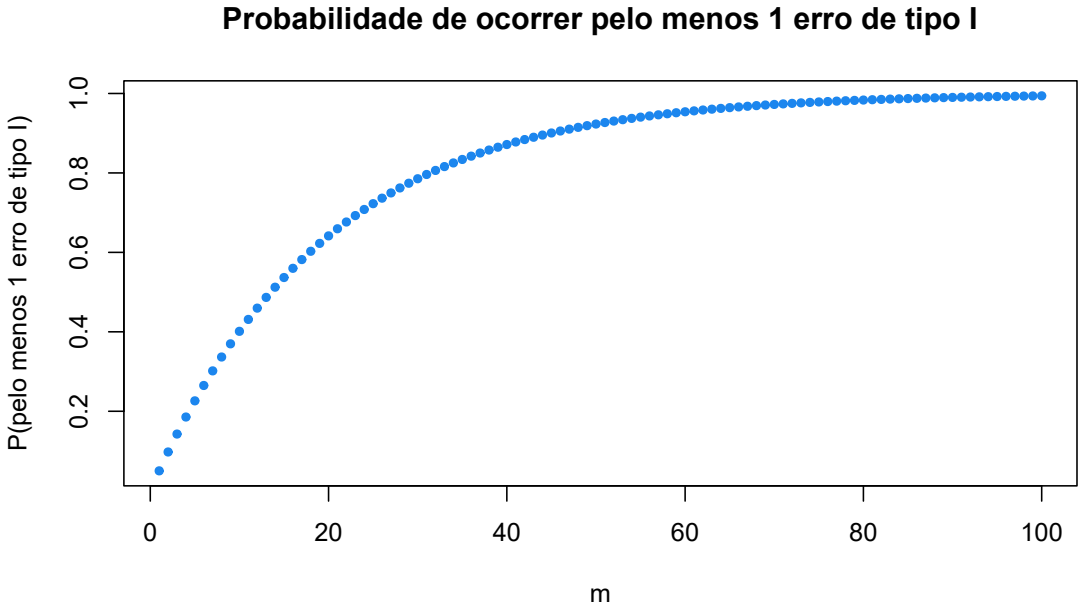


FIGURA 3.1: Problema dos testes de hipóteses múltiplas

Para solucionar este problema, existem vários métodos que tornam os testes individuais mais conservadores, de modo a minimizar o número total de erros de tipo I, mantendo a taxa de erro global. Estas técnicas exigem um limiar de significância mais rigoroso para as comparações individuais, de modo a compensar o número de inferências que vão ser realizadas. No entanto, estes procedimentos têm uma redução na potência dos testes individuais e não são fáceis de comparar entre si.

Tipicamente, as várias hipóteses de um teste de hipóteses múltiplas são consideradas independentes, embora existam métodos para situações de dependência. Esses métodos não serão alvo de estudo nesta dissertação.

Considere-se uma situação de testes múltiplos com m hipóteses nulas, denotadas por $H_{01}, H_{02}, ..., H_{0m}$, com valores-p correspondentes $p_1, p_2, ..., p_m$, respetivamente. O i-ésimo teste diz-se significativo se se rejeitar a sua hipótese nula H_{0i}. Considerando o mesmo nível de significância α em cada um dos testes individuais, o i-ésimo teste será significativo se $p_i < \alpha$.

Suponha-se que, destas hipóteses nulas, m_0 são verdadeiras e m_1 são falsas. Num contexto real, m_0 e m_1 não são conhecidos. Na Tabela 3.1 estão retratados os possíveis cenários para os m testes de hipóteses.

	Não rejeitar H_0	Rejeitar H_0	Total
H_0 Verdadeira	U	V	m_0
H_0 Falsa	T	S	m_1
Total	$m - R$	R	m

TABELA 3.1: Erros cometidos em m testes de hipóteses

- R representa o número total de rejeições;
- V e T são o número de erros cometidos do tipo I e II, respetivamente;
- S e U são o número de hipóteses nulas corretamente rejeitadas e corretamente não rejeitadas, respetivamente.

O número de hipóteses nulas rejeitadas R é aleatório assim como V, T, S e U. Para além disso, os valores de V, T, S e U são desconhecidos.

Com esta notação, a proporção de falsas rejeições é $\frac{V}{R}$ com $R > 0$ e a proporção de falsas não rejeições é $\frac{T}{m-R}$ com $m - R > 0$.

Como foi definido anteriormente, a taxa de erro de família é a probabilidade de se cometer pelo menos um erro de tipo I e pode ser descrita através das variáveis apresentadas na Tabela 3.1 do seguinte modo:

$$FWER = P(V \geq 1) \tag{3.1}$$

3.1 Métodos baseados nos valores-p ordenados

Muitos dos métodos que controlam FWER são baseados na ideia de valores-p ordenados. Ou seja, antes de realizar qualquer ajuste, ordena-se os m valores-p: $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$, com as hipóteses nulas correspondentes $H_{0(1)}, H_{0(2)}, ..., H_{0(m)}$. Os procedimentos que se seguem baseiam-se nesta estrutura.

Os primeiros métodos de ajuste dos testes de hipóteses múltiplas foram desenvolvidos com o intuito de controlar a taxa de erro de família, e ainda são comummente usados nos dias de hoje [3]. Têm a desvantagem de serem excessivamente conservadores, resultando em testes com baixa potência.

3.1.1 Método simples de Bonferroni

Um procedimento clássico que garante o controlo da FWER é o "Procedimento de Bonferroni"[12]. Relembrando que a probabilidade de cometer pelo menos um erro de tipo I em m testes individualmente é $\alpha' = 1-(1-\alpha)^m$, e que, para α pequeno, $(1-\alpha)^m \approx 1 - m\alpha$, obtém-se $\alpha' = m\alpha$ e portanto $\alpha = \frac{\alpha'}{m}$. Assim, para qualquer $q > 0$ este método efetua cada teste a um nível de significância de $\frac{q}{m}$.

$$\begin{aligned} &\text{Se para o teste i o valor-p que lhe corresponde satisfaz } p_{(i)} < \frac{q}{m}, \\ &\text{então, rejeita-se a hipótese nula } H_{0(i)}. \end{aligned} \tag{3.2}$$

O procedimento garante que FWER $< q$.

3.1.2 Método de Holm

Um método mais potente do que o método simples de Bonferroni inclui um algoritmo sequencial e foi proposto por Sture Holm [13] em 1979 da seguinte maneira:

$$\begin{aligned} &\text{Definindo-se } k = \min\left\{ i : p_{(i)} > \frac{q}{m+1-i} \right\}, \\ &\text{então, rejeita-se todas } H_{0(i)}\ i = 1, 2, ..., k-1. \end{aligned} \tag{3.3}$$

O processo consiste nas seguintes etapas: na primeira etapa, $H_{0(1)}$ é rejeitado se $p_{(1)} \leq q/m$. Se $H_{0(1)}$ não for rejeitado, todas as outras hipóteses também não serão rejeitadas; Na segunda etapa, $H_{0(2)}$ é rejeitado se $p_{(2)} \leq q/(m-1)$. Continuando o processo até alguma etapa j, $H_{0(j)}$ é rejeitado se e só se todos $H_{0(i)}$ tiverem sido rejeitados, $i < j$, e $p_{(j)} \leq q/(m-j+1)$.

O controlo pelos métodos apresentados, Bonferroni e Holm, não requer nenhuma suposição sobre a dependência entre os testes.

3.1.3 Método de Hochberg

Em 1988, Yosef Hochberg [14] apresentou um método muito semelhante ao de Holm mas mais potente. O método começa por analisar $H_{0(m)}$. Depois aplica-se sucessivamente a seguinte regra:

$$\begin{aligned} &\text{Definindo-se } k = \max\left\{ i : p_{(i)} \leq \frac{i}{m} q \right\} \\ &\text{então, rejeita-se todas } H_{0(i)}\ i = 1, 2, ..., k. \end{aligned} \tag{3.4}$$

Tal como as outras estratégias de controlo, esta também mantém a taxa de erro a um nível q.

3.1.4 Método de Hommel

O procedimento de Hommel [15], publicado também em 1988, é mais potente do que o de Hochberg. A taxa de erro é também controlada a um nível q.

$$\begin{aligned} &\text{Define-se } k = \max\left\{ i : p_{(m-i+j)} > \frac{jq}{i}, \forall\, j = 1, ..., i \right\}. \\ &\text{Se existir um tal } k\text{, rejeitam-se todas } H_{0(i)} \text{ com } p_i \leq \frac{q}{k}; \\ &\text{Se não existir um tal } k\text{, rejeitam-se todas as hipóteses.} \end{aligned} \tag{3.5}$$

O controlo destes dois métodos foi provado para testes independentes.

O controlo da FWER é apropriado quando se quer proteger contra qualquer rejeição errada de uma hipótese nula. No entanto, em muitos problemas de multiplicidade, o número de rejeições erradas deve ser levado em consideração e não apenas a questão de se ter cometido algum erro. O uso da correção de Bonferroni pode ser muito restrito e até pode levar à não rejeição de hipóteses nulas que são realmente falsas.

Exemplo 3.1. Considere-se um teste com 5 hipóteses de onde se obtiveram os seguintes valores-p ordenados: $p_{(1)} = 0.002$, $p_{(2)} = 0.013$, $p_{(3)} = 0.028$, $p_{(4)} = 0.041$ e $p_{(5)} = 0.106$.

Para controlar a FWER a 0.05, o procedimento de Bonferroni usa $\frac{0.05}{5} = 0.01$ e, portanto, rejeita-se a hipótese nula com valor-p mais baixo $(p_{(1)} = 0.002)$.

Usando o método de controlo de Holm com $q = 0.05$, compara-se sequencialmente cada $p_{(i)}$ com $\frac{0.05}{5+1-i}$, começando com $p_{(1)}$. O primeiro valor-p a satisfazer a restrição é $p_{(2)}$, porque

$$p_{(2)} = 0.013 > \frac{0.05}{5+1-2} = 0.0125$$

Portanto, rejeita-se a hipótese nula com o valor-p correspondente $p_{(1)} = 0.002$.

Pelo método de Hochberg, compara-se cada $p_{(i)}$ com $\frac{0.05 \times i}{5}$, começando com $p_{(5)}$. O primeiro valor-p que satisfaz a restrição é $p_{(2)}$, com

$$p_{(2)} = 0.028 < \frac{0.05 \times 3}{5} = 0.03$$

Assim, rejeita-se as três hipóteses nulas com os valores-p mais baixos $(p_{(1)} = 0.002$, $p_{(2)} = 0.013$ e $p_{(3)} = 0.028)$.

Por último, o método de Hommel usa $p_{(i)} \leq \frac{0.05}{4} = 0.0125$. Porque, $k = 4$ é o maior valor que satisfaz $\left\{ p_{(m-k+j)} > \frac{jq}{k}, \forall\, j = 1, ..., k \right\}$ e, portanto, rejeita-se apenas a hipótese nula com $p_{(1)} = 0.002$.

3.2 Fator de Bayes

Na Seção 2.3 viu-se a aplicação do Fator de Bayes a um teste de hipóteses simples. Numa situação de testes de hipóteses múltiplas, em que m hipóteses serão testadas, nada muda. Simplesmente, aplica-se o método m vezes.

Como se observou na Figura 2.2, o BF depende do tamanho da amostra. Quanto maior for o tamanho, melhor será o desempenho do método. Em contraste, os testes múltiplos com base nos valores-p (por exemplo, o método de Bonferroni ou o de BH) dependem principalmente, do número m de testes.

Assim, tem-se que:

- A decisão bayesiana é baseada na estatística de cada teste e no tamanho da amostra, n, mas não no número de testes, m.
- A decisão dos métodos clássicos para testes múltiplos baseia-se na estatística de cada teste e no número de testes, m, mas não no tamanho da amostra, n.

Capítulo 4

Taxa de falsas descobertas

4.1 Definição

Num teste de hipóteses múltiplas, a proporção de erros cometidos pela falsa rejeição de hipóteses nulas é,

$$Q = \frac{V}{V+S} \tag{4.1}$$

onde Q, V e S são as variáveis aleatórias previamente definidas na Tabela 3.1.

Quando $V + S = 0$, define-se $Q = 0$ uma vez que, se nenhuma hipótese nula foi rejeitada, então não é possível cometer erros de falsa rejeição. Portanto, podemos definir a variável aleatória que representa a proporção das hipóteses nulas rejeitadas erradamente da seguinte maneira,

$$Q = \begin{cases} \frac{V}{R}, & R > 0 \\ 0, & R = 0 \end{cases} \tag{4.2}$$

No contexto estatístico, descoberta refere-se à rejeição de uma hipótese nula. Portanto, uma descoberta falsa é uma rejeição incorreta de uma hipótese e taxa de falsas descobertas (em inglês, false discovery rate - FDR) [6] é definida como a proporção esperada dessas rejeições, ou seja,

$$E(Q) = \begin{cases} E\left(\frac{V}{R}\right), & R > 0 \\ 0, & R = 0 \end{cases} \tag{4.3}$$

onde $E(.)$ denota o valor esperado. A FDR pode ser expressa de forma equivalente como

$$\begin{aligned} FDR &= E\left(\tfrac{V}{R}|R>0\right) P(R>0) + E\left(\tfrac{V}{R}|R=0\right) P(R=0) \\ &= E\left(\tfrac{V}{R}|R>0\right) P(R>0) \end{aligned} \tag{4.4}$$

4.2 Propriedades

Existem algumas relações entre a FDR e a FWER [6, 16] dependendo de m_0 - o número de hipóteses nulas verdadeiras. Se este for estritamente inferior ao número total de hipóteses nulas, então a FDR é menor ou igual à FWER. Se m_0 for igual ao número total de hipóteses nulas, então as duas taxas são iguais:

1. $m_0 = m \Rightarrow FDR = FWER$
2. $m_0 < m \Rightarrow FDR \leq FWER$

De facto, no primeiro caso, tem-se que todas as hipóteses nulas são verdadeiras e, portanto, $V = R$. Assim,

$$\frac{V}{R} = \begin{cases} 0, & V = 0 \\ 1, & V \geq 1 \end{cases} \tag{4.5}$$

e portanto,

$$\begin{aligned} FDR = E\left(\frac{V}{R}\right) &= 0 * P(V = 0) + 1 * P(V \geq 1) \\ &= P(V \geq 1) = FWER \end{aligned} \tag{4.6}$$

Se nem todas as hipóteses nulas forem verdadeiras (segundo caso), então $V \leq R$ e

$$\begin{aligned} FDR = E\left(\frac{V}{R}\right) &= E\left(\frac{V}{R} | V \geq 1\right) \times P(V \geq 1) \\ &\leq P(V \geq 1) \end{aligned} \tag{4.7}$$

dado que $V \leq R$.

A Equação 4.7 mostra que FDR é sempre menor ou igual que FWER, implicando que qualquer abordagem que controle FWER também controlará FDR. O contrário, no entanto, não se verifica, ou seja, o controlo da FDR geralmente não implica o controlo da FWER. Controlar FDR é uma condição menos rigorosa do que controlar FWER e, portanto, é de esperar que os procedimentos que controlam apenas a FDR sejam mais potentes.

4.3 Procedimento de controlo

Nesta secção serão abordados procedimentos que controlam globalmente a FDR a um nível q, não necessariamente controlando as probabilidades de erro de tipo I a um certo nível para cada teste.

Imagine-se que se pretende testar $H_1, H_2, ..., H_m$ com base nos valores-p que lhes são correspondentes $p_1, p_2, ..., p_m$. Considerem-se os valores-p ordenados: $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ e denote-se por $H_{0(i)}$ a hipótese nula que corresponde a $p_{(i)}$. O procedimento publicado por Benjamini e Hochberg para controlar a FDR [6] é o seguinte:

$$\begin{aligned} &\text{Para um dado valor de q, determine-se } k = max\left\{i : p_{(i)} \leq \frac{i}{m}q\right\}; \\ &\text{De seguida, rejeita-se todos } H_{0(i),\, i=1,...,k}. \end{aligned} \tag{4.8}$$

O procedimento começa, então, por comparar $p_{(m)}$ com q, e se for menor todas as hipóteses são rejeitadas. Se $p_{(m)} > q$, prossegue-se para valores-p menores até que um deles satisfaça a condição. O procedimento termina, se não for antes, comparando $p_{(1)}$ com $\frac{q}{m}$.

Benjamini e Hochberg, em [6], provam que, para estatísticas de teste independentes e para qualquer configuração das hipóteses nulas falsas, o procedimento acima controla a FDR ao nível q.

Exemplo 4.1. Considere-se um estudo cujo objetivo consiste em identificar os genes, de entre uma lista de 10 genes, que têm uma expressão genética diferente em duas populações distintas. Para cada gene possui-se a expressão genética de seis indivíduos, três de cada população.

Considere-se os valores-p referentes a cada gene (valores esses que foram calculados cada um por um teste de hipóteses simples adequado) representados na Tabela 4.1.

Gene	A	B	C	D	E	F	G	H	I	J
Valor-p	0.13	0.051	0.0001	0.37	0.80	0.0009	0.026	0.51	0.036	0.029

TABELA 4.1: Valores-p associados a cada gene.

Usando o procedimento de Benjamini-Hochberg, é agora necessário ordenar os valores-p observados - Tabela 4.2.

Índice	1	2	3	4	5	6	7	8	9	10
Gene	C	F	G	J	I	B	A	D	H	E
Valor-p	0.0001	0.0009	0.026	0.029	0.036	0.051	0.13	0.37	0.51	0.80

TABELA 4.2: Valores-p ordenados associados a cada gene.

Considerando $q = 0.05$, verifica-se que $p_{(10)} = 0.8 > 0.05$ e, portanto, procede-se para valores-p mais baixos. O primeiro valor-p a satisfazer a condição é $p_{(2)}$:

$$p_{(2)} = 0.0009 \leq \frac{2}{10}0.05 = 0.01$$

Assim, rejeita-se $H_{(1)}$ e $H_{(2)}$ que correspondem aos genes C e F. Para estes genes, as expressões genéticas médias para as duas amostras são consideradas significativamente diferentes, e, portanto, pode-se concluir que, a este nível, as duas populações apresentam expressões genéticas diferentes. Para os restantes genes, nada se pode concluir.

4.3.1 Valor-p ajustado

Pode-se olhar para o algoritmo de uma outra forma, utilizando os chamados valores-p ajustados. Ao testar uma única hipótese, o valor-p fornece mais informação do que apenas rejeitar ou não a hipótese nula. A extensão desse conceito ao contexto de testes múltiplos não é necessariamente direta. Um conceito que permite a generalização a partir do teste de uma única hipótese para o contexto múltiplo é o valor-p ajustado. O valor-p ajustado para uma hipótese nula em particular é o menor nível de erro nominal de tipo I (para o teste múltiplo de todas as m hipóteses) em que se rejeitaria essa hipótese nula.

Assim, o procedimento descrito em 4.8 pode ser feito da seguinte maneira:

- Ordena-se os valores-p observados: $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$
- Calcula-se os valores-p ajustados pelo método de BH: $p_{a(i)} = \frac{p_{(i)} * m}{i}$
- Compara-se os valores-p ajustados com o nível de rejeição de FDR - q. Os valores-p ajustados menores do que q rejeitam a hipótese nula que lhes corresponde.

4.3.2 Comportamento assintotico do método BH

Entre 1995 e 2001, as razões pelas quais o método apresentado por Benjamini e Hochberg tinha tanto sucesso eram desconhecidas, para além de se saber que controlava FDR. Só em 2001 é que Christopher Genovese e Larry Wasserman ([17]) descobriram que, para além de outras características, o método BH devolve um número de rejeições que se posiciona entre o número de rejeições usando valores-p não ajustados (método bruto) e usando valores-p ajustados pelo método de Bonferroni, desde que o número m de testes a realizar seja *suficientemente grande*. Esse resultado assintotico encontra-se esquematizado na Figura 4.1 que serve de base às considerações e raciocínios efetuados por Genovese e Wasserman.

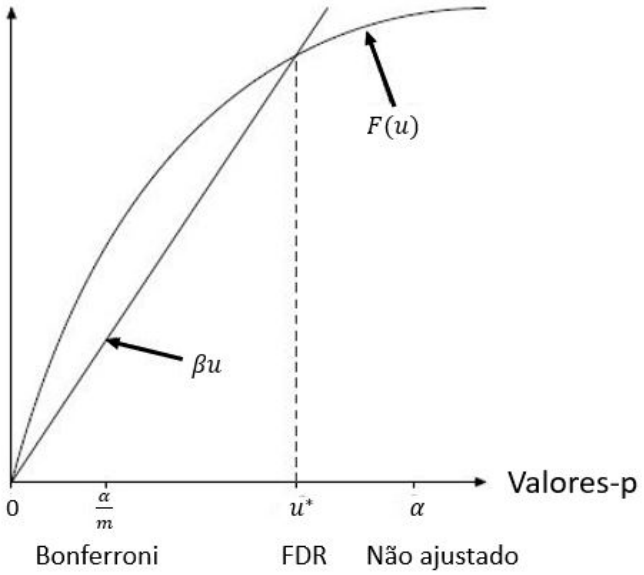


FIGURA 4.1: Interpretação geométrica do comportamento assintótico do procedimento BH. Adaptado de [17].

Na figura, F corresponde à função de distribuição dos valores-p sob a hipótese alternativa, u^* representa o limite crítico obtido pela implementação do método BH e β é o declive de uma reta a definir abaixo. Os autores assumem o cenário simples em que a distribuição dos valores-p sob as hipóteses alternativas é a mesma (iguais hipóteses alternativas, independentes de m).

Genovese e Wasserman provam que, assintoticamente, o método BH rejeita todos os valores-p menores ou iguais a u^* em que u^* satisfaz

$$F(u^*) = \beta u^*$$

com $\beta = \left(\frac{1}{\alpha} - A_0\right) / (1 - A_0)$, onde A_0 representa a proporção de hipóteses nulas verdadeiras. Apesar de u^* depender de F e A_0, os autores conseguem encontrar um estimador consistente para u^* que é independente de F (é *distribution free*); mais precisamente

$$D\frac{\alpha}{m} \xrightarrow{p} u^*.$$

onde $D = \max\left\{i : p_{(i)} < i\frac{\alpha}{m}\right\}$ é o "ponto de decisão"do método Benjamini Hochberg (ver 4.8). A prova deste resultado consiste em encontrar limites inferior a_m e superior b_m que satisfazem a condição $P\left(a_m \leq D \leq b_m\right) \underset{m\to+\infty}{\longrightarrow} 1$.

Mostram ainda que, para *m suficientemente grande* e condições bastante genéricas sobre F (estritamente côncava e $F'(0) > \beta$), se tem

$$\frac{\alpha}{m} \leq u^* \leq \alpha,$$

onde $\frac{\alpha}{m}$ corresponde ao valor crítico associado ao método de Bonferroni e α corresponde ao valor crítico para os valores-p não ajustados.

Os detalhes de todos estes resultados podem ser encontrados em [17].

Capítulo 5

Estudo Computacional

Nesta dissertação, foi desenvolvido um estudo computacional de simulação com o objetivo de estudar o comportamento dos erros de tipo I e de tipo II para diferentes números de hipóteses m, números de hipóteses verdadeiras m_0 e tamanhos do efeito. Para essa finalidade, foi criado um código, onde se estudam diferentes famílias de hipóteses. Cada família integra m pares de hipóteses para comparação de médias entre duas populações:

$$H_{0i} : \mu_{1i} = \mu_{2i}$$
$$H_{1i} : \mu_{1i} \neq \mu_{2i}$$

onde μ_{ji} designa a média da população $j = 1, 2$ incluída na hipótese $i = 1, ..., m$. Ambas as populações são normalmente distribuídas e independentes com variâncias iguais a 1.

Das m hipóteses nulas tem-se m_0 verdadeiras. Por esse motivo, geraram-se duas amostras independentes de tamanho 5 de distribuições normais com $\sigma^2 = 1$ cada para cada uma dessas hipóteses verdadeiras, com médias iguais a 0. As restantes m_1 hipóteses nulas são falsas e, portanto, para cada uma, geraram-se duas amostras também de populações normais e independentes com $\sigma^2 = 1$ de tamanho 5: uma com média 0 e outra com média igual a $\frac{5}{4}$, $\frac{2\times5}{4}$, $\frac{3\times5}{4}$ ou 5. Assim, o tamanho do efeito para as hipóteses nulas falsas varia entre $\frac{5}{4}$, $\frac{2\times5}{4}$, $\frac{3\times5}{4}$ ou 5. Aplicam-se testes-t, e vão ser analisadas as rejeições ou não rejeições das hipóteses nulas através dos valores-p brutos e dos valores-p ajustados pelo método de Bonferroni e pelo método de Benjamini-Hochberg. Nesta simulação, o nível de significância foi fixado em 0.05.

Primeiramente, na Seção 5.1, vai ser analisado o comportamento das não rejeições de H_{0i} quando esta é falsa, ou seja, os erros de tipo II. De seguida, na Seção 5.2 analisa-se o comportamento dos erros de tipo I, isto é, as rejeições de H_{0i} quando esta é verdadeira.

Para cada um destes erros, apresentam-se dois estudos distintos. O primeiro estudo analisa os erros mantendo a proporção de hipóteses nulas verdadeiras em $A_0 = \frac{m_0}{m} = \frac{1}{4}$ e variando o número de hipóteses; como número total de hipóteses considera-se $m = 16$, 32 e 64. O segundo estudo explora os erros mantendo o número de hipóteses total em $m = 64$ e variando a proporção de hipóteses nulas verdadeiras. As proporções consideradas foram $A_0 = \frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{4}$.

Esta simulação foi inspirada pelo trabalho desenvolvido por Benjamini e Hochberg [6], tendo sido extendida para a comparação de duas amostras, que é muito usado na prática. Os valores $A_0 = 0$ e $A_0 = 1$ não foram considerados nesta simulação, visto tratarem-se de situações extremas que podem ser consideradas num trabalho futuro.

Na Tabela 5.1 é apresentado o número de erros de tipo I e de tipo II (a verde e a vermelho, respetivamente), para uma simulação específica com $m = 64$ e $m_0 = \frac{m}{4} = 16$, com um tamanho do efeito igual a $\frac{2\times5}{4}$. Nenhum método foi aplicado, portanto as rejeições e não rejeições foram obtidas a partir dos valores-p brutos.

Hipótese nula	Não rejeitada	Rejeitada
Falsa	6	42
Verdadeira	16	0

TABELA 5.1: Tabela de erros de uma simulação com $m = 64$ e $m_0 = \frac{m}{4} = 16$.

Através desta tabela pode-se estimar os erros de tipo I e de tipo II. Realizando um estudo sistemático com 2000 simulações para cada método - bruto, Bonferroni e BH, e contando o número de erros obtidos individualmente para cada uma destas simulações, consegue-se estudar as distribuições desses erros.

Na Seção 5.3, será explorada a aplicação do fator de Bayes a testes de hipóteses múltiplas. O fator de Bayes é aplicado a cada hipótese individualmente para identificação das evidências a favor ou contra a hipótese nula. Tal como anteriormente, aqui também será feita uma variação no número de hipóteses e nos tamanhos do efeito, bem como no tamanho da amostra. Este método será comparado com dois dos métodos já considerados - bruto e BH.

Por fim, na Seção 5.4, um conjunto de dados reais da Genética será analisado por estes métodos.

A biblioteca *ggplot2* do R foi usada, na obtenção de todos os gráficos que serão apresentados neste capítulo.

5.1 Erros de tipo II

Lembra-se que num teste de hipóteses, um erro de tipo II corresponde a não rejeições (erradas) de hipóteses nulas falsas.

5.1.1 Aumento de m com proporção A_0 fixa

Analisando, nas 2000 simulações realizadas, os erros de tipo II em função do tamanho do efeito para os 3 métodos considerados, obtêm-se os histogramas apresentados na Figura 5.1, para $m = 16$ testes e um número de hipóteses nulas verdadeiras $m_0 = \frac{m}{4} = 4$.

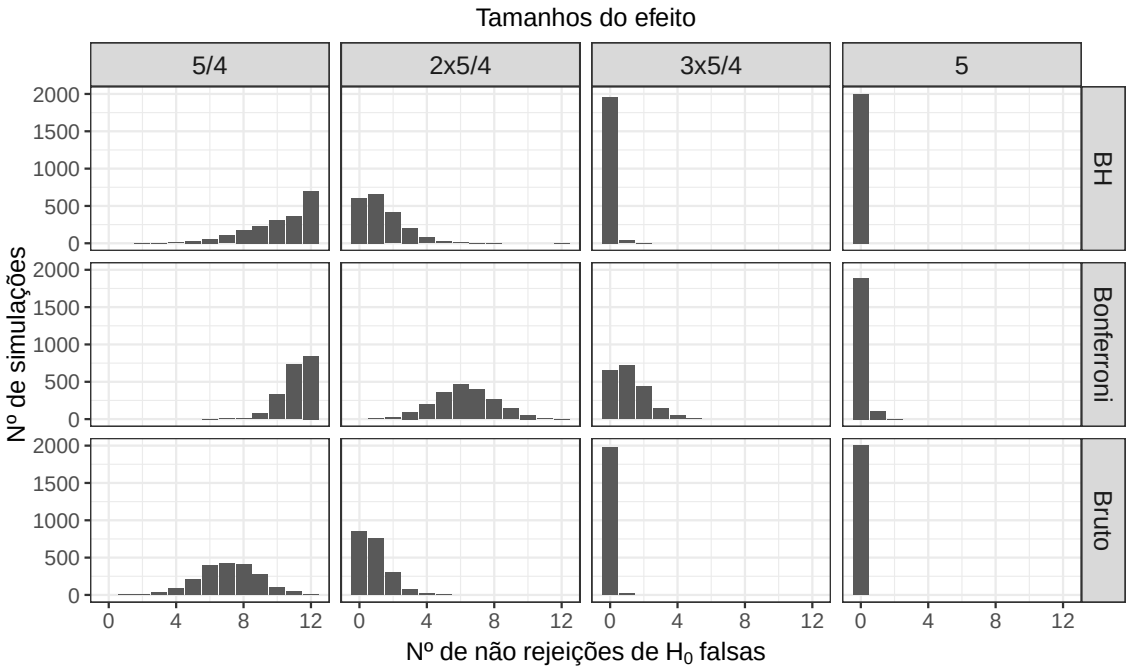


FIGURA 5.1: Distribuição dos erros de tipo II em 2000 simulações para $m = 16$ e $m_0 = 4$.

Pode-se observar que, à medida que o tamanho do efeito aumenta, o número de erros tende para 0 em todos os métodos estudados; para os tamanhos do efeito estudados, os 3 métodos resultaram numa potência próxima. Quando o tamanho do efeito é $\frac{5}{4}$ a distribuição dos erros está mais próxima de 0 no método bruto do que no método de BH, sendo o método de Bonferroni o pior nesta situação. Este resultado vai-se refletir numa maior potência para o caso bruto em relação ao BH, e numa maior potência deste em relação ao Bonferroni.

Na Figura 5.2 são apresentadas as distribuições do número de erros de tipo II para $m = 32$ testes com $\frac{1}{4} \times 32 = 8$ das hipóteses nulas verdadeiras.

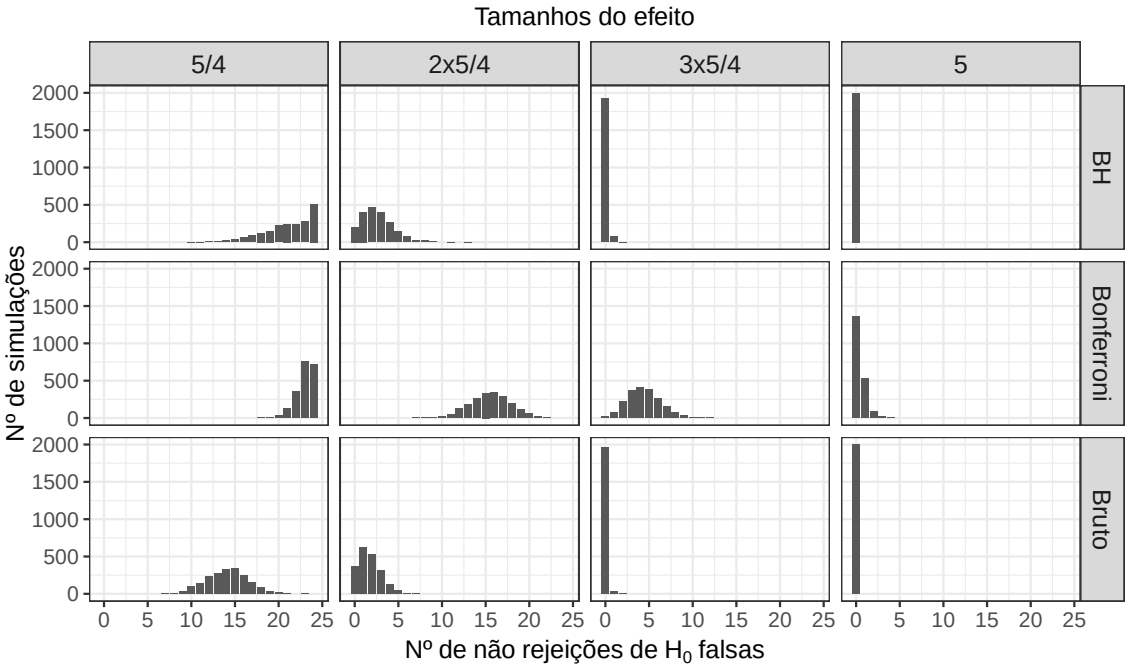


FIGURA 5.2: Distribuição dos erros de tipo II em 2000 simulações para $m = 32$ e $m_0 = 8$.

Tal como para $m = 16$, a distribuição dos erros de tipo II para $m = 32$ também se aproxima de 0 à medida que o tamanho do efeito aumenta. Contudo, para efeitos pequenos, a distribuição dos erros de tipo II está mais afastado de 0 na situação $m = 32$ testes. Isto é de esperar visto que o número de hipóteses falsas aumenta também - a proporção de hipóteses falsas é que se mantém constante. Este resultado influencia a distribuição dos erros de tipo II, não alterando contudo os resultados obtidos para a potência, dado que esta é calculada em função do número de hipóteses nulas falsas existentes.

Por fim, aumenta-se o número de hipóteses nulas para $m = 64$. Na Figura 5.3 são apresentadas as distribuições dos erros de tipo II para $m = 64$ em que $\frac{1}{4} \times 64 = 16$ hipóteses nulas são verdadeiras.

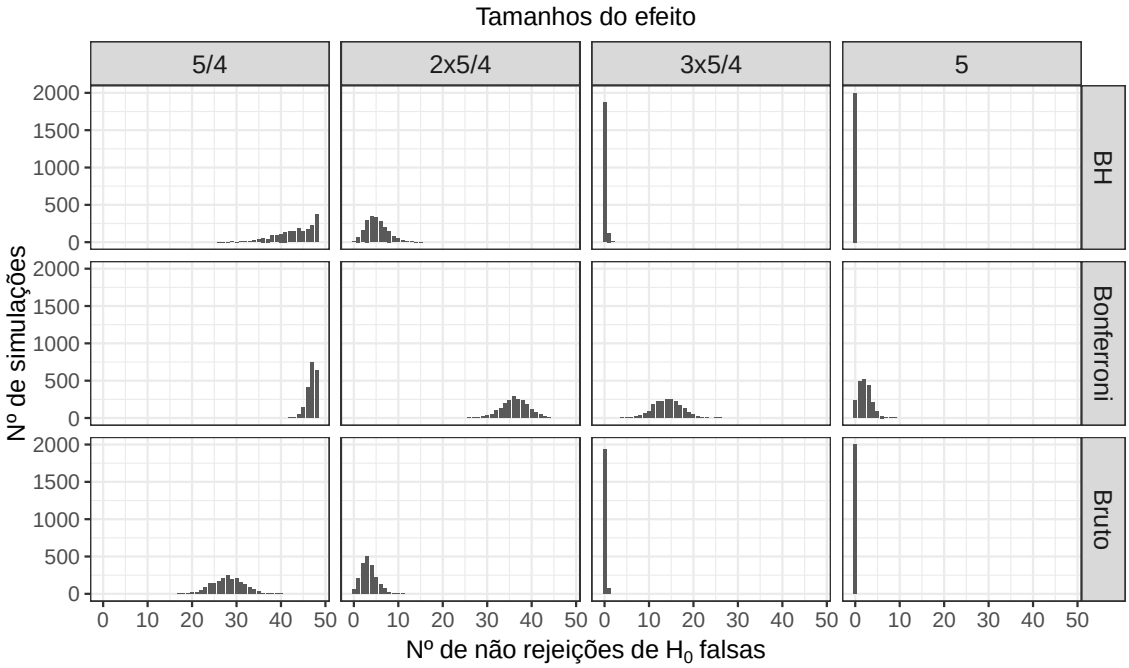


FIGURA 5.3: Distribuição dos erros de tipo II em 2000 simulações para $m = 64$ e $m_0 = 16$.

As conclusões retiradas são semelhantes às obtidas para $m = 16$ e $m = 32$, nomeadamente no que toca à relação entre a potência e o tamanho do efeito.

Na Figura 5.4 está representada a evolução da potência média nas 2000 simulações em função do tamanho do efeito, para um número total de hipóteses $m = 16$, 32 e 64.

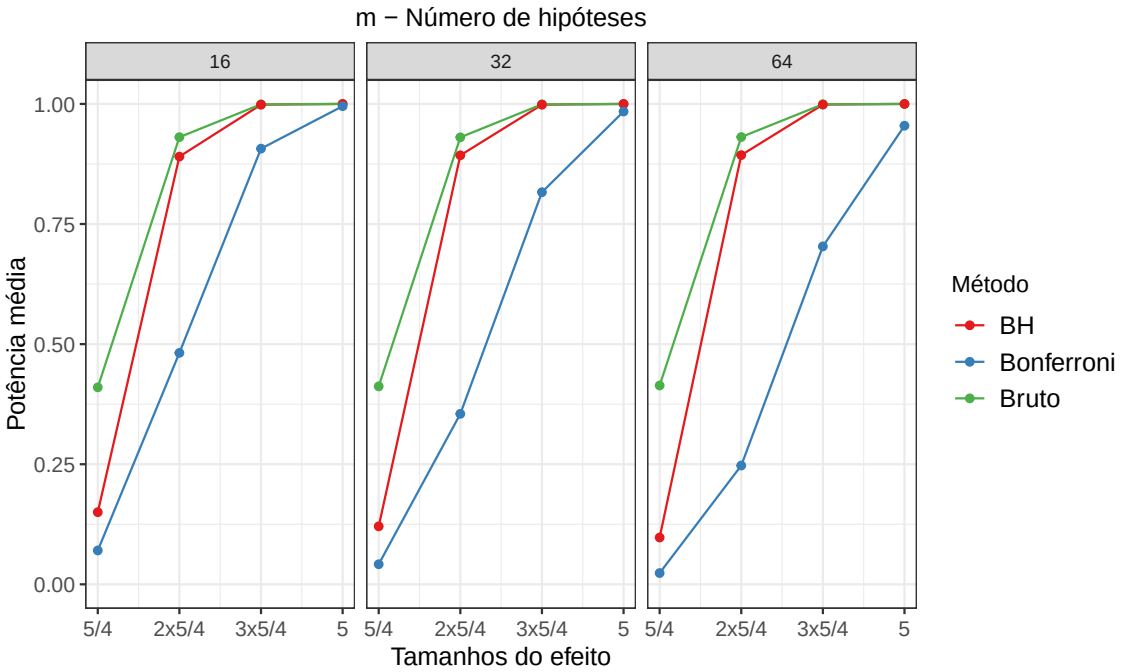


FIGURA 5.4: Potência média vs. tamanho do efeito para 2000 simulações com $m = 16$, 32 e 64.

Constata-se que o método bruto conduz a uma potência maior do que o método de BH quando o efeito é pequeno, mas essa diferença vai-se esbatendo à medida que o tamanho do efeito aumenta, chegando ao mesmo valor para tamanho do efeito $\geq \frac{3\times5}{4}$. Já o método de Bonferroni tem uma potência menor do que a dos métodos bruto e BH, para qualquer tamanho do efeito e número de hipóteses. Também se observa que a potência nos três métodos aumenta à medida que o tamanho do efeito se afasta de 0. Considerando o tamanho do efeito fixo, a potência decresce com o aumento do número de hipóteses nos métodos de BH e Bonferroni, enquanto que no método bruto permanece constante.

5.1.2 Aumento da proporção A_0 com m fixo

Considere-se agora o caso de $m = 64$ (fixo), e a proporção do número de hipóteses nulas verdadeiras m_0 a variar no conjunto $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$. O caso de $m_0 = \frac{1}{4} \times m = 16$ está já exposto na Figura 5.3. Os resultados obtidos para $m_0 = \frac{1}{2} \times m = 32$ e $m_0 = \frac{3}{4} \times m = 48$ podem ser vistos na Figura 5.5 e na Figura 5.6, respetivamente.

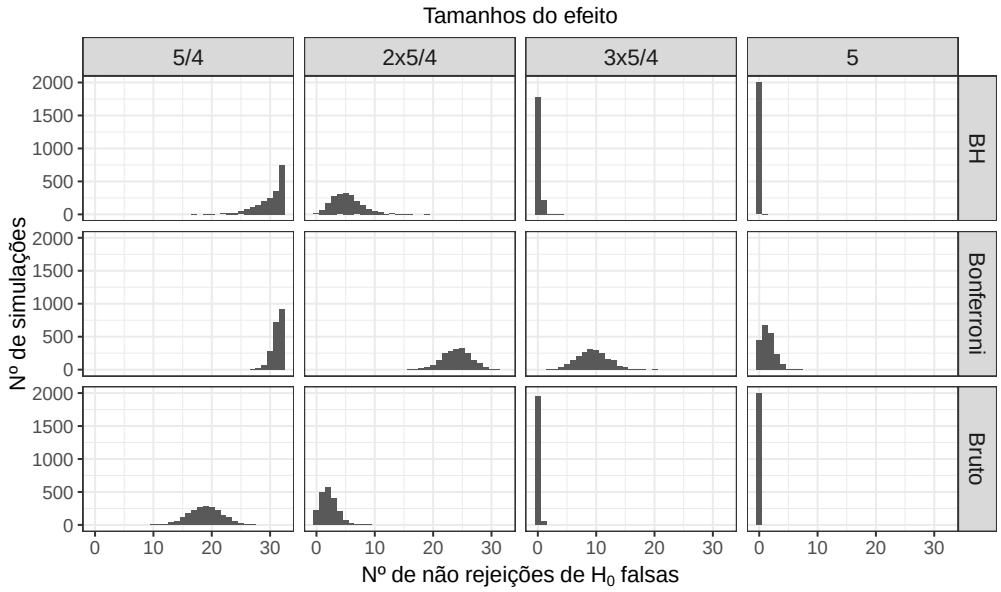


FIGURA 5.5: Distribuição dos erros de tipo II em 2000 simulações para $m = 64$ e $m_0 = 32$.

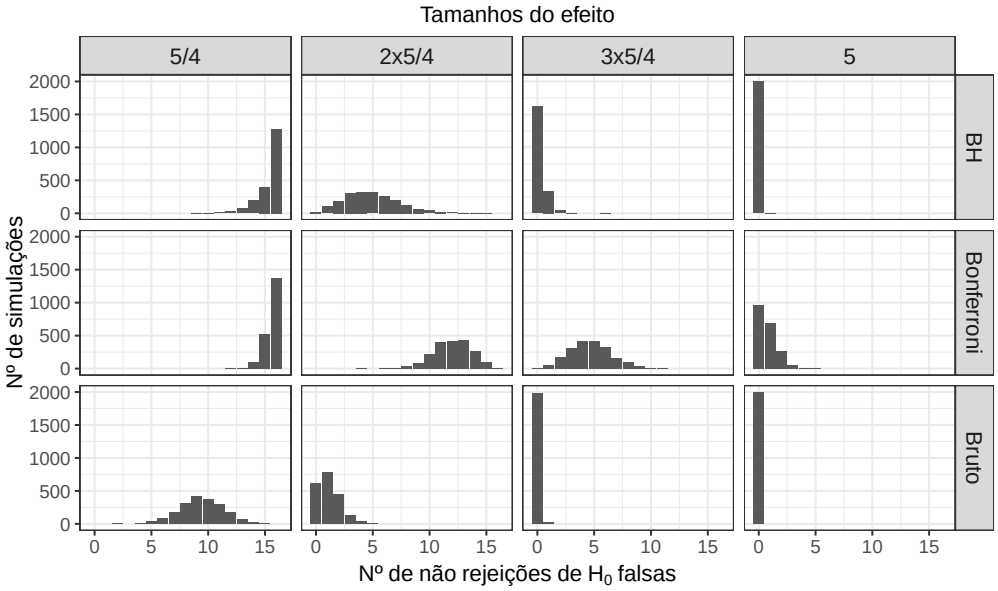


FIGURA 5.6: Distribuição dos erros de tipo II em 2000 simulações para $m = 64$ e $m_0 = 48$.

À medida que o número m_0 de hipóteses nulas verdadeiras aumenta, os erros de tipo II diminuem para os três métodos estudados. Este resultado é expectável, uma vez que o número $m - m_0$ de hipóteses nulas falsas diminui. Observa-se que o método de BH comete mais erros de tipo II (portanto tem menor potência) do que o método bruto, com a diferença a esbater-se à medida que o tamanho do efeito aumenta. Por seu lado, o método de Bonferroni, apresenta sempre mais erros de tipo II do qualquer um dos outros métodos, para qualquer tamanho do efeito considerado.

A Figura 5.7 apresenta a potência média para as 2000 simulações em função do tamanho do efeito, para diferentes valores de m_0.

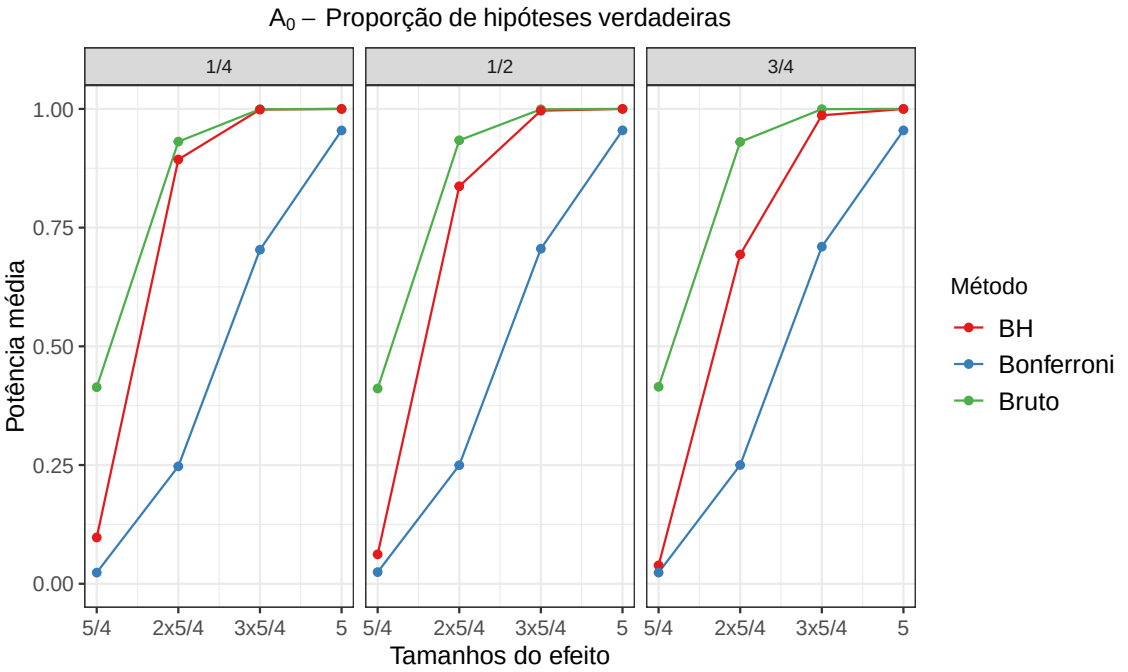


FIGURA 5.7: Potência média vs. tamanho do efeito para 2000 simulações com $m = 64$ e $m_0 = 16$, 32 e 48.

À medida que a proporção de hipóteses nulas verdadeiras aumenta e para um tamanho do efeito constante, a potência obtida pelo método de BH vai diminuindo aproximando-se da potência do método de Bonferroni; por sua vez, os métodos bruto e de Bonferroni mantêm a mesma potência. Tal como na Figura 5.4, a potência obtida pelo método de BH é mais baixa do que a obtida pelo método bruto para tamanhos do efeito reduzidos, mas superior à obtida pelo método de Bonferroni. Novamente, quando o tamanho do efeito se torna elevado, os métodos bruto e de BH tornam-se semelhantes em termos de potência.

5.2 Erros de tipo I

A análise feita na secção anterior foi também construída para os erros de tipo I, com as mesmas variações do número total de hipóteses nulas, de hipóteses nulas verdadeiras e de tamanhos do efeito. Relembra-se que um erro de tipo I corresponde a rejeições (erradas) de hipóteses nulas verdadeiras.

5.2.1 Aumento de m com proporção A_0 fixa

Para $m = 16$ hipóteses nulas testadas com $\frac{1}{4} \times 16 = 4$ verdadeiras, obteve-se o gráfico apresentado na Figura 5.8.

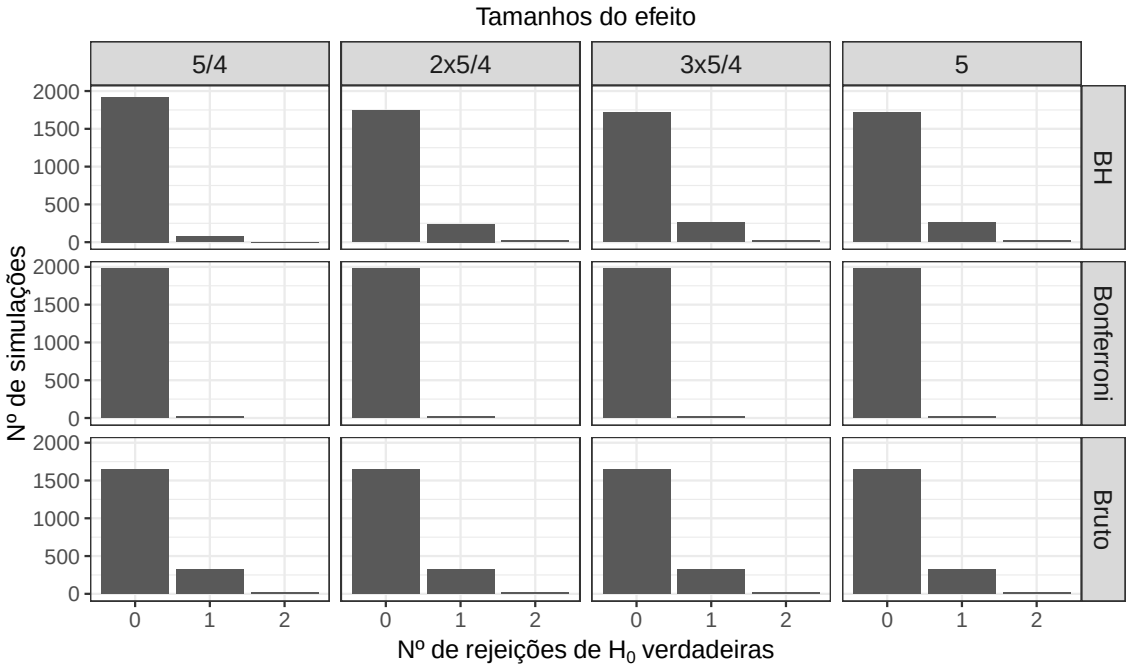


FIGURA 5.8: Distribuição dos erros de tipo I em 2000 simulações para $m = 16$ e $m_0 = 4$.

Os erros de tipo I ocorrem quando há rejeições erradas de H_0. É possível ver que, à medida que o tamanho do efeito aumenta, o número de rejeições de H_0 verdadeiras também aumenta, isto é, o número de testes que não apresentam rejeições de H_0 verdadeiras diminui. Será feito um estudo mais específico desta situação na Subseção 5.2.3.

Também é notório que o método de BH apresenta melhores resultados do que o método bruto no que diz respeito à rejeição de hipóteses verdadeiras. No entanto, o método de Bonferroni é o melhor método de entre os estudados no que toca ao número de erros de tipo I. O mesmo se pode concluir para $m = 32$ e para $m = 64$, com a mesma proporção de $\frac{1}{4}$ de hipóteses verdadeiras, como se pode confirmar pelas Figura 5.9 e Figura 5.10.

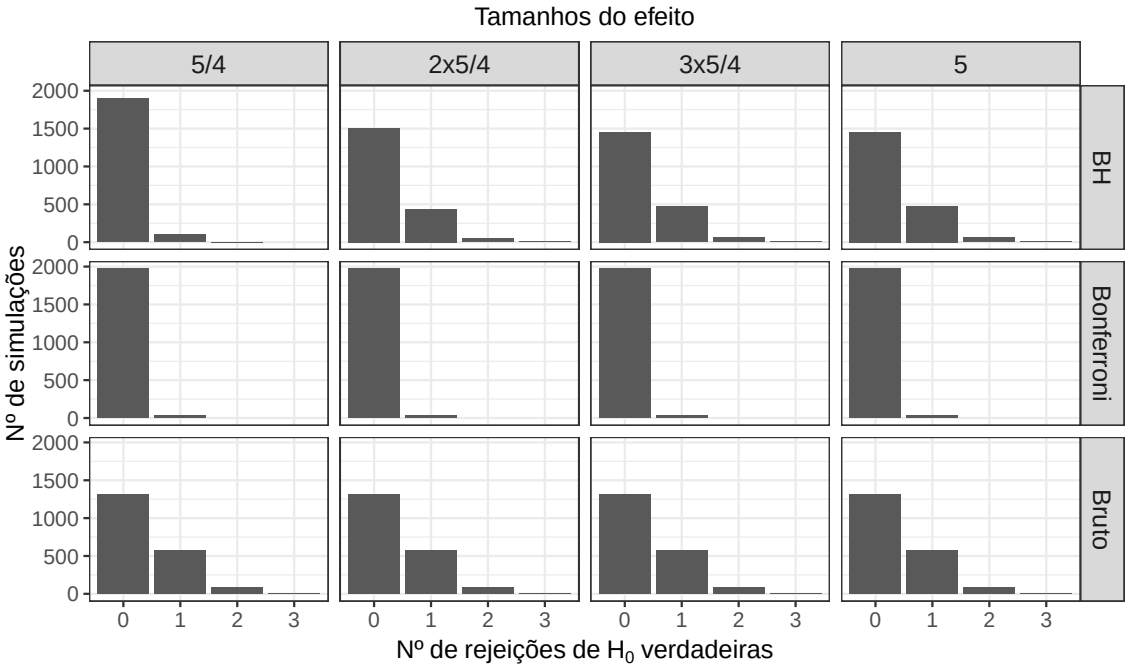


FIGURA 5.9: Distribuição dos erros de tipo I em 2000 simulações para $m = 32$ e $m_0 = 8$.

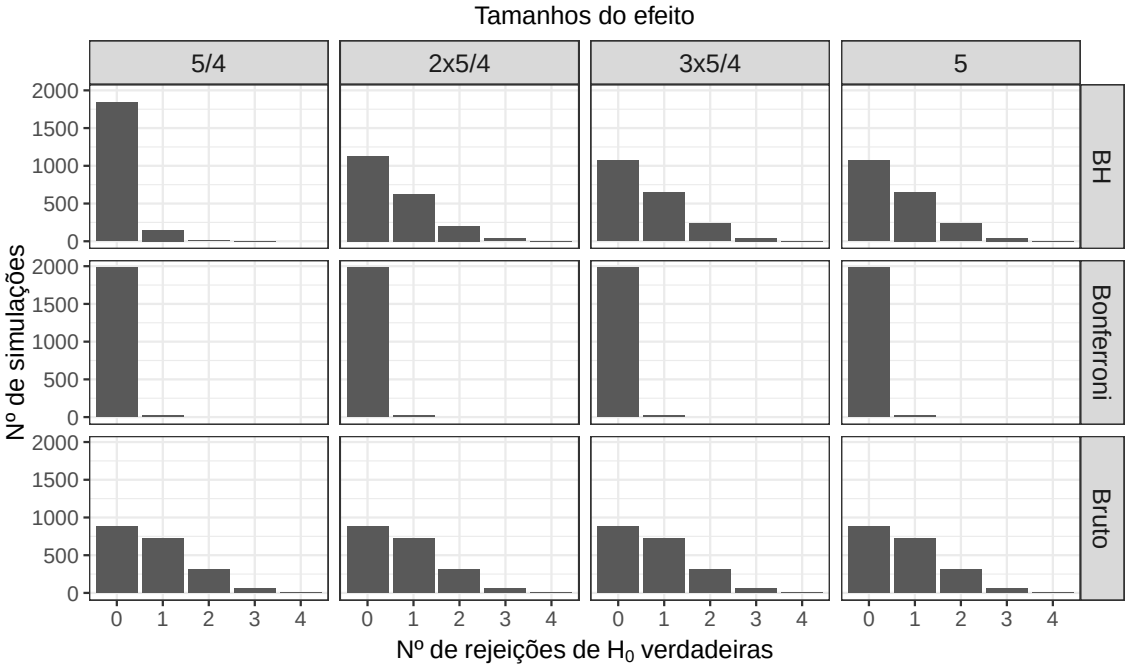


FIGURA 5.10: Distribuição dos erros de tipo I em 2000 simulações para $m = 64$ e $m_0 = 16$.

Tendo agora em consideração o aumento do número total de hipóteses, verifica-se que o número de rejeições de hipóteses verdadeiras aumenta. Isto acontece devido ao aumento do número de hipóteses verdadeiras m_0 em função de m, levando a uma maior possibilidade de se errar.

5.2.2 Aumento da proporção A_0 com m fixo

Por último, fixa-se o número de hipóteses em $m = 64$ e considera-se a proporção de hipóteses nulas verdadeiras como sendo $\frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{4}$ de m. Para esta situação, constata-se (Figura 5.8, Figura 5.11 e Figura 5.12) que o número de erros de tipo I aumenta em função do número de hipóteses nulas verdadeiras, pela razão apresentada atrás. Este resultado ocorre para os três métodos considerados.

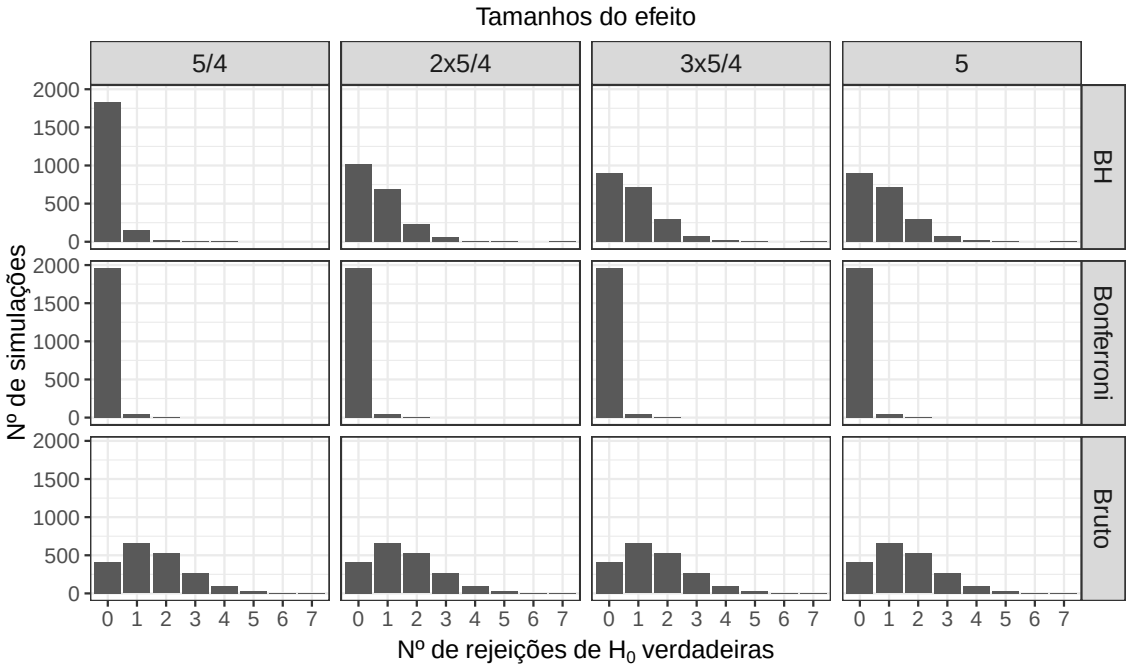


FIGURA 5.11: Distribuição dos erros de tipo I em 2000 simulações para $m = 64$ e $m_0 = 32$.

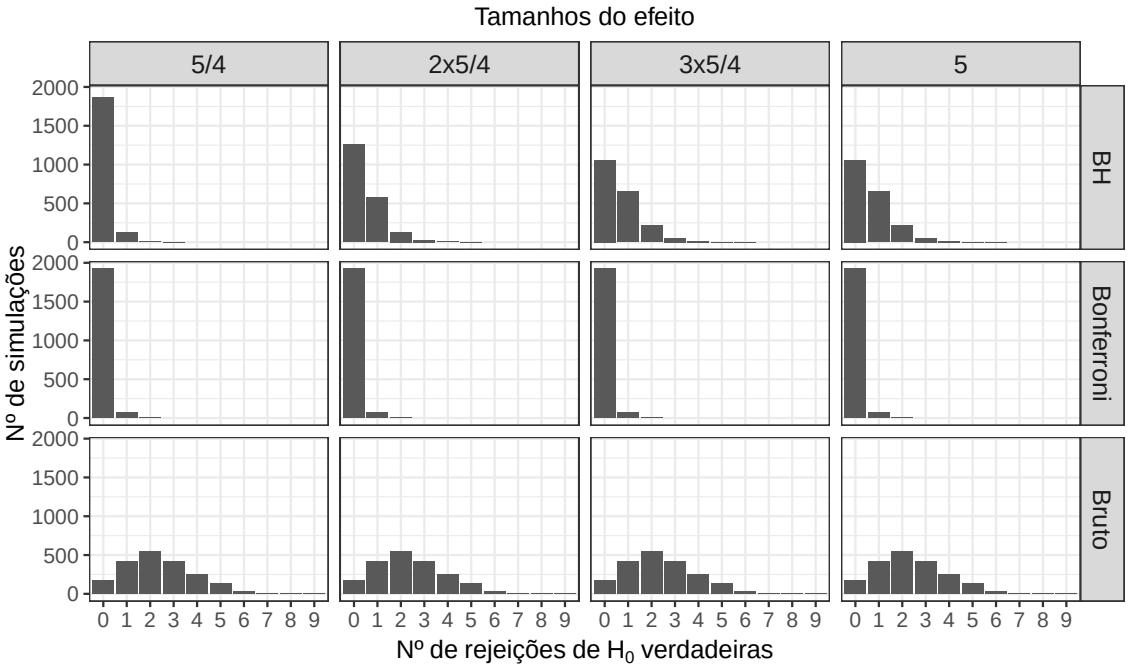


FIGURA 5.12: Distribuição dos erros de tipo I em 2000 simulações para $m = 64$ e $m_0 = 48$.

Relativamente aos erros de tipo I, observa-se que o método BH tem um melhor desempenho do que o método bruto. Esta ocorrência é de esperar porque o método BH controla o valor esperado da proporção das falsas rejeições, isto é, a proporção de erros cometidos pela falsa rejeição de hipóteses nulas. Contudo, o método de Bonferroni, por ser um método mais conservativo, e portanto, com menos rejeições, apresenta melhores resultados para este tipo de erros do que os métodos bruto e de BH.

Pode-se concluir que o método BH, comparativamente com o método bruto, reduz o número de erros de tipo I mas em consequência aumenta o número de erros de tipo II reduzindo, assim, a potência. Como o método BH rejeita menos hipóteses do que o método bruto, pois tem uma área de rejeição menor (pode ser observada no gráfico da Figura 5.13), então é natural que tenha menos erros de rejeição de hipóteses nulas quando estas são verdadeiras e mais erros de não rejeição de hipóteses nulas quando estas são falsas. O método de Bonferroni apresenta uma região de rejeição menor do que a do método de BH e, portanto, apresenta menores erros de tipo I e mais de erros de tipo II do que os métodos bruto e de BH.

Realça-se que o método bruto apresenta os melhores resultados para o número de erros de tipo II, e os piores resultados para o número de erros de tipo I. Contrariamente, o método de Bonferroni é o melhor método para diminuir os erros de tipo I, mas o pior no que toca a erros de tipo II. Assim, conclui-se que o método de BH é um método mais

moderado, que equilibra o número de erros de tipo II e de tipo I, o que pode ser justificado pelo facto deste método controlar não apenas a probabilidade de se cometer pelo menos uma falsa rejeição - o que acontece no método de Bonferroni - mas também o valor esperado da proporção de falsas rejeições.

5.2.3 Distribuição dos valores-p

Para melhor se compreender a situação referida na Subseção 5.2.1, foi feita uma simulação com 10 experiências individuais. Considera-se uma família de hipóteses que compara as médias de duas populações:

$$H_{0i} : \mu_{1i} = \mu_{2i}$$
$$H_{1i} : \mu_{1i} \neq \mu_{2i}$$

onde μ_{ji} designa a média da variável $i = 1, ..., m$ na população $j = 1, 2$. Para cada variável, as populações são normalmente distribuídas e independentes com variâncias iguais a 1.

Para cada uma das m_0 hipóteses nulas verdadeiras, geram-se duas amostras de tamanho 5 com médias iguais a 0. Para cada uma das m_1 hipóteses nulas falsas geram-se duas amostras de tamanho 5 uma com média igual a 0 e a outra com média igual a $\frac{5}{10}$, $\frac{5}{4}$, $\frac{2\times5}{4}$, $\frac{3\times5}{4}$ ou 5. Assumem-se populações normalmente distribuídas e independentes, com igual variância (= 1). Após aplicar o teste-t a cada par de hipóteses H_{0i} e H_{1i} para $i = 1, ..., m$, retiram-se os valores-p brutos associados. De seguida, ordenam-se esses valores-p e representam-se num gráfico. Esta simulação é repetida 10 vezes.

O estudo inclui 16, 64 ou 1000 hipóteses, com 5 efeitos de tamanho diferente ($\frac{5}{10}$, $\frac{5}{4}$, $\frac{2\times5}{4}$, $\frac{3\times5}{4}$, 5) em cada situação. Salienta-se que a amostra correspondente às hipóteses nulas verdadeiras mantém-se ao longo das linhas, isto é, os valores-p correspondentes às hipóteses nulas verdadeiras são os mesmos quando se aumenta o tamanho do efeito das hipóteses nulas falsas.

Atente-se, por exemplo, na coluna 2 da Figura 5.13. Quando o valor-p bruto de uma hipótese nula verdadeira é muito baixo, para um efeito de $\frac{5}{10}$, esse valor-p não é rejeitado pelo método ajustado de Benjamini e Hochberg - reta a vermelho. No entanto, à medida que se aumenta o tamanho do efeito, os valores-p correspondentes às hipóteses nulas falsas diminuem e portanto começam a ocupar os índices mais baixos, como seria de esperar. Por isso, o valor-p da hipótese nula verdadeira que antes não era rejeitado vai aumentar em ordem, alcançando, assim, a zona da rejeição de BH. Quanto ao método bruto, dado

que a zona de rejeição é igual em todos os índices - reta a azul - as hipóteses nulas verdadeiras com baixo valor-p que não são rejeitadas para tamanhos do efeito reduzidos vão permanecer não rejeitados quando se aumenta o tamanho do efeito. O mesmo acontece com o método de Bonferroni - reta a verde. No entanto, como este tem uma zona de rejeição muito menor relativamente aos outros dois métodos, tem menos rejeições, como já se tinha referido anteriormente.

Na Figura 5.14 é apresentado um gráfico em moldes semelhantes aos descritos anteriormente, alterando apenas o número de hipóteses para $m = 64$. E na Figura 5.15 para $m = 1000$. As conclusões retiradas são essencialmente independentes do valor de m.

Para todos os números de hipóteses estudados ($m = 16$, $m = 64$ e $m = 1000$), também se observa que o número de hipóteses nulas falsas rejeitadas aumenta à medida que o tamanho do efeito se afasta de 0, conseguindo-se atingir a rejeição de todas as hipóteses nulas falsas para o caso em que o efeito é igual a 5, ou seja, para um tamanho do efeito "elevado". O mesmo se verifica para o número de hipóteses nulas verdadeiras, embora em números menores. Para além disso, quando o tamanho do efeito é muito próximo de 0, o número de rejeições é muito baixo.

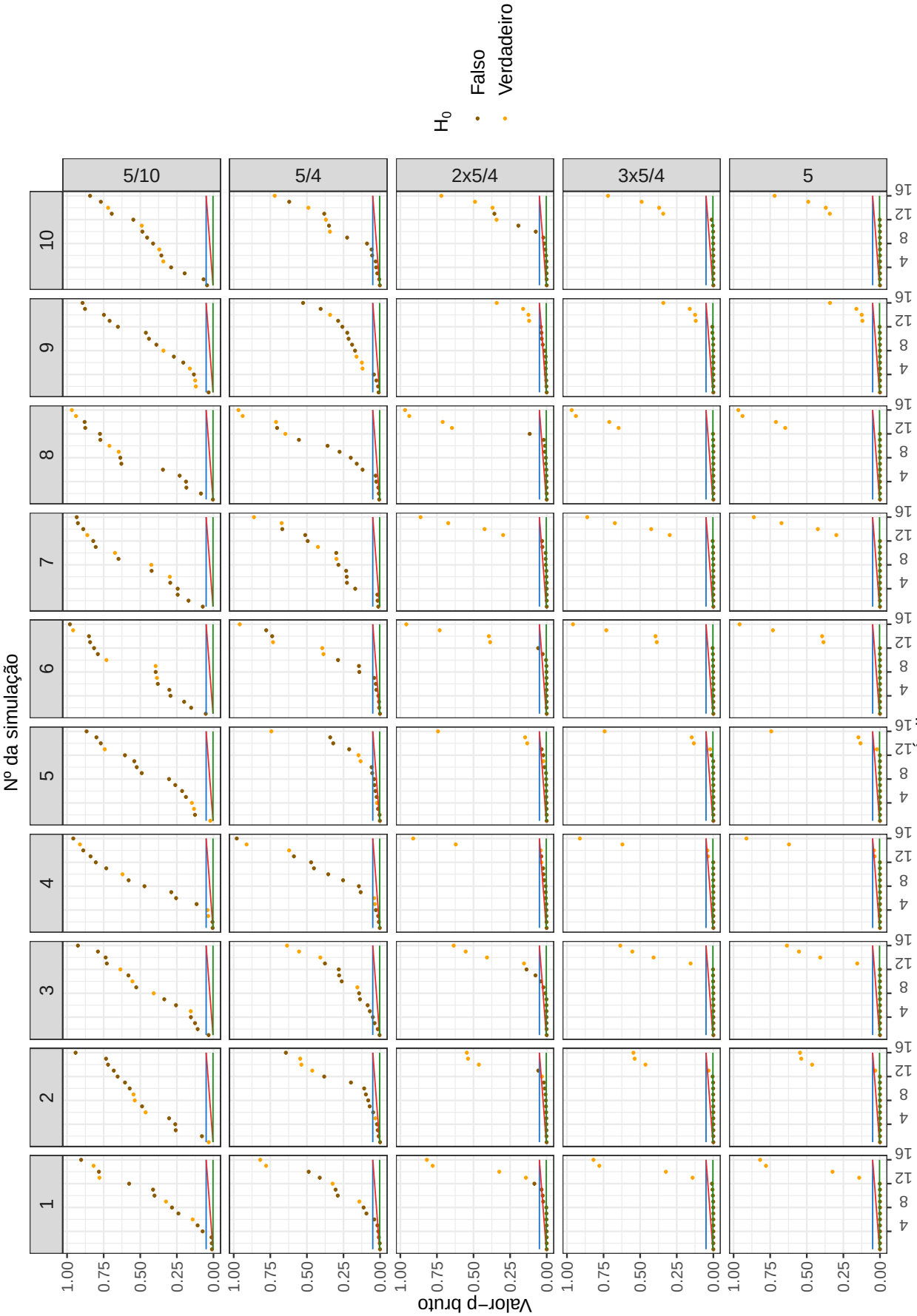


FIGURA 5.13: Distribuição dos valores-p em 10 simulações para $m = 16$ e as retas de corte dos métodos bruto (reta azul), de BH (reta vermelha) e de Bonferroni (reta verde).

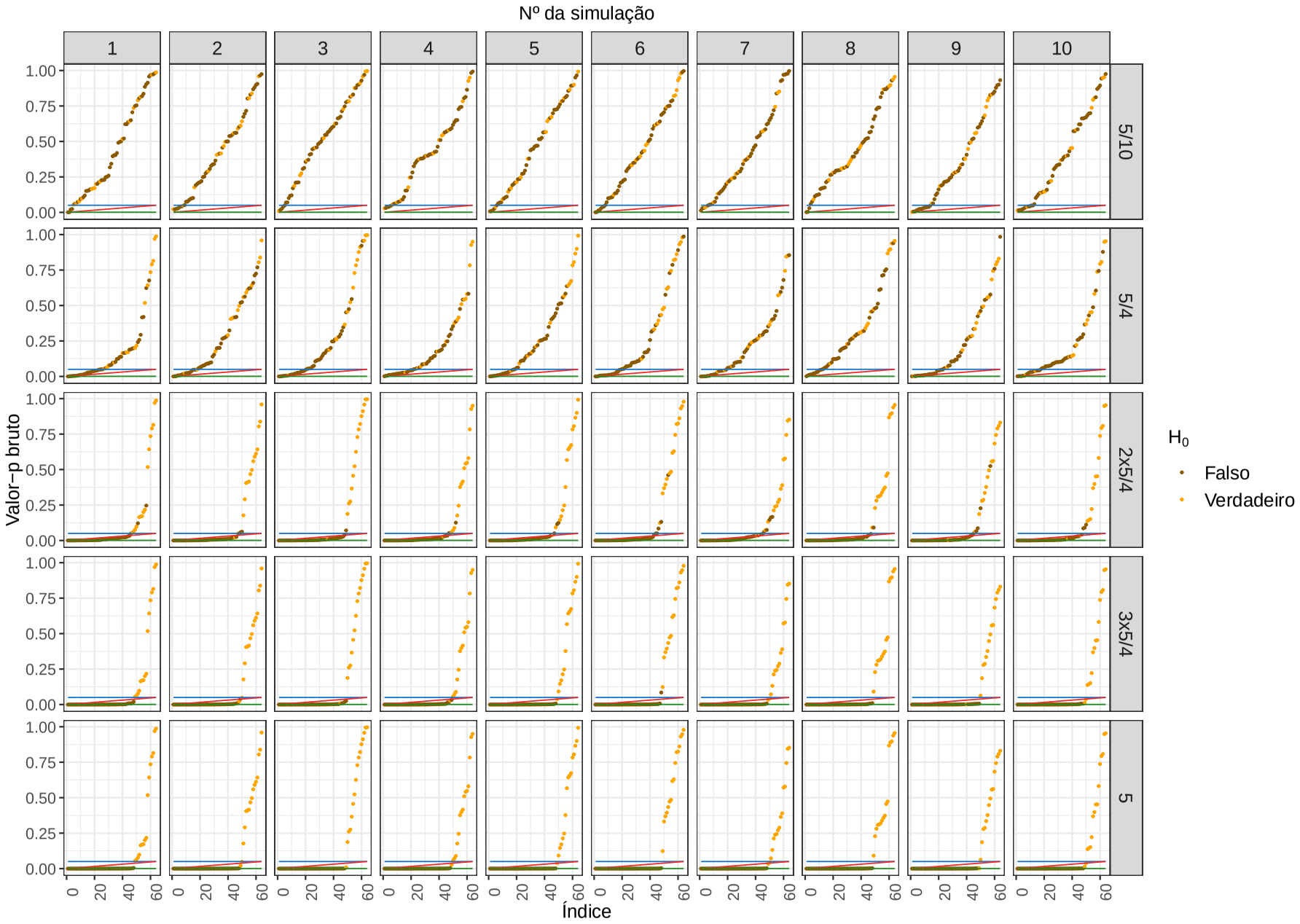


FIGURA 5.14: Distribuição dos valores-p em 10 simulações para $m = 64$ e as retas de corte dos métodos bruto (reta azul), de BH (reta vermelha) e de Bonferroni (reta verde).

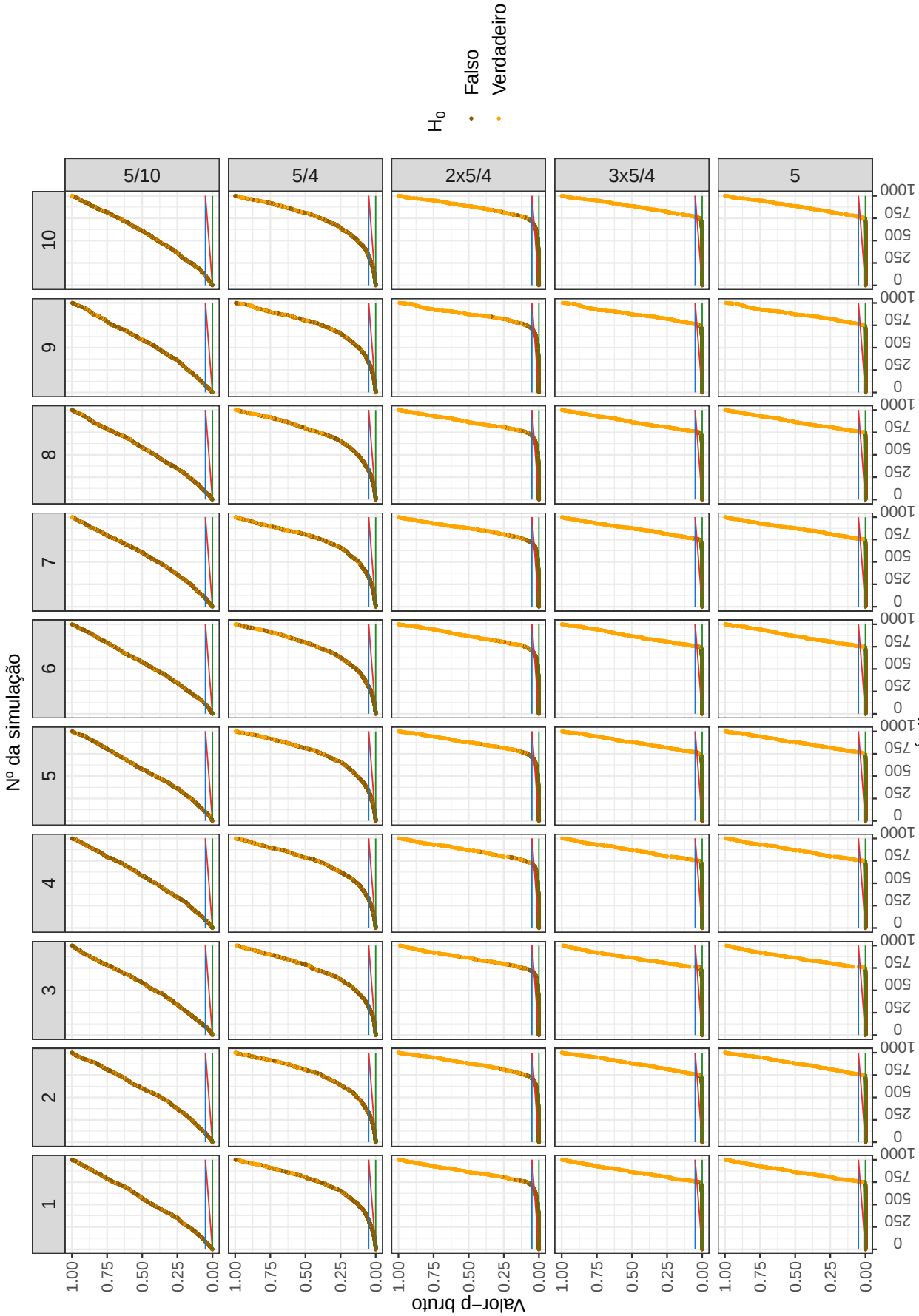


FIGURA 5.15: Distribuição dos valores-p em 10 simulações para $m = 1000$ e as retas de corte dos métodos bruto (reta azul), BH (reta vermelha) e de Bonferroni (reta verde).

5.3 Fator de Bayes em testes de hipóteses múltiplas

Foi aplicado o fator de Bayes a uma família de hipóteses, testando m igualdades de médias de duas populações distintas. Considere-se uma família de hipóteses que compara as médias entre duas populações:

$$H_{0i} : \mu_{1i} = \mu_{2i}$$
$$H_{1i} : \mu_{1i} \neq \mu_{2i}$$

onde μ_{ji} designa a média da variável $i = 1, ..., m$ na população $j = 1, 2$. Ambas as populações são normalmente distribuídas e independentes com variâncias iguais a 1. O estudo foi realizado aplicando o teste-t bayesiano a cada par de hipóteses – usando o pacote *BayesFactor* no R - de onde se obtém o fator de Bayes. Aplicando da mesma forma o teste-t nestas hipóteses, obtiveram-se os valores-p brutos. Os valores-p de BH obtiveram-se por aplicação do método aos valores-p brutos. O propósito da experiência consiste em comparar as evidências adquiridas pelo fator de bayes com as rejeições ou não rejeições das hipóteses obtidas pelos valores-p, brutos e ajustados por BH, fazendo variar o tamanho de cada uma das amostras (5 e 50), o número de hipóteses a testar (16 e 64) e o tamanho do efeito ($\frac{5}{10}$, $\frac{5}{4}$, $\frac{2\times5}{4}$, $\frac{3\times5}{4}$ e 5).

Para cada uma das m_0 hipóteses nulas verdadeiras, geram-se duas amostras de tamanho 5 ou 50 com médias iguais que tomam o valor 0. Para cada uma das m_1 hipóteses falsas gera-se duas amostras de tamanho 5 ou 50 em que uma tem média igual a 0 e a outra média igual a $\frac{5}{10}$, $\frac{5}{4}$, $\frac{2\times5}{4}$, $\frac{3\times5}{4}$ ou 5. As amostras são geradas a partir de populações normalmente distribuídas, independentes, e com igual variância ($= 1$).

Nas seguintes figuras estão representados o fator de Bayes BF e o valor-p correspondentes a cada teste de hipóteses na família de testes, tamanhos de amostra e número de hipóteses referidos anteriormente. Os valores de BF maiores do que 10 não estão representados nos gráficos, pois são demasiado grandes por comparação com os mais pequenos, e não permitem perceber o que se passa com estes últimos. No entanto, todos os pontos que não se encontrem reproduzidos nos gráficos são referentes a hipóteses nulas falsas com valor-p muito baixo. Portanto, são pontos relativos a hipóteses nulas falsas que são rejeitados pelo método bruto e pelo método de BH e, para além disso, com evidências fortes a favor de H_{1i}.

Ao analisar o gráfico da Figura 5.16, relativo a 16 testes e com amostras de tamanho 5, observa-se que, quando o tamanho do efeito é muito baixo, há vários valores de BF sem

evidências suficientes a favor da rejeição ou não rejeição. No entanto, à medida que se aumenta o tamanho do efeito; (1) os valores de BF correspondentes às hipóteses nulas falsas aumentam, deslocando-se da zona "Não há evidências"para a zona "Há evidências fortes a favor de H_1"; (2) os valores de BF correspondentes às hipóteses nulas verdadeiras mantêm-se entre $1/3$ e 3, continuando a não haver evidências a favor de nenhuma das hipóteses. Para além disso, é notório que, para efeitos baixos, há hipóteses nulas falsas que são rejeitadas pelo método bruto, mas não rejeitadas pelo BF, o que vai de encontro ao que foi observado na Figura 2.2 (b), pois o BF exige efeitos maiores para considerar evidências a favor de H_1.

Quando se aumenta o número de hipóteses de 16 para 64 - Figura 5.16 e Figura 5.17, respetivamente - são retiradas conclusões análogas. Pelo que se pode concluir que, o BF não é influenciado pelo aumento de testes a realizar.

Alterando o tamanho amostral de 5 para 50 - Figura 5.16 e Figura 5.18 - pode-se ver que o BF já corresponde a evidências moderadas a favor de H_0 para as hipóteses nulas verdadeiras e evidências fortes a favor de H_1 para as hipóteses nulas falsas, pois para um efeito igual a $\frac{5}{4}$, o BF correspondente a todas as hipóteses nulas falsas é já muito maior do que 10. O mesmo se pode concluir quando o número total de hipóteses m é igual a 64 - Figura 5.17 e Figura 5.19.

A existência de hipóteses nulas verdadeiras na zona sem evidências ou até mesmo de evidências a favor da H_1, na Figura 5.19, pode acontecer. Na primeira coluna da Figura 2.2 (b) observa-se que, mesmo com um tamanho de amostra grande, o método não é totalmente eficaz na deteção de hipóteses nulas verdadeiras, existindo alguma percentagem destas hipóteses que tenham resultado em ausência de evidência - zona branca - ou em evidências a favor da H_1 - 1% a vermelho.

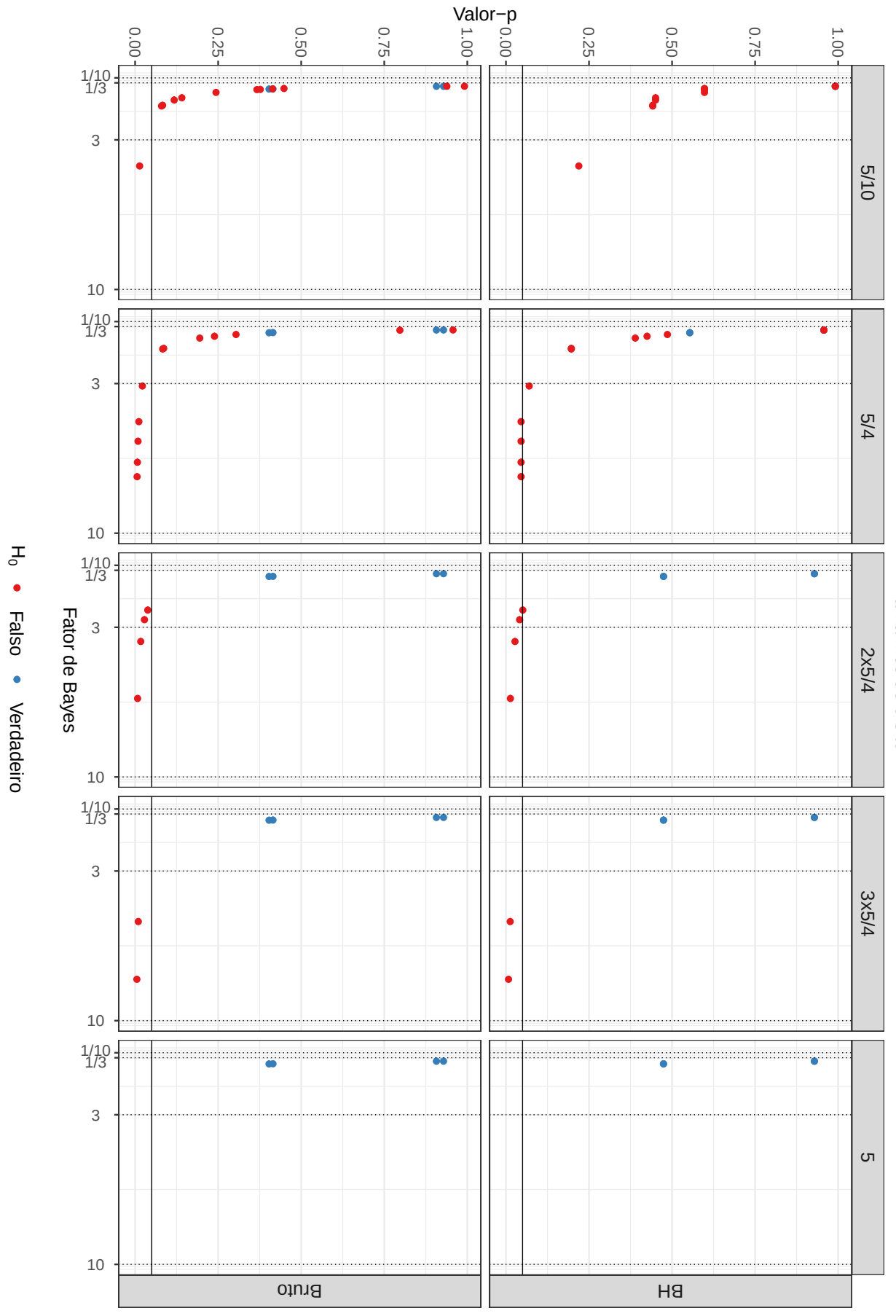


FIGURA 5.16: Fator de Bayes num teste de 16 hipóteses com amostras de tamanho 5

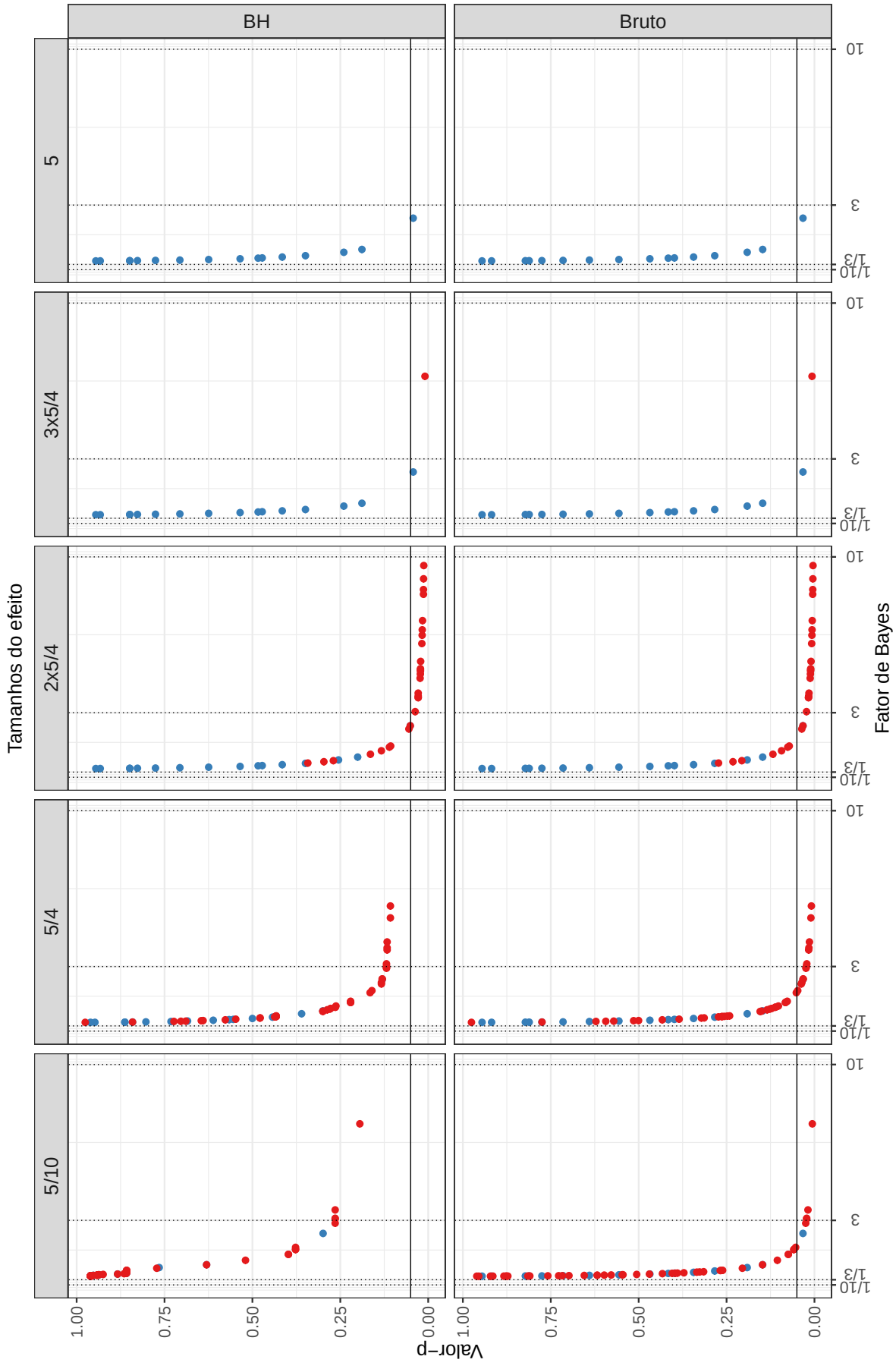


FIGURA 5.17: Fator de Bayes num teste de 64 hipóteses com amostras de tamanho 5

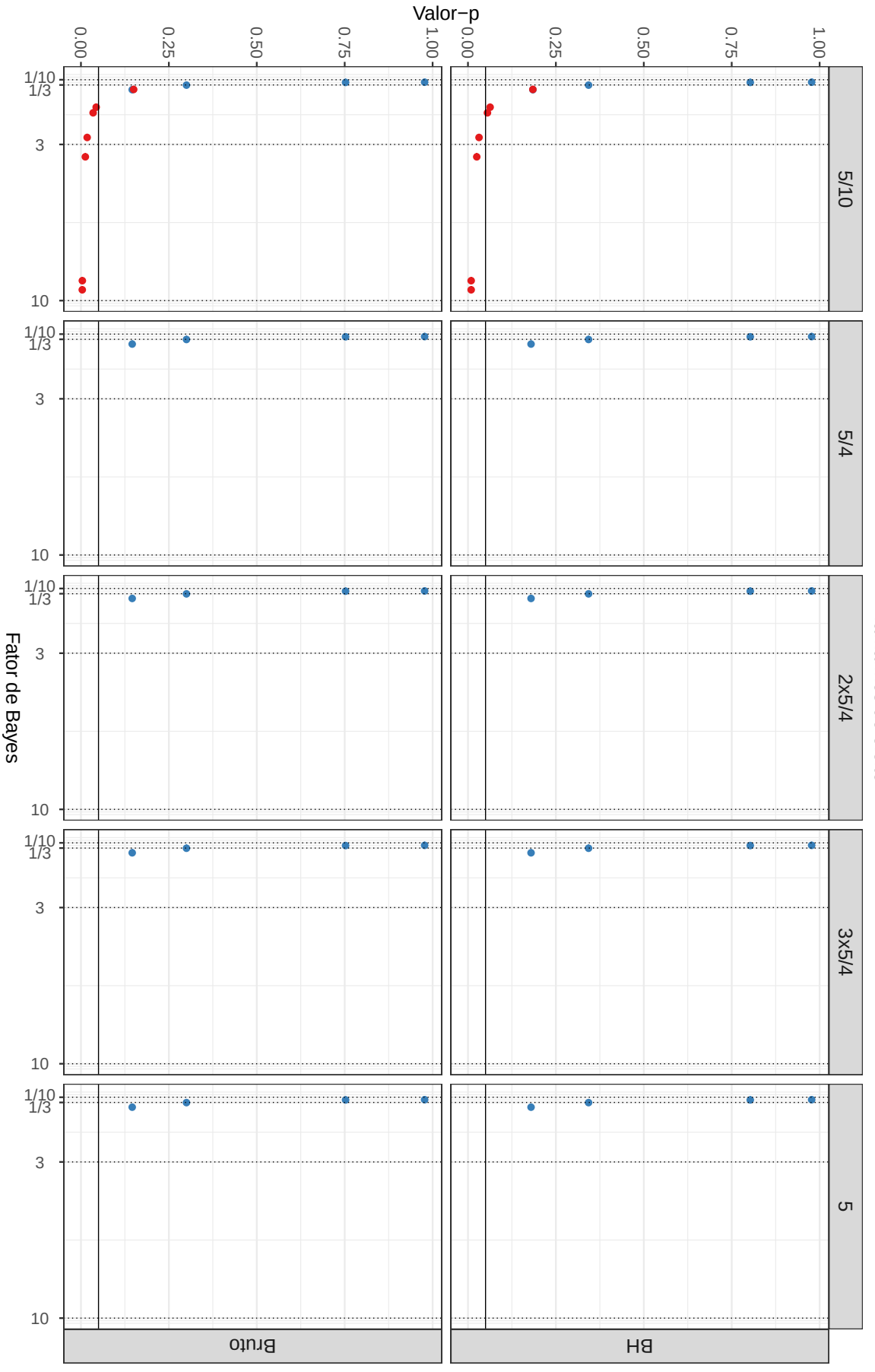


FIGURA 5.18: Fator de Bayes num teste de 16 hipóteses com amostras de tamanho 50

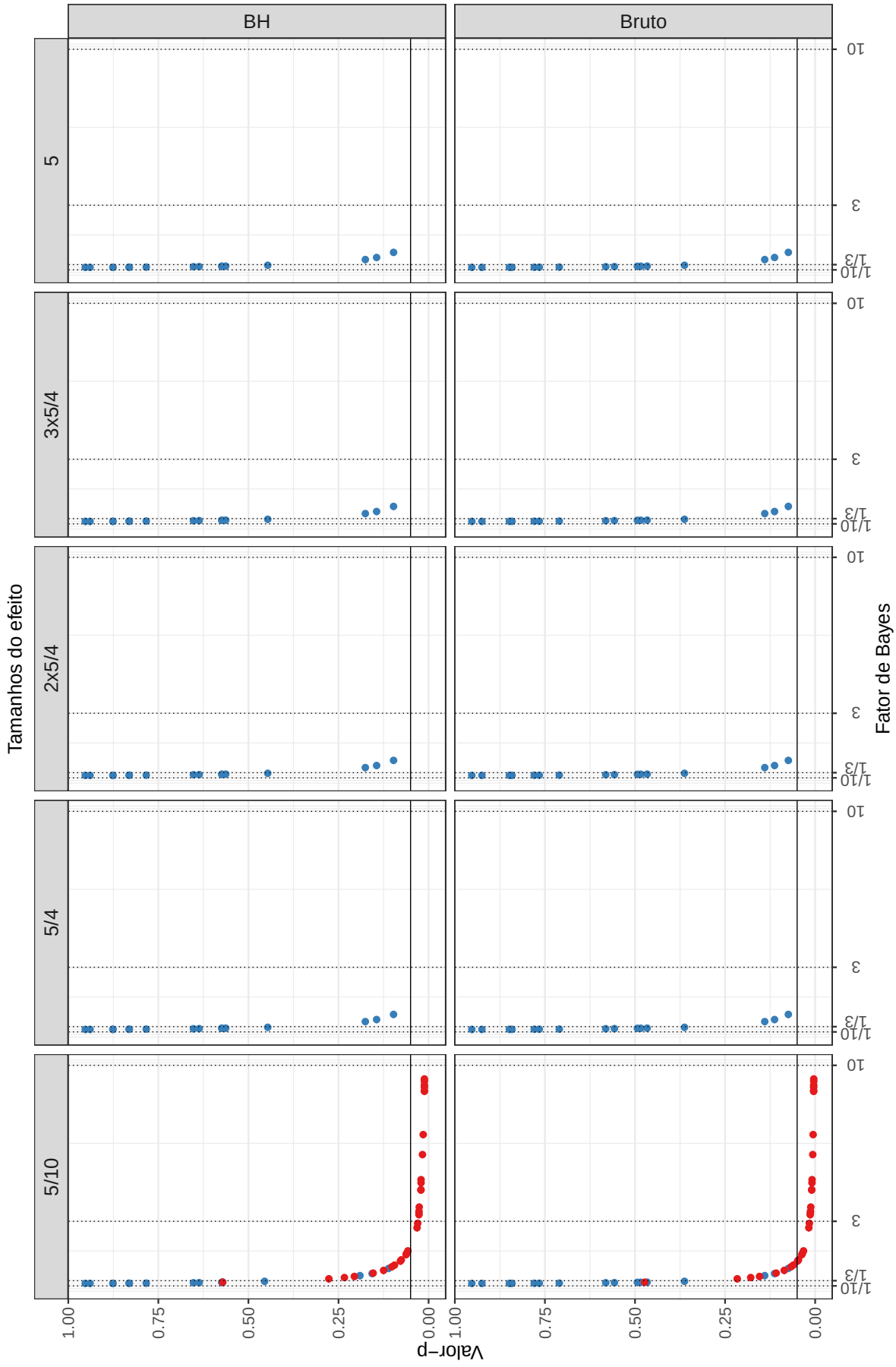


FIGURA 5.19: Fator de Bayes num teste de 64 hipóteses com amostras de tamanho 50

5.4 Estudo num conjunto de dados real

O estudo que a seguir se apresenta foi realizado a partir de uma base de dados retirada do site: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5583. O conjunto de dados consiste em expressões genéticas obtidas de *microarrays* de dois tipos de camundongos, selvagens e geneticamente modificados com HDAC[1], originários da mesma ninhada. Para cada um dos 12488 genes, existem dados de 3 animais selvagens e de 3 animais geneticamente modificados.

A Figura 5.20 compara as expressões genéticas médias de cada gene entre os dois tipos de camundongo. A reta vermelha apresentada é a reta $y = x$, portanto, pontos próximos dessa linha mostram expressões genéticas médias próximas.

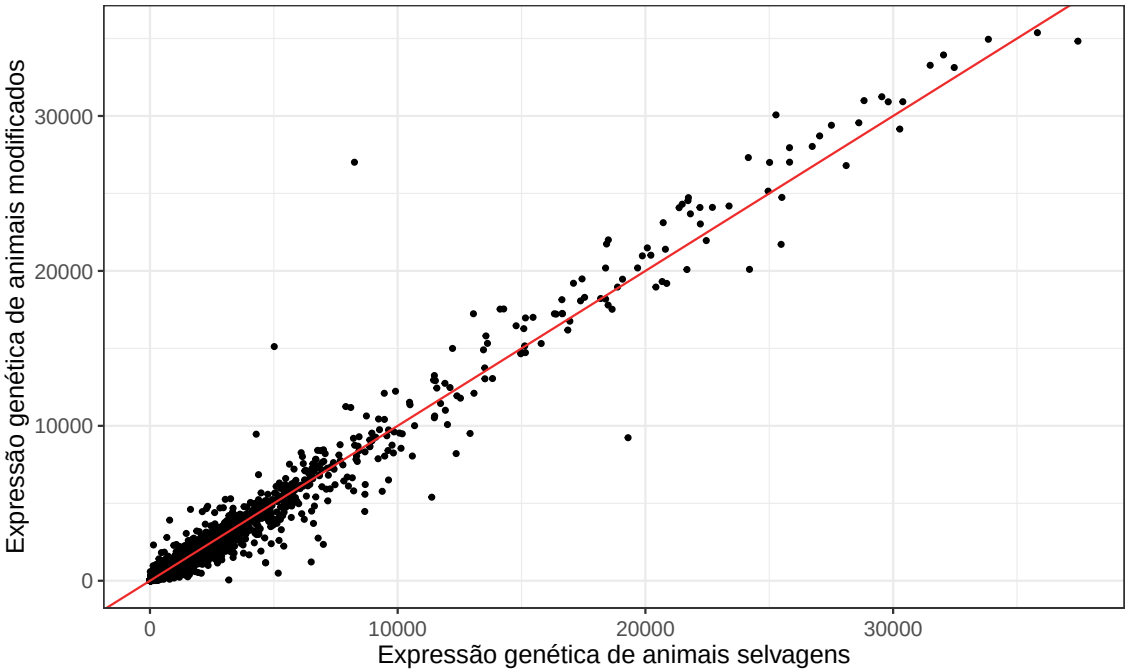


FIGURA 5.20: Gráfico de dispersão entres as expressões genéticas de animais modificados e de animais selvagens

As expressões genéticas médias são relativamente próximas entre as duas amostras de camundongo, existindo no entanto alguns genes que se conseguem diferenciar na sua expressão entre as duas amostras. Serão esses os genes que irão ser rejeitados pelos métodos bruto e de BH?

O objetivo do que se segue é exatamente a identificação dos genes que têm expressões médias diferentes para os dois tipos de camundongos. Aplica-se um teste bruto de hipóteses

[1] do inglês *Histone deacetylase*

múltiplas para comparação de expressões médias entre camundongos selvagens e camundongos modificados pelo HDAC. Os valores-p BH são obtidos por aplicação do método aos valores-p brutos

Para cada teste i, i = 1, . . . , 12488, as hipóteses consideradas foram

$$H_{0i} : \mu_{Selv_i} = \mu_{HDAC_i}$$

$$H_{1i} : \mu_{Selv_i} \neq \mu_{HDAC_i}$$

Os histogramas dos valores-p brutos e ajustados pelo método BH são apresentados na Figura 5.21. Como seria de esperar, o método BH exibe um maior número de genes correspondentes a valores-p mais altos do que o método bruto, rejeitando desse modo um menor número de hipóteses nulas.

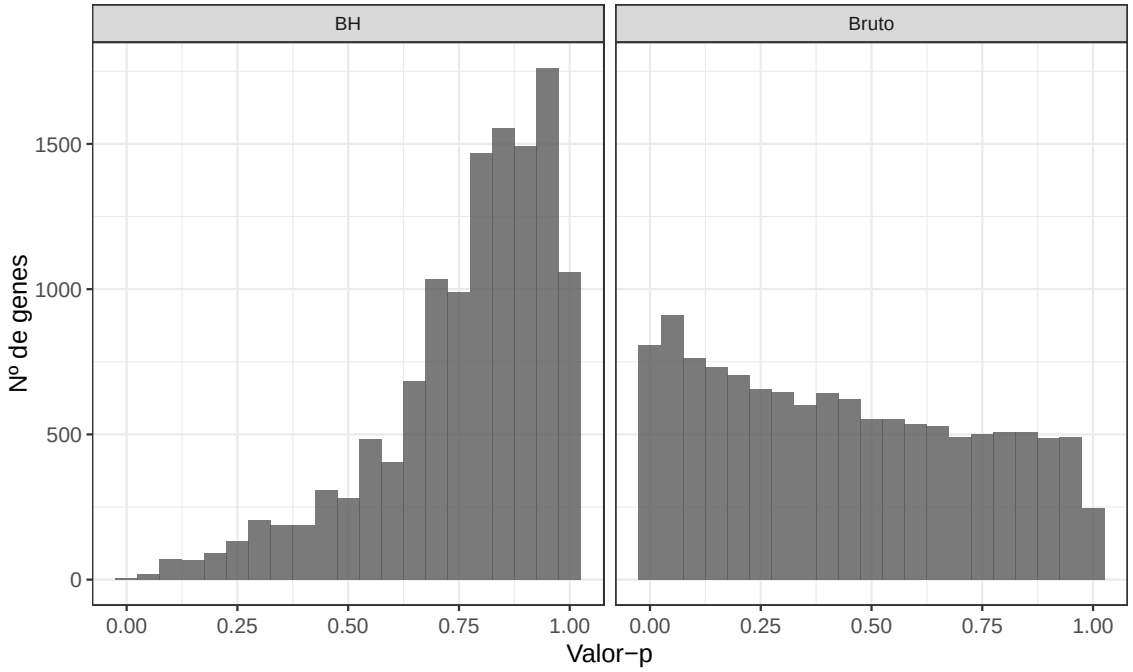


FIGURA 5.21: Histograma dos valores-p obtidos da comparação genética entre os dois tipos de camundongos, ajustados pelo método BH (à esquerda) e não ajustados (à direita).

Para uma melhor perceção da situação de rejeição dos testes, apresentam-se na Figura 5.22 os valores-p ordenados por ordem crescente. As retas representadas são as retas de corte para o método bruto (a azul) e para o método BH (a vermelho); os valores-p abaixo das mesmas conduzirão a rejeições pelo método em questão.

Na sub-figura está apresentado um zoom dos primeiros genes para que se consiga ter uma melhor perspetiva da rejeição pelo método ajustado de BH (para se poder representar

a sub-figura é necessário o pacote *patchwork* no R).

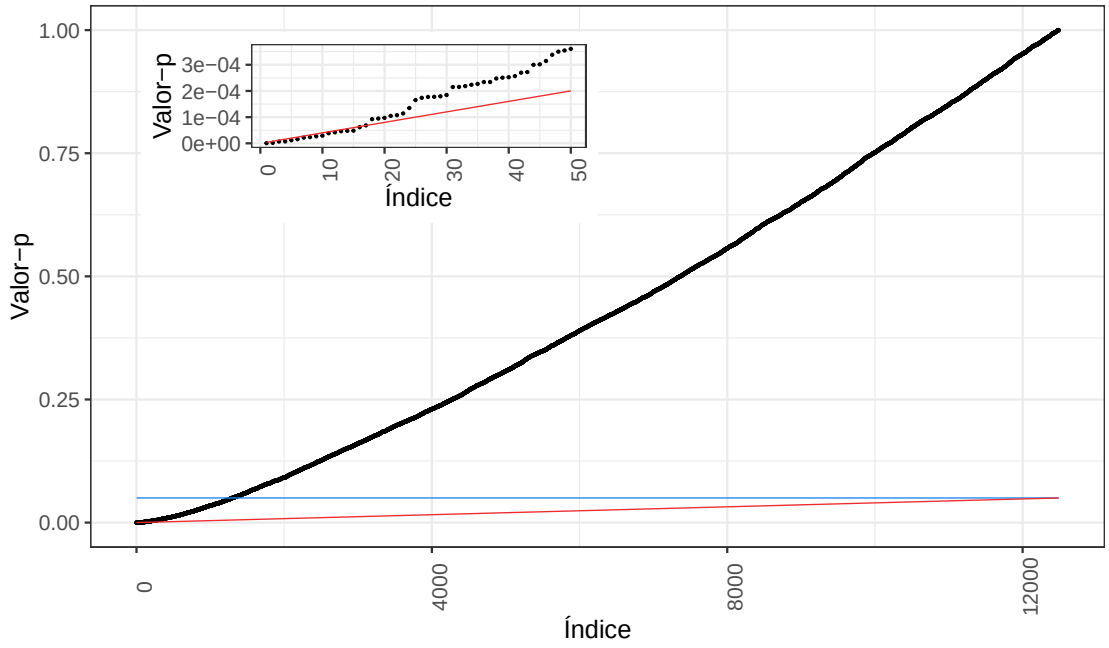


FIGURA 5.22: Valores-p brutos ordenados e retas de corte para o método bruto (reta azul) e para o método BH (reta vermelha).

Do que se pode concluir da Figura 5.22, o método BH rejeita menos genes do que o método bruto. Mais precisamente, o procedimento BH rejeita 16 genes (0.13%) e o bruto rejeita 1298 genes (10.39%).

Atente-se agora na Figura 5.23 onde está representado um gráfico de dispersão dos genes colorido de acordo com a rejeição ou não rejeição de H_0 pelos dois métodos tratados.

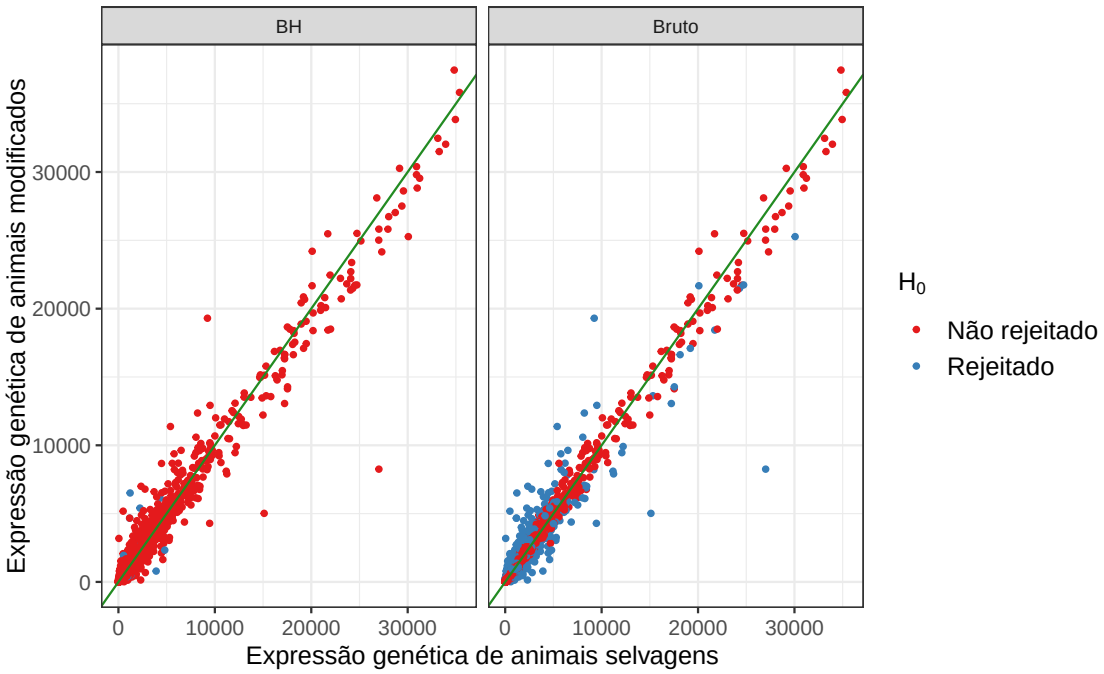


FIGURA 5.23: Gráfico de dispersão entre as expressões genéticas dos dois tipos de camundongos.

Quando se usa o método bruto, existem pontos próximos da reta $y = x$ que são rejeitados, enquanto que pontos mais afastados da mesma não o são. Tal pode ser explicado pelos valores dos desvios padrões. De facto, se dois genes tiverem uma diferença média muito próxima mas com um desvio padrão muito baixo, será possível distinguir a expressão genética dos mesmos para os dois grupos em causa, enquanto que se dois genes tiverem uma diferença média maior mas com um desvio padrão elevado, poderá não ser possível distingui-los, e portanto esses genes não serão rejeitados.

Na Figura 5.24 observa-se a relação entre a estatística de teste e o valor-p correspondente a cada gene. É de salientar que a estatística de teste já inclui o desvio padrão, anteriormente referido, no seu cálculo.

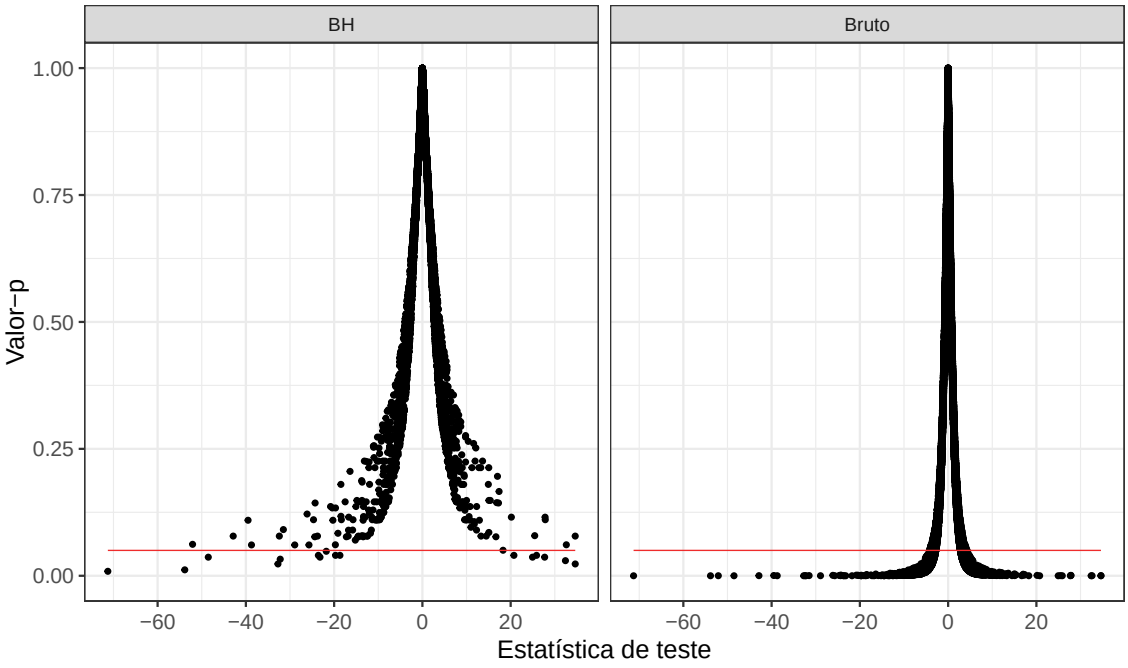


FIGURA 5.24: Relação entre os valores-p e as respetivas estatísticas de teste.

Constata-se que os valores da estatística de teste mais afastados de zero são os que são rejeitados pelos dois métodos. No entanto, o método BH pesa as caudas da distribuição rejeitando assim menos hipóteses do que o método bruto.

Se se aplicar o Fator de Bayes a este conjunto de dados é de esperar, pelo que foi estudado em seções anteriores, que não haja quaisquer evidências a favor de H_0, pois o tamanho das amostras é muito baixo. A Figura 5.25 apresenta o gráfico de dispersão do conjunto de dados colorido de acordo com o grau de evidência associado a cada gene.

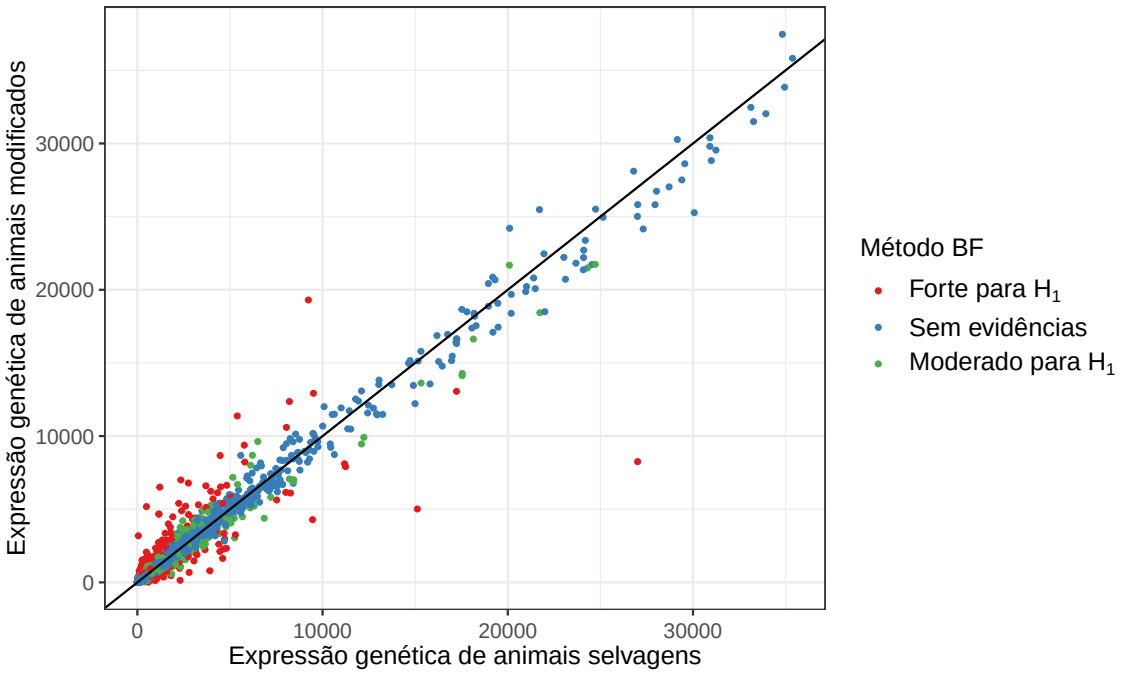


FIGURA 5.25: Gráfico de dispersão entre as expressões genéticas dos dois tipos de camundongos.

De facto, observa-se que não existem quaisquer genes que tenham evidências a favor de H_0. Numericamente, existem 733 genes (5.87%) que apresentam evidências moderadas - a verde - para a diferença das expressões genéticas médias e 361 (2.89%) com evidências fortes - a vermelho. Os restantes não apresentam qualquer evidência e estão representados a azul.

Do estudo feito, é possível observar que existem 16 genes que são rejeitados pelos dois métodos aplicados. Assim, podemos afirmar que, para esses genes, existe realmente uma diferença significativa na sua expressão. No entanto, para os restantes genes que são rejeitados pelo método bruto mas não o são pelo método BH nada podemos afirmar sobre a diferença de expressão entre os dois tipos de camundongos.

Capítulo 6

Considerações finais

O problema da inferência múltipla é reconhecido há muitos anos, mas só recentemente é que os testes de hipóteses múltiplas se tornaram numa área mais intensa de investigação devido à sua maior utilização em estudos nomeadamente na área da Genómica. As abordagens mais utilizadas continuam a ser as clássicas para o controlo da FWER, embora as abordagens que controlam a FDR comecem a ganhar grande popularidade. Apesar de muita investigação já realizada, não há indicações objetivas sobre que método utilizar em cada problema.

Pelos resultados obtidos, comparando os três métodos estudados, concluiu-se que o método de Benjamini-Hochberg é um método intermédio, no que toca a erros de tipo I e II, entre o método bruto - que comete mais erros de tipo I - e o método de Bonferroni - que comete mais erros de tipo II. No entanto, uma vez que a importância de cada um destes erros depende do estudo em questão, não há evidências para a escolha do melhor método. Assim, continua em aberto a questão sobre qual é o melhor método para um estudo de hipóteses múltiplas. Contudo, é notório que a abordagem que exige o controlo da FDR pode ser desenvolvida num processo simples e potente.

Ao aplicar o fator de Bayes aos testes de hipóteses múltiplas, verificou-se que este não é influenciável pelo número de testes, mas apenas pelo tamanho das amostras. Este método será uma boa estratégia a aplicar quando o tamanho da amostra é considerável, de forma a poder suportar evidências a favor da hipótese nula.

É importante lembrar que existem outros métodos que não foram aqui considerados. Assim, seria interessante realizar estudos similares aos observados nesta dissertação, com outras metodologias, no sentido de clarificar vantagens e desvantagens dos diferentes métodos.

Bibliografia

[1] J. Neyman, E. S. Pearson, and K. Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[2] R. A. Fisher, *Statistical Method for Research Workers*. Oliver & Boyd, 1925.

[3] S. R. Austin, I. Dialsingh, and N. S. Altman, "Multiple hypothesis testing: A review," *J. Indian.Soc. Agric. Stat.*, vol. 68, no. 2, pp. 303–314, 2014.

[4] D. Pascovici, D. C. L. Handler, J. X. Wu, and P. A. Haynes, "Multiple testing corrections in quantitative proteomics: A useful but blunt tool," *Proteomics*, vol. 16, no. 18, pp. 2448–2453, 2016.

[5] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics*. Springer, 2005.

[6] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[7] A. Agresti, C. Franklin, and B. Klingenberg, *Statistics: The Art and Science of learning from data*. Pearson, 2018.

[8] C. Keysers, V. Gazzola, and E.-J. Wagenmakers, "Using bayes factor hypothesis testing in neuroscience to establish evidence of absence." *Nat. Neurosci.*, vol. 23, pp. 788–799, 2020.

[9] R. E. Kass and A. E. Raftery, "Bayes factor," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[10] H. Jeffreys, *Theory of probability*, 3rd ed. Oxford University Press, 1961.

[11] S. K. Mathur, *Statistical Bioinformatics with R*. Elsevier, 2010.

[12] J. P. Shaffer, "Multiple hypothesis testing," *Annual Review of Psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[13] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[14] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[15] G. Hommel, "A stagewise rejective multiple test procedure based on a modified bonferroni test," *Biometrika*, vol. 75, no. 2, pp. 383–386, 1988.

[16] A. S. Foulkes, *Applied Statistical Genetics with R*. Springer, 2009.

[17] C. Genovese and L. Wasserman, "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 499–517, 2002.