

LGN5830 - Biometria de Marcadores Genéticos

Tópico 3: Mapas Genéticos I Segregação Mendeliana

Antonio Augusto Franco Garcia

<http://about.me/augusto.garcia>
augusto.garcia@usp.br

Departamento de Genética
ESALQ/USP
2015



Conteúdo

- 1 Teste de Segregação
 - Testes de Hipóteses
 - p-valores
 - Teste de Aderência
- 2 Múltiplos Testes
 - Princípios
 - Correção de Bonferroni
 - False Discovery Rate (FDR)
- 3 Referências

Etapas Cartoon Guide to Statistics

- Hipóteses

Step 1. FORMULATE ALL HYPOTHESES.

H₀ THE NULL HYPOTHESIS, IS
USUALLY THAT THE
OBSERVATIONS ARE THE RESULT
PURELY OF CHANCE.

H_a THE ALTERNATE HYPOTHESIS,
IS THAT THERE IS A REAL
EFFECT, THAT THE
OBSERVATIONS ARE THE
RESULT OF THIS REAL EFFECT,
PLUS CHANCE VARIATION.



- Estatística do Teste

Step 2. THE TEST STATISTIC.
IDENTIFY A STATISTIC THAT WILL ASSESS
THE EVIDENCE AGAINST THE NULL
HYPOTHESIS.



- Obtenha o *p*-valor

Step 3. P-VALUE:

A PROBABILITY STATEMENT WHICH ANSWERS THE QUESTION: IF THE NULL HYPOTHESIS WERE TRUE, THEN WHAT IS THE PROBABILITY OF OBSERVING A TEST STATISTIC AT LEAST AS EXTREME AS THE ONE WE OBSERVED?



- Tome a decisão

Step 4. COMPARE THE P-VALUE TO A FIXED SIGNIFICANCE LEVEL, α .

α ACTS AS A CUT-OFF POINT
BELOW WHICH WE AGREE THAT AN
EFFECT IS STATISTICALLY SIGNIFI-
CANT. THAT IS, IF

$U \in \mathcal{A}$

THEN WE RULE OUT THE NULL
HYPOTHESIS H_0 AND AGREE THAT
SOMETHING ELSE IS GOING ON.



Exemplo: Diferença entre duas médias

- Uma população homogênea foi genotipada com um marcador dominante
- Deseja-se saber se o peso dos indivíduos é diferente em função do genótipo do marcador (presença/ausência)
- Dados:

$\mu_1 = 17.5$	$n_1 = 20$	$\sigma_1^2 = 7.4$
$\mu_2 = 15.0$	$n_2 = 18$	$\sigma_2^2 = 6.9$

- $H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 \neq \mu_2$
- Estadística:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Exemplo: Diferença entre duas médias

$\mu_1 = 17.5$	$n_1 = 20$	$\sigma_1^2 = 7.4$
$\mu_2 = 15.0$	$n_2 = 18$	$\sigma_2^2 = 6.9$

- $t_{obs} = 2.88, t_{0.05;36} = 2.028$
- Conclusão?

SÃO EQUIVALENTES: falsa descoberta (false discovery), erro tipo I, falso positivo

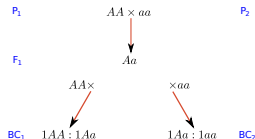
Testes de Hipóteses

Definição

O *p*-valor é definido como a probabilidade de observar valores mais extremos da estatística do teste sob H_0 do que o valor observado. Sendo T a estatística do teste que assume valores positivos,

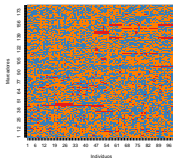
$$p = Pr(T \geq T_{obs} | H_0)$$

Retrocruzamentos Fundamentos



Dados

Cana-de-açúcar (1:1), Garcia et al. 2006



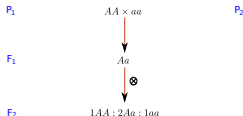
Retrocruzamentos

	AA	Aa
Freq. esperada	1/2	1/2
n. esp.	$n/2$	$n/2$
n. obs.	n_1	n_2

$$\chi^2 = \sum \frac{(n_{obs} - n_{esp})^2}{n_{esp}} = \frac{(n_1 - n/2)^2}{n/2} + \frac{(n_2 - n/2)^2}{n/2} = \frac{(n_1 - n_2)^2}{n} \sim \chi_1^2$$

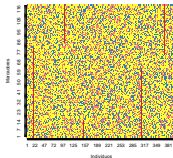
Quantos GLs?

1 (para $\theta = 1/2$)



Dados

Milho (1:2:1), Sibov et al. 2003

 F_2

	AA	Aa	aa
Freq. esperada	1/4	1/2	1/4
n. esp.	n/4	n/2	n/4
n. obs.	n_1	n_2	n_3

$$\chi^2 = \sum \frac{(n_{obs} - n_{esp})^2}{n_{esp}} = \frac{(n_1 - n/4)^2}{n/4} + \frac{(n_2 - n/2)^2}{n/2} + \frac{(n_3 - n/4)^2}{n/4} \sim \chi_2^2$$

Quantos GL?

Dois: θ_1 e θ_2 (multinomial)

RILs

	AA	aa
Freq. esperada	1/2	1/2
n. esp.	$n/2$	$n/2$
n. obs.	n_1	n_2

$$\chi^2 = \sum \frac{(n.obs - n.esp)^2}{n.esp} = \frac{(n_1 - n_2)^2}{n} \sim \chi_1^2$$

1 GL



2010 FIFA World Cup

Teams	Stage	Date	Prediction	Result	Outcome
 Germany vs  Australia	Group stage	13 June 2014	4-0	Correct	
 Germany vs  SOUTH AFRICA	Group stage	18 June 2014	0-1	Correct	
 Ghana vs Germany	Group stage	23 June 2014	0-1	Correct	
 Germany vs  England	Round of 16	27 June 2014	4-1	Correct	
 Argentina vs Germany	Quarter-finals	2 July 2014	0-4	Correct	
 Germany vs Spain	Semi-finals	7 July 2014	0-1	Correct	
 Uruguay vs Germany	3rd place play-off	10 July 2014	2-0	Correct	
 Netherlands vs Germany	Final	11 July 2014	0-1	Correct	

Acaso?

- Qual a probabilidade de observar este evento (excluindo empates)?
- $(\frac{1}{2})^8$
- 1 em 256
- A fama do polvo começou após acertar o resultado de *Alemanha vs Inglaterra*
- Porém, com 178 indivíduos, há grande chance de alguém acertar o resultado de uma série de 8 jogos
 - Qual a probabilidade de encontrar duas pessoas que fazem aniversário na mesma data, numa sala com pessoas tomadas ao acaso?
 - Com 57 pessoas, a probabilidade é 99%! (<http://goo.gl/5irBA>)

Múltiplos Testes

- Mapeamento Genético: normalmente, os testes são realizados repetidas vezes
- $1 - \alpha$: probab. de não cometer erro tipo I em um teste
- $(1 - \alpha)^m$: prob. de não cometer *erro tipo I* nos m testes
- Note que estamos assumindo que os m testes são **independentes**
- α^* : erro **conjunto** tipo I
- Logo, $1 - \alpha^* = (1 - \alpha)^m$ e $\alpha^* = 1 - (1 - \alpha)^m$

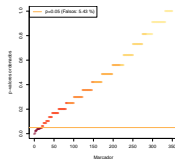
Múltiplos Testes

Exemplo - Mouse Data

- $m = 14$
- $\alpha = 0.05$
- Qual a probabilidade de ocorrer pelo menos um falso positivo nos 14 testes?
- Resp.: $\alpha^* = 0.51$

Múltiplos Testes

Simulação: 350 marcadores, $n=300$ (RC)



Bonferroni

Sidák

$$1 - \alpha^* = (1 - \alpha)^m$$

$$\sqrt[m]{1 - \alpha^*} = 1 - \alpha$$

$$\alpha = 1 - \sqrt[m]{1 - \alpha^*}$$

Bonferroni

$$\alpha^* = 1 - (1 - \alpha)^m = m\alpha - \binom{m}{2}\alpha^2 + \binom{m}{3}\alpha^3 - \binom{m}{4}\alpha^4 + \dots$$

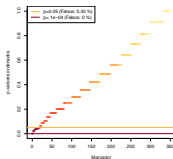
$$\alpha \approx \frac{\alpha^*}{m}$$

Exemplo - Mouse Data

- $m = 14$
- $\alpha^* = 0.05$
- Qual valor de α deve ser usado em cada teste para garantir esse valor global de 5%?
- Resp.: $\alpha = 0.00357$ (menor que 0.05)

Bonferroni

Simulação: 350 marcadores, $n=300$ (RC)



Múltiplos testes - mapas genéticos

Pontos para reflexão

- Os m testes são independentes no caso dos mapas genéticos?
- São graves as consequências de não descartar marcas que não segregam mendelianamente?
 - SIM Fu e Ritland. 1994, Lorieux et al. 1995a, b; Vogl e Xu 2000; Luo e Xu 2003; Luo et al. 2005; Wang et al. 2005
 - NÃO Zhao-Bang Zeng
 - TALVEZ Xu, S. 2008. Quantitative trait locus mapping can benefit from segregation distortion. *Genetics* 180 (4): 2201-2208.
- A correção de Bonferroni é conhecidamente **conservativa**.

"Naive approach" para mapas genéticos

- Assumindo independência condicional (propriedade markoviana)

$$\begin{array}{c} \underbrace{M_1, M_2, \dots, M_i}_{37.5 \text{ cM}} \underbrace{M_{i+1}, M_{i+2}, \dots, M_m}_{37.5 \text{ cM}} \\ \underbrace{M_1, M_2, \dots, M_i}_{(1-\alpha) \cdot 1 \dots 1} \underbrace{M_{i+1}, M_{i+2}, \dots, M_m}_{(1-\alpha) \cdot 1 \dots 1} \\ 1 - \alpha^* = (1 - \alpha)^2 = (1 - \alpha)^{m^*} \end{array}$$

Número de Testes

$$m^* = \frac{L}{37.5}$$

Resultados possíveis

m p-valores

	Signif.	Não signif.	Total
H_0 verdadeiro	F	$m_0 - F$	m_0
H_a verdadeiro	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Razão de Falsas Descobertas (FDR)

- False Discovery Rate:** alternativa para controle do erro tipo I
- Seu uso é frequente em experimentos de expressão gênica, SNPs (genômica), etc (e várias outras áreas)
- Motivação: usar $\alpha = 0.05$ (ou $\alpha = 0.01$) fornece muitos falso positivos; usar α^* elimina muitos positivos verdadeiros
- Princípio: dados os resultados significativos, determina-se quantos deles (proporção) são verdadeiramente significativos ($1 - FDR$)

	Signif.	Não signif.	Total
H_0 verdadeiro	F	$m_0 - F$	m_0
H_a verdadeiro	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Definição

FDR: É a proporção esperada de falsas descobertas dentre as hipóteses H_0 rejeitadas

$$\frac{\text{n. falsos positivos}}{\text{n. testes significativos}} = \frac{F}{F + T} = \frac{F}{S}$$

$$FDR = E \left[\frac{F}{F + T} \right] = E \left[\frac{F}{S} \right]$$

FDR

	Signif.	Não signif.	Total
H_0 verdadeiro	F	$m_0 - F$	m_0
H_a verdadeiro	T	$m_1 - T$	m_1
Total	S	$m - S$	m

- Seja t o threshold (limiar) usado para considerar os p -valores como significativos ($0 < t \leq 1$)
- Para m muito grande (p. ex., milhares):

$$FDR(t) = E \left[\frac{F(t)}{S(t)} \right] \approx \frac{E[F(t)]}{E[S(t)]}$$

- Uma estimativa de $E[S(t)]$ é o número $S(t)$ observado (isto é, o número observado de p -valores menores ou iguais a t)
- $E[F(t)] = m_0 \cdot t$

FDR

	Signif.	Não signif.	Total
H_0 verdadeiro	F	$m_0 - F$	m_0
H_a verdadeiro	T	$m_1 - T$	m_1
Total	S	$m - S$	m

- Note que m_0 não é conhecido!
- É usual considerar $\pi_0 = m_0/m$, e não m_0 (fácil interpretação)

Estimativa do FDR

$$\widehat{FDR}(t) = \frac{\widehat{\pi}_0 \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{S(t)}$$

FDR

- $\pi_0 = \frac{m_0}{m}$: pode ser estimado com base na distribuição dos p -valores sob H_0

Simulação - 350 locos (1:1) sob H_0



FDR

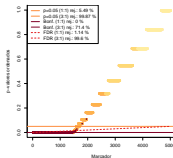
Cálculo

$$\widehat{FDR}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)}$$

$$t = \frac{\widehat{FDR}(t) \cdot S(t)}{\hat{\pi}_0 m}$$

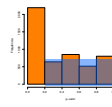
Simulação

100 ind., 5000 marc. (3500 1:1 e 1500 3:1), teste para 1:1



π_0

p-valores



Estimativas de π_0

- Análise visual
 $\pi_0 = \frac{0.8 \times 750}{0.2 \times 2200 + 0.8 \times 750} = 0.59$
- Software Q-VALUE
<http://genome.org/qvalue/>

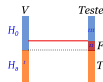
```
library(qvalue)
q <- qvalue(x[,1])
q
$pi0
[1] 0.5648856
$nulls
[1] 1.756017e+11 9.618898e+11 ...
```

Valor Real

$$\pi_0 = \frac{m_0}{m} = \frac{3500}{3500+1500} = 0.70$$

FDR vs Erro Tipo I

	Signif.	Não signif.	Total
H_0 verdadeiro	F	$m_0 - F$	m_0
H_a verdadeiro	T	$m_1 - T$	m_1
Total	S	$m - S$	m



- Erro Tipo I:
 $\frac{F}{T+T}$
- FDR: $\frac{F}{T+T}$

Comparação

$$t = \frac{FDR(t).i}{\pi_0 m}$$

$$\alpha = \frac{\alpha^*}{m}$$

Atenção

FDR é conservativo

Dados reais

Hedenfalk et al. 2001




- Expressão diferencial de 3226 genes (câncer)
- Usando p-valor 0.001 para determinar significância, encontraram 51 genes diferencialmente expressos (sugestivos), sendo apenas 9-11 deles tomados como diferencialmente expressos
- Com base nos q-valores (limiar 0.05), Storey e Tibshirani (2003) encontraram evidências de que 160 genes são diferencialmente expressos (sendo que cerca de 8 desses 160 possivelmente sejam falsos positivos)

Considerações Finais

Alguns pontos

- **FDR**: balanço entre o número de falsos positivos e o número de positivos verdadeiros
- Interessante para estudos exploratórios (ex: expressão gênica), em que não faz sentido preocupar-se em demasia com os genes sob H_0
- Não é recomendado para mapas genéticos ou QTLs (!)
- Pode ser interessante para Mapeamento Associativo
- Trabalhos recentes consideram o problema da dependência dos testes (discutiremos oportunamente)

Principais Referências

-  Storey, John D.; Tibshirani, Robert
Statistical significance for genomewide studies
Proc. Nat. Acad. Sci. 100: 9440-9445, 2003
-  Storey, John D.
False Discovery Rates
International Encyclopedia of Statistical Science 1: 504-508, 2011
-  Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S.
Posterior error probabilities and false discovery rates: two sides of the same coin
Journal of Proteome Research 7: 40-44, 2008