

MAE0418 - Estatística Documentária  
Lista 3  
Equilíbrio de Hardy-Weinberg e o sistema ABO  
Manual de Uso

Fernando Henrique Ferraz Pereira da Rosa  
Guilherme Miguel Mitne  
Vagner Aparecido Pedro Junior

24 de outubro de 2004

## **Resumo**

Descrevemos no presente trabalho a modelagem através de técnicas de análise categorizada de dados estudadas no curso de Estatística Documentária, do equilíbrio de Hardy-Weinberg para tabelas do sistema de classificação ABO.

Estimamos os parâmetros do modelo através de máxima verossimilhança, utilizando procedimentos iterativos. Essa estimação foi implementada no pacote estatístico R[1]. Descrevemos o uso dessa implementação e discutimos alguns exemplos.

# 1 Metodologia

## 1.1 O equilíbrio de Hardy-Weinberg

Seja uma população com muitos indivíduos em que cada um possui um par particular de genes. Cada gene desse par pode ser de um dado tipo (por exemplo, a ou A). Suponhamos ainda que os acasalamentos entre os indivíduos dessa população sejam aleatórios ou não seletivos, que a população não esteja sujeita a migrações nem a mutações constantes. Segundo o equilíbrio de Hardy-Weinberg[2], as frequências e razões genotípicas nessa população serão constantes de geração para geração.

Notemos que as suposições para a validade da lei são muito restritivas, e que na maioria dos casos portanto ela só pode ser aplicada a populações teóricas. Entretanto, ela é uma importante ferramenta para os geneticistas no estudo de populações naturais. Através das fugas do equilíbrio de Hardy-Weinberg nessas populações, é possível estudar seu comportamento e modificações. É portanto de grande interesse a possibilidade de se delinear um método quantitativo para verificar se uma certa população está de acordo com a lei de Hardy-Weinberg para um dado par de genes.

## 1.2 O modelo probabilístico considerado

Consideremos o sistema ABO de classificação de tipo sanguíneo. Existem quatro fenótipos diferentes, “A”, “B”, “O” e “AB”, e três genes diferentes envolvidos:  $i$ ,  $I_a$  e  $I_b$ . Os pares de genes e os fenótipos se relacionam de acordo com a Tabela 1.

Fenótipo	Genótipos
O	$ii$
A	$I_a I_a, I_a i$
B	$I_b I_b, I_b i$
AB	$I_a I_b$

Tabela 1: Relação entre fenótipos e genótipos

Seja  $N$  o tamanho da população considerada. Vamos modelar a probabilidade de ocorrência de cada um dos fenótipos através de uma distribuição multinomial com parâmetros  $N, \theta_O, \theta_A, \theta_B, \theta_{AB}$ .

## 1.3 O modelo estrutural considerado

Sejam agora  $\beta_A, \beta_B$  e  $\beta_O$  as frequências relativas na população dos genes  $I_a, I_b$  e  $i$  respectivamente. Sob o equilíbrio de Hardy-Weinberg, temos que os parâmetros  $\theta_x, x = A, B, AB, O$  do modelo probabilístico considerado podem ser expressos em função dos parâmetros  $\beta_x, x = A, B, O$ . Consideremos então o modelo estrutural:

$$H : \underline{\theta} = \underline{\theta}(\underline{\beta})$$

com:

$$H : \begin{pmatrix} \theta_O \\ \theta_A \\ \theta_B \\ \theta_{AB} \end{pmatrix} = \begin{pmatrix} \beta_O^2 \\ \beta_A^2 + 2\beta_A\beta_O \\ \beta_B^2 + 2\beta_B\beta_O \\ 2\beta_A\beta_B \end{pmatrix}$$

onde, dada a restrição natural em  $\underline{\beta}$ , podemos ainda escrever  $\beta_O = 1 - \beta_A - \beta_B$ .

A verossimilhança sob o equilíbrio de Hardy-Weinberg é dada por:

$$\begin{aligned} L(\underline{\beta} | \underline{n}) &= \frac{N!}{n_O!n_A!n_B!n_{AB}!} \theta_O^{n_O} \theta_A^{n_A} \theta_B^{n_B} \theta_{AB}^{n_{AB}} \\ &\propto \beta_O^{2n_O} (\beta_A^2 + 2\beta_A\beta_O)^{n_A} (\beta_B^2 + 2\beta_B\beta_O)^{n_B} (2\beta_A\beta_B)^{n_{AB}} \quad (1) \end{aligned}$$

## 1.4 Descrição dos métodos iterativos

A partir da verossimilhança dada em 1, podemos então obter os estimadores de máxima verossimilhança de  $\underline{\beta}$ . Pelo princípio da invariância, podemos obter a partir desse estimador o estimador de máxima verossimilhança de  $\underline{\theta}$ , de acordo com o modelo estrutural.

Não é possível obter os estimadores de máxima verossimilhança de forma explícita. Consideraremos então três métodos iterativos para obtenção de  $\hat{\underline{\beta}}$ .

### 1.4.1 Multiplicadores de Lagrange

Usando o método dos multiplicadores de Lagrange, obtemos a seguinte relação de recorrência para encontrar  $\hat{\underline{\beta}}$ :

$$\begin{cases} \hat{\beta}_j &= p_j \frac{\hat{\beta}_j + \hat{\beta}_O}{\hat{\beta}_j + 2\hat{\beta}_O} + \frac{p_{AB}}{2}, \quad j = A, B \\ \hat{\beta}_O &= p_0 + \sum_{j=A,B} \frac{p_j \hat{\beta}_O}{\hat{\beta}_j + 2\hat{\beta}_O} \end{cases}$$

onde  $p_j = n_j / N$ .

### 1.4.2 Newton-Raphson

Seja  $H(\underline{\beta}, \underline{n})$  a matriz Hessiana calculada em  $\underline{\beta}$  e  $\underline{n}$ , e  $U(\underline{\beta}, \underline{n})$  a função score, calculada também nos mesmos parâmetros. O método de Newton-Raphson nos provê a seguinte relação de recorrência para encontrar  $\hat{\underline{\beta}}$ :

$$\underline{\beta}^{(k)} = \underline{\beta}^{(k-1)} - \left[ H(\underline{\beta}^{(k-1)}, \underline{n}) \right]^{-1} U(\underline{\beta}^{(k-1)}, \underline{n}).$$

### 1.4.3 Scoring de Fisher

Seja  $\mathcal{I}(\underline{\beta}) = E(-H(\underline{\beta}, \underline{n}))$  a matriz de informação de Fisher. O método de Scoring de Fisher, define a seguinte relação iterativa para encontrarmos o estimador de máxima verossimilhança:

$$\underline{\beta}^{(k)} = \underline{\beta}^{(k-1)} + \left[ \mathcal{I}(\underline{\beta}^{(k-1)}) \right]^{-1} U(\underline{\beta}^{(k-1)}, \underline{n}).$$

#### 1.4.4 Chutes Iniciais

Os três métodos propostos precisam de um valor inicial para que seja começado o processo iterativo. Nos três casos, consideraremos o estimador dos momentos para obter esse chute. Temos:

$$\begin{aligned}\beta_A^{(0)} &= 1 - \sqrt{p_O + p_B} \\ \beta_B^{(0)} &= 1 - \sqrt{p_O + p_A} \\ \beta_O^{(0)} &= 1 - \beta_A^{(0)} - \beta_B^{(0)}\end{aligned}$$

## 2 Uso do Programa

Implementamos o programa através de um conjunto de funções no R, na qual a interface principal para análise dos dados é a função `hardy.weinberg()`. Ela tem os parâmetros:

```
> args(hardy.weinberg)

function (dados, eps.min = 1e-04, n.max = 30, method = c("lagrange",
  "newton", "fisher"))
NULL
```

O parâmetro `dados` é o vetor com os dados (na ordem O,A,B,AB) a serem analisados. Os parâmetros `eps.min` e `n.max` controlam os critérios de convergência. Eles são por padrão 0.0001 e 30. Por fim o parâmetro `method` controla qual dos três métodos iterativos será usado para obter as estimativas de máxima verossimilhança.

Ao executar a função com os parâmetros desejados é criado um objeto da classe `hardy.weinberg`, que pode ser mostrado na tela ou guardado para futuras análises. O objeto criado contém todas as informações a respeito da análise, como o valor das estatísticas, os valores escolhidos de `n.max` e `eps.min` e o número de iterações realizadas.

Para entrar com os dados, basta criar um vetor com os valores de  $\tilde{n}$ , na ordem O, A, B, AB:

```
> sangue1 <- c(4578, 4219, 890, 313)
```

Por padrão, no caso de a função ser chamada somente com o argumento `dados`, é utilizado o método de Multiplicadores de Lagrange.

```
> m1 <- hardy.weinberg(sangue1)
```

O comando acima vai fazer a análise dos dados contidos em `sangue1`, definido conforme a linha anterior, utilizando o método de Multiplicadores de Lagrange com  $n$  e  $\epsilon$  padrões. O resultado fica guardado no objeto `m1`. Podemos então ver o resultado da análise imprimindo esse objeto na tela:

```
> m1
```

### Equilíbrio de Hardy-Weinberg para o sistema ABO

#### Dados Utilizados

O	A	B	AB	N
4578	4219	890	313	10000

#### Método de Estimação Utilizado

Multiplicadores de Lagrange

eps.min = 1e-04      n.max = 30

n efetivo = 2

#### Chutes iniciais para beta

A	B	O
0.26054074	0.06207676	0.67738250

#### Estimativas finais de beta

A	B	O
0.26063895	0.06210068	0.67726036

#### Estimativas finais de teta

O	A	B	AB
0.45868160	0.42097353	0.08797316	0.03237171

#### Valores Esperados

O	A	B	AB
4587	4210	880	324

#### Estatísticas de Teste

	obs	gl	P-valor
$Q_V$	0.5155106	1	0.4727630
$Q_P$	0.5119967	1	0.4742758
$Q_N$	0.5227503	1	0.4696708

Criamos também um método `plot` para a classe `hardy.weinberg`, de forma que além do resultado poder ser mostrado na tela, pode ser feito um gráfico a partir dele (vide Figura 1). O gráfico contém os gráficos de barra dos valores esperados e observados para cada grupo, os P-valores para as estatísticas  $Q_V$ ,  $Q_P$  e  $Q_N$ , com o gráfico da distribuição Qui-Quadrado com 1 grau de liberdade e a área preenchida indicando o P-Valor.

Por fim, criamos um método `latex`, que permite que seja gerado um relatório em  $\text{\LaTeX}$  a partir da análise, automaticamente. O output por padrão é impresso na tela, e daí pode ser copiado e colado em um documento em  $\text{\LaTeX}$  existente. Para salvar o resultado em um arquivo, basta usar-se o parâmetro `file`. O comando abaixo gera o relatório e salva-o no rquivo `m1.tex`, no diretório corrente. Pode-se então inserir esse documento dentro de um documento  $\text{\LaTeX}$ :

```
> latex(m1, file = "m1.tex")
```

```
> plot(m1)
```

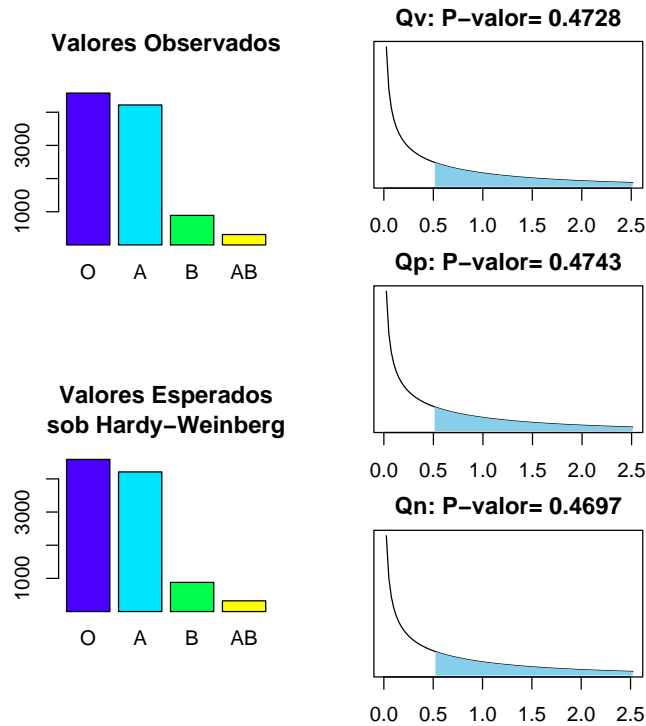


Figura 1: Gráficos para a análise dos dados `sangue1`

Note-se que deve ser definido no cabeçalho do documento o seguinte comando, para que a notação vetorial seja mostrada corretamente:

```
\newcommand{\stackunder}[2]{ \renewcommand{\arraystretch}{0.3}
\displaystyle \begin{array}[t]{1} {#1}\_{#2}\end{array}
\renewcommand{\arraystretch}{1}}
```

Inserindo os comandos diretamente no presente documento temos:

```
> latex(m1)
```

## Equilíbrio de Hardy-Weinberg para o sistema ABO

Dados considerados:

$$\tilde{n}' = (4578, 4219, 890, 313)$$

Método de estimação utilizado: Multiplicadores de Lagrange. Critérios de convergência:  $\epsilon_{\min} = 1e - 04$ ,  $n_{\max} = 30$ . Número de iterações realizadas:  $n = 2$ .

$$\tilde{\beta}^{(0)} = \begin{pmatrix} 0.26054 \\ 0.062077 \\ 0.67738 \end{pmatrix} \quad \tilde{\beta} = \begin{pmatrix} 0.26064 \\ 0.062101 \\ 0.67726 \end{pmatrix} \quad \tilde{\theta} = \begin{pmatrix} 0.45868 \\ 0.42097 \\ 0.087973 \\ 0.032372 \end{pmatrix}$$

Estimativas dos valores esperados:

$$\tilde{\mu}' = (4587, 4210, 880, 324)$$

Estatísticas de teste:

	obs	gl	P-valor
$Q_V$	0.51551	1	0.47276
$Q_P$	0.512	1	0.47428
$Q_N$	0.52275	1	0.46967

### 3 Exemplos de Uso

Nos exemplos abaixo, assume-se que o script `hw.R` esteja no diretório em que o R esteja sendo rodado. Caso não, deve ser especificado o seu caminho completo como parâmetro da função `source()`. Por exemplo: `source("a://hw.R")` ou `source("c://home/hw.R")`.

```
> source("hw.R")
> duodenal <- c(298, 214, 39, 13)
```

Com os comandos acima lemos o script na sessão corrente do R e guardamos os dados da Tabela de distribuição no vetor `duodenal`. Para fazer a análise através do método de Multiplicadores de Lagrange, fazemos:

```
> ex1 <- hardy.weinberg(duodenal, method = "lagrange")
> latex(ex1)
```

## Equilíbrio de Hardy-Weinberg para o sistema ABO

Dados considerados:

$$\tilde{n}' = (298, 214, 39, 13)$$

Método de estimação utilizado: Multiplicadores de Lagrange. Critérios de convergência:  $\epsilon_{\min} = 1e - 04$ ,  $n_{\max} = 30$ . Número de iterações realizadas:  $n = 2$ .

$$\tilde{\beta}^{(0)} = \begin{pmatrix} 0.22701 \\ 0.047214 \\ 0.72578 \end{pmatrix} \quad \tilde{\hat{\beta}} = \begin{pmatrix} 0.22688 \\ 0.047188 \\ 0.72593 \end{pmatrix} \quad \tilde{\hat{\theta}} = \begin{pmatrix} 0.52697 \\ 0.38088 \\ 0.070736 \\ 0.021412 \end{pmatrix}$$

Estimativas dos valores esperados:

$$\tilde{\hat{\mu}}' = (297, 215, 40, 12)$$

Estatísticas de teste:

	obs	gl	P-valor
$Q_V$	0.09432	1	0.75876
$Q_P$	0.0959	1	0.7568
$Q_N$	0.09135	1	0.76247

Para usarmos Newton-Raphson, com  $\text{eps}=0.00001$ , fazemos<sup>1</sup>:

```
> ex2 <- hardy.weinberg(duodenal, eps.min = 1e-05, method = "newton")  
> latex(ex2)
```

---

<sup>1</sup>Note-se que estamos trabalhando diretamente com a saída em L<sup>A</sup>T<sub>E</sub>X do programa, pois o Manual está escrito em nessa linguagem. Caso deseje-se, é possível ver o output que seria gerado na tela do R, com o comando `print(obj)` ou ainda simplesmente `obj`, numa sessão interativa.

## Equilíbrio de Hardy-Weinberg para o sistema ABO

Dados considerados:

$$\tilde{n}' = (298, 214, 39, 13)$$

Método de estimação utilizado: Newton-Raphson. Critérios de convergência:  $\epsilon_{\min} = 1e - 05$ ,  $n_{\max} = 30$ . Número de iterações realizadas:  $n = 2$ .

$$\tilde{\beta}^{(0)} = \begin{pmatrix} 0.22701 \\ 0.047214 \\ 0.72578 \end{pmatrix} \quad \tilde{\beta} = \begin{pmatrix} 0.22688 \\ 0.047188 \\ 0.72593 \end{pmatrix} \quad \tilde{\theta} = \begin{pmatrix} 0.52698 \\ 0.38088 \\ 0.070737 \\ 0.021412 \end{pmatrix}$$

Estimativas dos valores esperados:

$$\tilde{\mu}' = (297, 215, 40, 12)$$

Estatísticas de teste:

	obs	gl	P-valor
$Q_V$	0.09432	1	0.75876
$Q_P$	0.0959	1	0.7568
$Q_N$	0.09135	1	0.76247

Por fim, para usarmos o método de Scoring de Fisher:

```
> ex3 <- hardy.weinberg(duodenal, method = "fisher")  
> latex(ex3)
```

## Equilíbrio de Hardy-Weinberg para o sistema ABO

Dados considerados:

$$\tilde{n}' = (298, 214, 39, 13)$$

Método de estimação utilizado: Scoring de Fisher. Critérios de convergência:  $\epsilon_{\min} = 1e - 04$ ,  $n_{\max} = 30$ . Número de iterações realizadas:  $n = 2$ .

$$\tilde{\beta}^{(0)} = \begin{pmatrix} 0.22701 \\ 0.047214 \\ 0.72578 \end{pmatrix} \quad \tilde{\beta} = \begin{pmatrix} 0.22688 \\ 0.047188 \\ 0.72593 \end{pmatrix} \quad \tilde{\theta} = \begin{pmatrix} 0.52698 \\ 0.38088 \\ 0.070737 \\ 0.021412 \end{pmatrix}$$

Estimativas dos valores esperados:

$$\tilde{\mu}' = (297, 215, 40, 12)$$

Estatísticas de teste:

	obs	gl	P-valor
$Q_V$	0.09432	1	0.75876
$Q_P$	0.0959	1	0.7568
$Q_N$	0.09135	1	0.76247

É possível também obter componentes individuais ao invés da saída completa. Por exemplo, para comparar os valores estimados de beta entre os três métodos:

```
> ex1$beta
      A      B      O
0.226883 0.047188 0.725929
```

```
> ex2$beta
      A      B      O
0.226881 0.047188 0.725931
```

```
> ex3$beta
      A      B      O
0.226881 0.047188 0.725931
```

Para ter uma lista dos componentes que podem ser obtidos:

```
> str(ex1)
```

```

List of 10
 $ dados      : Named num [1:5] 298 214 39 13 564
  ..- attr(*, "names")= chr [1:5] "0" "A" "B" "AB" ...
 $ beta       : Named num [1:3] 0.2269 0.0472 0.7259
  ..- attr(*, "names")= chr [1:3] "A" "B" "0"
 $ teta       : Named num [1:4] 0.5270 0.3809 0.0707 0.0214
  ..- attr(*, "names")= chr [1:4] "0" "A" "B" "AB"
 $ esperados  : Named num [1:4] 297.2 214.8 39.9 12.1
  ..- attr(*, "names")= chr [1:4] "0" "A" "B" "AB"
 $ estadisticas:'data.frame':      3 obs. of  3 variables:
  ..$ obs     : num [1:3] 0.0943 0.0959 0.0914
  ..$ gl      : num [1:3] 1 1 1
  ..$ P-valor: num [1:3] 0.759 0.757 0.762
 $ method     : chr "Multiplicadores de Lagrange"
 $ n          : num 2
 $ eps        : num 1e-04
 $ n.max      : num 30
 $ chutes     : Named num [1:3] 0.2270 0.0472 0.7258
  ..- attr(*, "names")= chr [1:3] "A" "B" "0"
 - attr(*, "class")= chr "hardy.weinberg"

```

Podemos então obter somente as estatísticas  $Q_V$  de cada um dos métodos, fazendo:

```
> ex1$estadisticas["Qv", ]
```

```

      obs gl P-valor
Qv 0.09432  1 0.75876

```

```
> ex2$estadisticas["Qv", ]
```

```

      obs gl P-valor
Qv 0.09432  1 0.75876

```

```
> ex3$estadisticas["Qv", ]
```

```

      obs gl P-valor
Qv 0.09432  1 0.75876

```

Podemos ainda obter os gráficos para qualquer uma das análises. Na Figura 2, temos os gráficos para o caso da análise *ex2*.

Tomemos agora o conjunto de dados:

```
> outros <- c(25, 424, 844, 12)
```

Fazendo a análise por Fisher-Scoring, temos:

```
> ex4 <- hardy.weinberg(outros, method = "fisher")
> latex(ex4)
```

> plot(ex2)

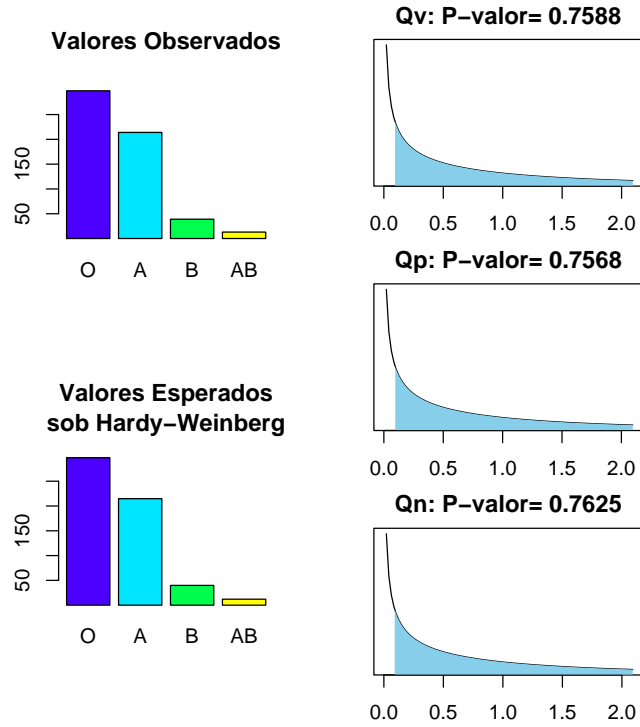


Figura 2: Gráficos para o ex2

## Equilíbrio de Hardy-Weinberg para o sistema ABO

Dados considerados:

$$\tilde{n}' = (25, 424, 844, 12)$$

Método de estimação utilizado: Scoring de Fisher. Critérios de convergência:  $\epsilon_{\min} = 1e - 04$ ,  $n_{\max} = 30$ . Número de iterações realizadas:  $n = 5$ .

$$\tilde{\beta}^{(0)} = \begin{pmatrix} 0.18397 \\ 0.41343 \\ 0.40259 \end{pmatrix} \quad \tilde{\beta} = \begin{pmatrix} 0.20506 \\ 0.45936 \\ 0.33558 \end{pmatrix} \quad \hat{\theta} = \begin{pmatrix} 0.11261 \\ 0.17968 \\ 0.51931 \\ 0.18839 \end{pmatrix}$$

Estimativas dos valores esperados:

$$\hat{\mu}' = (147, 234, 678, 246)$$

Estatísticas de teste:

	obs	gl	P-valor
$Q_V$	711.7	1	0
$Q_P$	517.63	1	0
$Q_N$	5269.8	1	0

e o respectivo gráfico na Figura 3. Observamos que rejeitamos a Hipótese de Equilíbrio de Hardy-Weinberg para essa população. Notemos ainda como o número de iterações realizadas foi maior até se obter a convergência.

```
> plot(ex4)
```

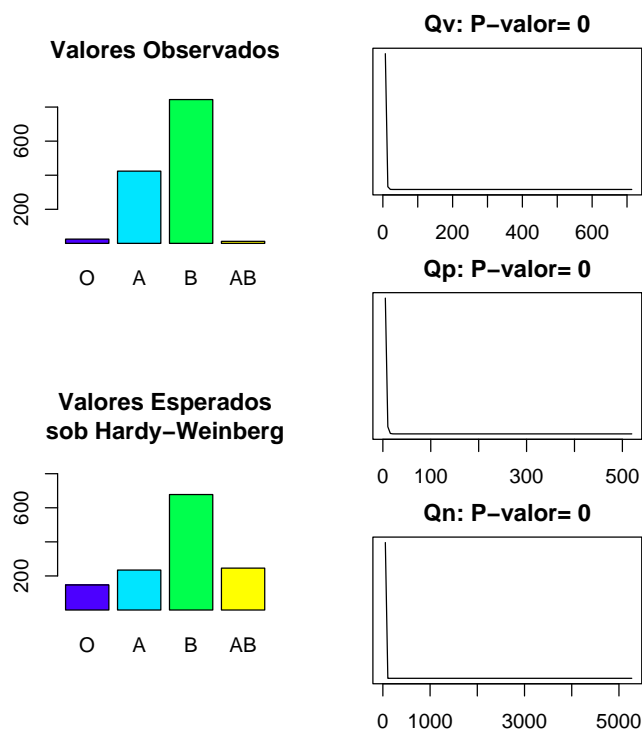


Figura 3: Gráficos para o ex4.

## Referências

- [1] R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2004. ISBN 3-900051-00-3. Disponível em: <<http://www.R-project.org>>.
- [2] UZUNIAN, A.; BINER, E. *Biologia 3*. São Paulo: Harbra, 1997.

## Sobre

A versão eletrônica desse arquivo pode ser obtida em <http://www.feferraz.net>

Copyright (c) 1999-2005 Fernando Henrique Ferraz Pereira da Rosa.  
É dada permissão para copiar, distribuir e/ou modificar este documento sob os termos da Licença de Documentação Livre GNU (GFDL), versão 1.2, publicada pela Free Software Foundation;

Uma cópia da licença em está inclusa na seção intitulada "Sobre / Licença de Uso".