

# Testes qui-quadrado e da razão de verossimilhanças

## 1. Simulações

São apresentados os passos para a geração de amostras em linguagem R e, a partir destas, o teste da hipótese da distribuição multinomial com probabilidades

$$\pi_1 = \theta^2, \pi_2 = \theta(1 - \theta), \pi_3 = \theta(1 - \theta) \text{ e } \pi_4 = (1 - \theta)^2, 0 < \theta < 1. \quad (1)$$

Para realizar o teste utilizamos as estatísticas  $-2 \log(\Lambda)$  e  $Q$  de Pearson. Inicialmente carregamos o pacote `lattice`, que inclui funções para os gráficos de quantis.

```
library(lattice)
```

Escolhemos o nível de significância nominal  $\alpha$  e calculamos o valor crítico obtido da distribuição de referência ( $\chi^2$  com 2 g.l.).

```
alfa <- 0.05  
(x2crit <- qchisq(1 - alfa, 2))
```

```
[1] 5.991465
```

Escolhendo o verdadeiro valor de  $\theta (= \theta_0)$  calculamos as probabilidades ( $\pi$ ) sob  $H_0$ .

```
teta0 <- 0.8  
pi1 <- teta0^2  
pi2 <- teta0 * (1 - teta0) # = pi3  
pi4 <- (1 - teta0)^2
```

Em seguida especificamos o tamanho amostral  $n$  e o número de repetições das simulações  $nrep$ .

```
n <- 200  
nrep <- 5000
```

Os dados correspondentes a todas as  $nrep$  repetições das simulações são gerados com a função `rmultinom` e são guardados em uma matriz  $4 \times nrep$  em que cada coluna representa uma amostra simulada.

```
dados <- rmultinom(nrep, size = n, prob = c(pi1, pi2, pi2, pi4))
```

As estimativas de máxima verossimilhança (EMV) irrestritas (ou seja, sob  $H_1$ ) de  $\pi$  e o logaritmo da função verossimilhança  $\log L_{\pi}$  (a menos de uma constante aditiva) são calculados por meio de funções matriciais. As EMV irrestritas de  $\pi$  são as proporções amostrais, que são obtidas dividindo cada elemento de `dados` por  $n$ . No cálculo do logaritmo da função verossimilhança devemos testar se algum valor gerado é igual a 0, pois neste caso tomamos  $n \log(n) = 0$  levando em conta que  $x \log(x) \rightarrow 0$  quando  $x \downarrow 0$ .

```
emvpi <- dados / n  
logLpi <- colSums(ifelse(dados > 0, dados * log(emvpi), 0))
```

As EMV de  $\pi$  sob  $H_0$  são calculadas com a expressão  $(2F_1 + F_2 + F_3) / (2n)$  aplicada às colunas de dados. Tendo estas estimativas podemos calcular as estimativas das probabilidades e o logaritmo da função verossimilhança  $\log L_{\text{piteta}}$  sob  $H_0$ .

```
emvteta <- apply(dados, 2, function(x) (2 * x[1] + x[2] + x[3]) / (2 * n))
piteta <- rbind(emvteta^2, emvteta * (1 - emvteta), emvteta * (1 - emvteta),
               (1 - emvteta)^2)
logLpiteta <- colSums(dados * log(piteta))
```

Os gráficos da Figura 1 sugerem uma boa aproximação da distribuição assintótica do EMV de  $\theta$ , que é normal com média  $\theta_0$  e variância  $\theta_0(1 - \theta_0) / (2n)$ . A hipótese de normalidade poderia ser formalmente testada (Como?).

```
hist(emvteta, main = "", freq = FALSE, xlab = expression(hat(theta)),
     ylab = "Densidade", cex.axis = 1.5, cex.lab = 1.5)
curve(dnorm(x, teta0, sqrt(0.5 * teta0 * (1- teta0) / n)), add = TRUE,
      col = "red")
box()

plot(ecdf(emvteta), main = "", xlab = expression(hat(theta)),
     ylab = "Função distribuição", pch = "*", cex.axis = 1.5, cex.lab = 1.5)
curve(pnorm(x, teta0, sqrt(0.5 * teta0 * (1- teta0) / n)), add = TRUE,
      col = "red")
```

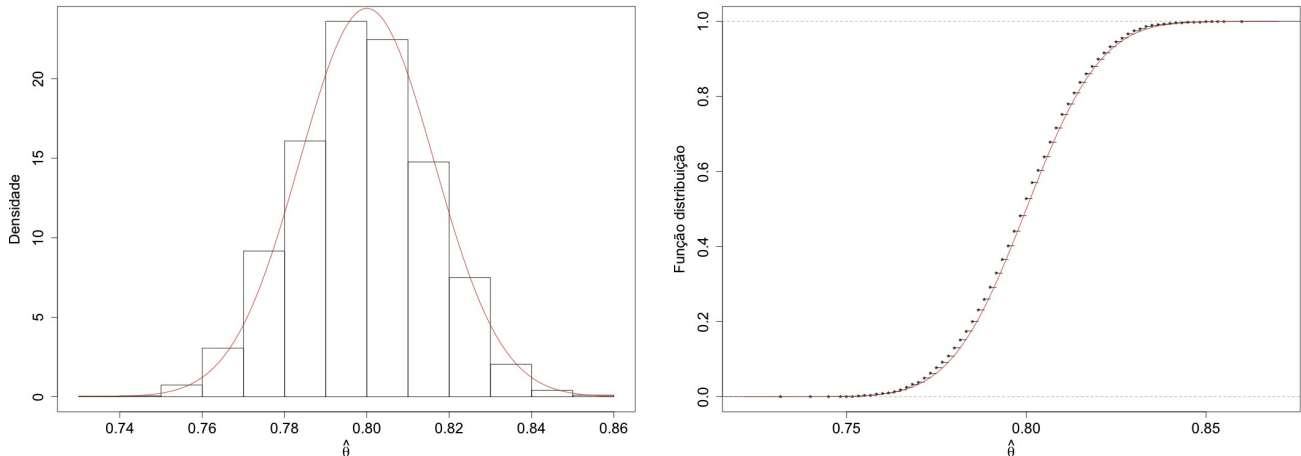


Figura 1. Esquerda: histograma e função densidade teórica. Direita: funções distribuição empírica e teórica.

Os resultados do teste com a estatística  $-2 \log(\Lambda)$  são apresentados em seguida.

```
loglam <- -2 * (logLpiteta - logLpi)
cat("\nResultados\n nível de significância =", alfa, "\n valor crítico =", x2crit)
cat("\n teta =", teta0, "\n pi sob H0 =", c(pi1, pi2, pi2, pi4))
cat("\n n =", n, "\n no. de repetições =", nrep)
cat("\n estatística - 2 log RV:")
cat("\n proporção de rejeição de H0 =", mean(loglam >= x2crit), "\n")
```

```

Resultados
nível de significância = 0.05
valor crítico = 5.991465
teta = 0.8
pi sob H0 = 0.36 0.24 0.24 0.16
n = 200
no. de repetições = 5000
estatística -2 log RV:
proporção de rejeição de H0 = 0.0494

```

Calculamos as frequências esperadas estimadas sob  $H_0$  e realizamos o teste com a estatística  $Q$ .

```

esp <- n * piteta
Q <- colSums((dados - esp)^2 / esp)
cat("\n estatística Q de Pearson:")
cat("\n proporção de rejeição de H0 =", mean(Q >= x2crit), "\n")

```

```

estatística Q de Pearson:
proporção de rejeição de H0 = 0.0504

```

Para este cenário (escolhas de  $\alpha$ ,  $\theta$ ,  $n$  e  $nrep$ ) as proporções de rejeição de  $H_0$  com  $-2 \log(\Lambda)$  e  $Q$  são próximas entre si e também são próximas do valor nominal ( $\alpha = 5\%$ ), indicando uma boa aproximação da distribuição assintótica das duas estatísticas de teste. Os gráficos de quantis da Figura 2 reforçam estas afirmações.

```

qq(rep(c("loglam", "Q"), each = nrep) ~ c(loglam, Q), xlab = "Q",
  ylab = expression(paste("-2 log", Lambda)), pch = 20,
  scales = list(cex = 1.5), main = "(a)")

qqmath(loglam, distribution = function(p) qchisq(p, df = 2), pch = 20,
  ylab = expression(paste("-2 log", Lambda)),
  xlab = expression(paste("Quantis ", chi[2]^2)),
  panel = function(x, ...) {
    panel.qqmathline(x, ...)
    panel.qqmath(x, ...)}, scales = list(cex = 1.5), main = "(b)")

qqmath(Q, distribution = function(p) qchisq(p, df = 2), pch = 20,
  ylab = "Q", xlab = expression(paste("Quantis ", chi[2]^2)),
  panel = function(x, ...) {
    panel.qqmathline(x, ...)
    panel.qqmath(x, ...)
  }, scales = list(cex = 1.5), main = "(c)")

```

## 2. Exemplo

Em uma amostra de  $n = 215$  observações as frequências são  $f_1 = 19$ ,  $f_2 = 62$ ,  $f_3 = 90$  e  $f_4 = 44$ .

```

dados <- c(19, 62, 90, 44)
n <- sum(dados)

```

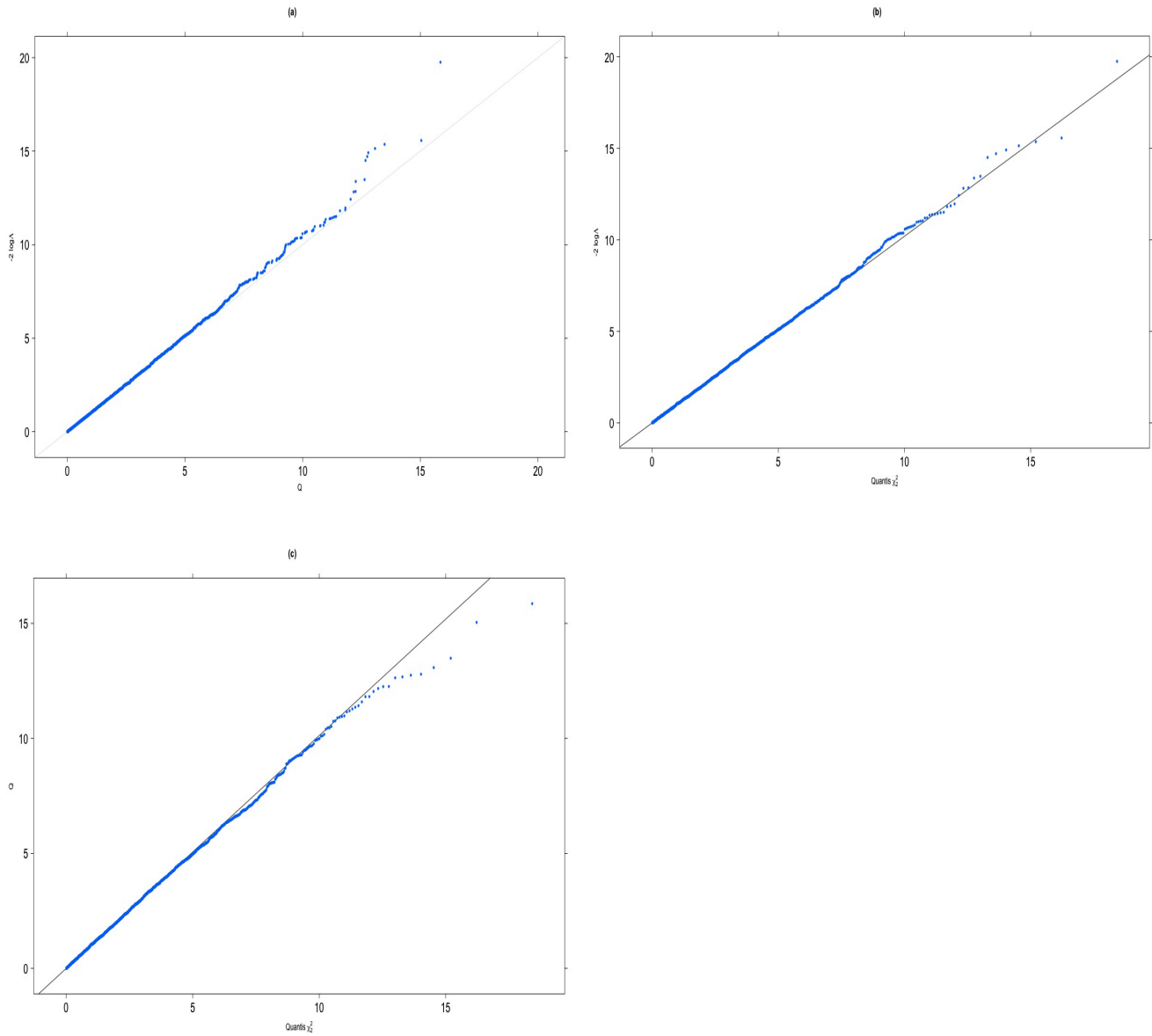


Figura 2. Gráficos de quantis. (a)  $-2 \log(\Lambda)$  e  $Q$ , (b)  $-2 \log(\Lambda)$  e (c)  $Q$ .

A EMV de  $\theta$  é apresentada abaixo.

```
emvteta = (2 * dados[1] + dados[2] + dados[3]) / (2 * n)
cat("\n dados:", dados)
cat("\n n =", n, "\n emv teta =", emvteta)
```

```
dados: 19 62 90 44
n = 215
emv teta = 0.4418605
```

O gráfico da função verossimilhança é mostrado na Figura 3.

```
logver <- function(theta) {  
  f123 * log(theta) + f234 * log(1 - theta)  
}  
  
f123 <- 2 * dados[1] + dados[2] + dados[3]  
f234 <- 2 * dados[4] + dados[2] + dados[3]  
  
maxlogver <- logver(emvteta)  
par(mai = c(1.2, 1.3, 0.1, 0.1))  
curve(logver, 0, 1, cex.lab = 1.5, cex.axis = 1.5, xlab =  
  expression(theta), ylab = expression(paste("log L(", theta, ")")))  
points(emvteta, maxlogver, pch = 20, col = "red")  
abline(h = maxlogver, lty = 2, col = "red")  
abline(v = emvteta, lty = 2, col = "red")
```

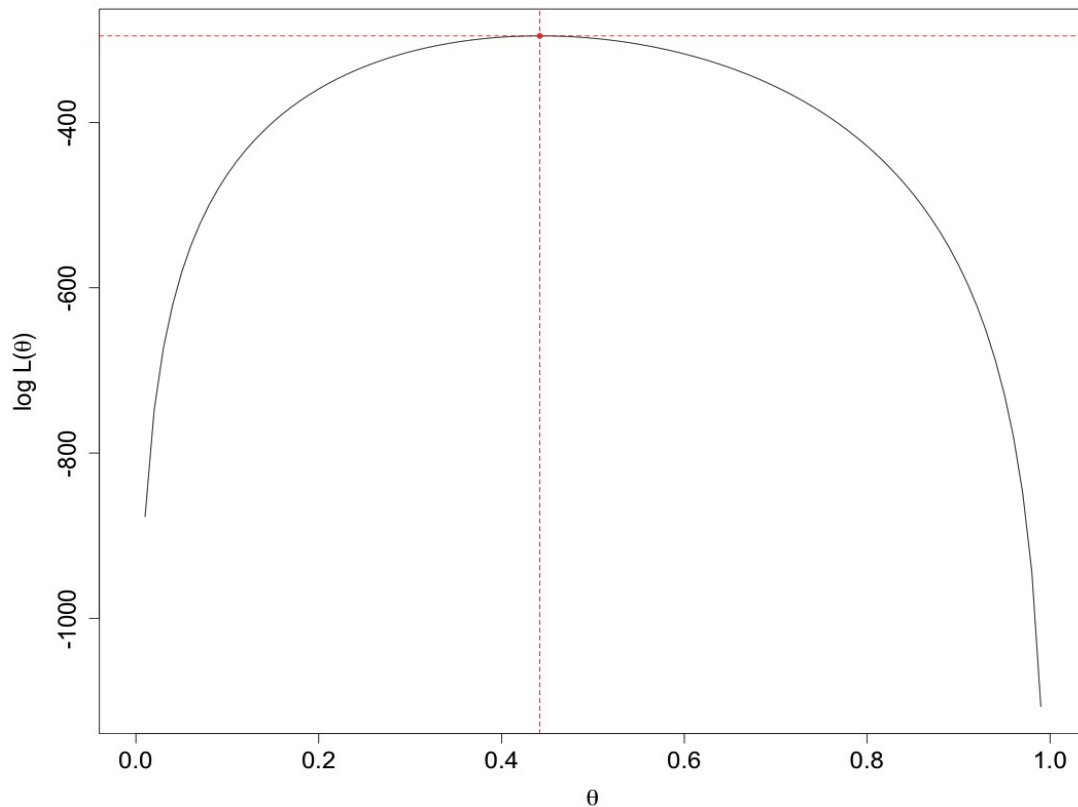


Figura 3. Função log-verossimilhança.

Por último realizamos o teste da hipótese na eq. (1).

```
emvpi <- dados / n
logLpi <- sum(ifelse(dados > 0, dados * log(emvpi), 0))
piteta <- c(emvteta^2, emvteta * (1 - emvteta), emvteta * (1 - emvteta),
  (1 - emvteta)^2)
logLpiteta <- sum(dados * log(piteta))
loglam <- -2 * (logLpiteta - logLpi)

esp <- n * piteta
Q <- sum((dados - esp)^2 / esp)

cat("\n -2 log RV =", loglam, "(p =", pchisq(loglam, 2, lower.tail =
  FALSE), ")")
cat("\n Q =", Q, "(p =", pchisq(Q, 2, lower.tail = FALSE), ")")

-2 log RV = 47.53287 (p = 4.768353e-11 )
Q = 47.7652 (p = 4.24539e-11 )
```

Neste exemplo os valores de  $-2 \log(\Lambda)$  e  $Q$  são próximos. Ambas as estatísticas de teste indicam diferenças significativas em relação à hipótese formulada ( $p < 0,0001$ ).

O cálculo da estatística  $Q$  de Pearson pode ser realizado com a função `chisq.test` em R utilizando o EMV do vetor de probabilidades sob  $H_0$  (`piteta`). Observe que o valor- $p$  refere-se ao teste de  $H_0$  simples com  $k - 1 = 4 - 1 = 3$  graus de liberdade ( $df = 3$ ).

```
(chisq.test(dados, p = piteta))
```

```
Chi-squared test for given probabilities
```

```
data: dados
X-squared = 47.7652, df = 3, p-value = 2.389e-10
```