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

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

Para realizar o teste utilizamos as estatísticas $G^2 = -2$ log(razão de versossimilhanças) e X^2 de Pearson. Inicialmente carregamos o pacote `lattice`, que inclui funções para os gráficos de quantis.

```
library(lattice)
```

Escolhemos o nível de significância nominal α e calculamos o valor crítico obtido da distribuição de referência χ^2 com 2 g.l. (3 – 1 = 2 g.l.).

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

```
[1] 5.991465
```

Escolhendo o verdadeiro valor de θ ($= \theta_0$) calculamos as probabilidades (π) sob a hipótese (1).

```
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 de π e o logaritmo da função verossimilhança `logLpi` (a menos de uma constante aditiva) são calculados por meio de funções matriciais. Conforme visto em sala de aula, as EMV irrestritas de π 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 $x \log(x) \to 0$ quando $x \downarrow 0$.

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

As contagens são denotadas por n_1, n_2, n_3 e n_4, com total $n = n_1 + n_2 + n_3 + n_4$. O estimador de máxima verossimilhança de θ sob a hipótese (1) é calculado abaixo.

A função verossimilhança é dada por

$$\begin{aligned} L(\theta) &\propto \theta^{2n_1}\{\theta(1-\theta)\}^{n_2}\{\theta(1-\theta)\}^{n_3}(1-\theta)^{n_4} \\ &= \theta^{2n_1}\{\theta(1-\theta)\}^{n_2+n_3}(1-\theta)^{n_4} \\ &= \theta^{2n_1+n_2+n_3}\,(1-\theta)^{n_2+n_3+2n_4}, \end{aligned}$$

em que a constante de proporcionalidade na primeira linha não envolve θ (escreva esta constante). Tomando logaritmo e omitindo uma constante aditiva obtemos

$$\ell(\theta) = (2n_1 + n_2 + n_3)\log(\theta) + (n_2 + n_3 + 2n_4)\log(1-\theta),$$

cuja derivada em relação a θ é

$$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{2n_1 + n_2 + n_3}{\theta} + \frac{n_2 + n_3 + 2n_4}{1-\theta} \times (-1). \qquad (2)$$

Igualando a derivada em (2) a 0 obtemos sucessivamente

$$\begin{aligned} \frac{2n_1 + n_2 + n_3}{\theta} &= \frac{n_2 + n_3 + 2n_4}{1-\theta}, \\ (2n_1 + n_2 + n_3)(1-\theta) &= (n_2 + n_3 + 2n_4)\theta, \\ 2n_1 + n_2 + n_3 - (2n_1 + n_2 + n_3)\theta &= (n_2 + n_3 + 2n_4)\theta, \\ 2n_1 + n_2 + n_3 &= (2n_1 + n_2 + n_3 + n_2 + n_3 + 2n_4)\theta \quad \text{e} \\ \theta &= \frac{2n_1 + n_2 + n_3}{2n}. \end{aligned} \qquad (3)$$

A partir da expressão (2) calculamos

$$\frac{\partial^2}{\partial\theta^2}\ell(\theta) = -\left\{\frac{2n_1 + n_2 + n_3}{\theta^2} + \frac{n_2 + n_3 + 2n_4}{(1-\theta)^2}\right\}, \qquad (4)$$

que assume somente valores negativos, para $\theta \in (0,1)$. Portanto, a solução em (3) maximiza $L(\theta)$, de modo que $\widehat{\theta} = (2n_1 + n_2 + n_3)/(2n)$.

A distribuição assintótica do EMV de θ é normal com média θ_0 e variância $\theta_0\,(1 - \theta_0) / (2n)$ calculada abaixo.

Levando em conta que as distribuições marginais dos componentes de um vetor com distribuição multinomial são binomiais, de acordo com a expressão (1) temos

$$\begin{aligned} &n_1 \sim \text{binomial}\big(n, \theta^2\big),\ n_2 \sim \text{binomial}\big(n, \theta(1-\theta)\big), \\ &n_3 \sim \text{binomial}\big(n, \theta(1-\theta)\big) \quad \text{e} \quad n_4 \sim \text{binomial}\big(n, (1-\theta)^2\big). \end{aligned} \qquad (5)$$

Com os resultados em (5) e a expressão (4) calculamos a informação de Fisher de θ, dada por

$$
\begin{aligned}
-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] &= \frac{2n\theta^2 + n\theta(1-\theta) + n\theta(1-\theta)}{\theta^2} \\
&\quad + \frac{n\theta(1-\theta) + n\theta(1-\theta) + 2n(1-\theta)^2}{(1-\theta)^2} \\
&= 2n\left\{\frac{\theta^2 + \theta(1-\theta)}{\theta^2} + \frac{\theta(1-\theta) + (1-\theta)^2}{(1-\theta)^2}\right\} \\
&= 2n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) = \frac{2n}{\theta(1-\theta)}.
\end{aligned}
$$

Portanto, para $\theta = \theta_0$, a variância assintótica é $\theta_0(1-\theta_0)/(2n)$.

Com a expressão $(2n_1 + n_2 + n_3) / (2n)$ aplicada às colunas de `dados` calculamos as EMV de θ. Tendo estas estimativas podemos calcular as estimativas das probabilidades e o logaritmo da função verossimilhança `logLpiteta` sob a hipótese (1).

```
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 θ. A hipótese de normalidade poderia ser formalmente testada (Como? Efetue o teste).

```
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")
```

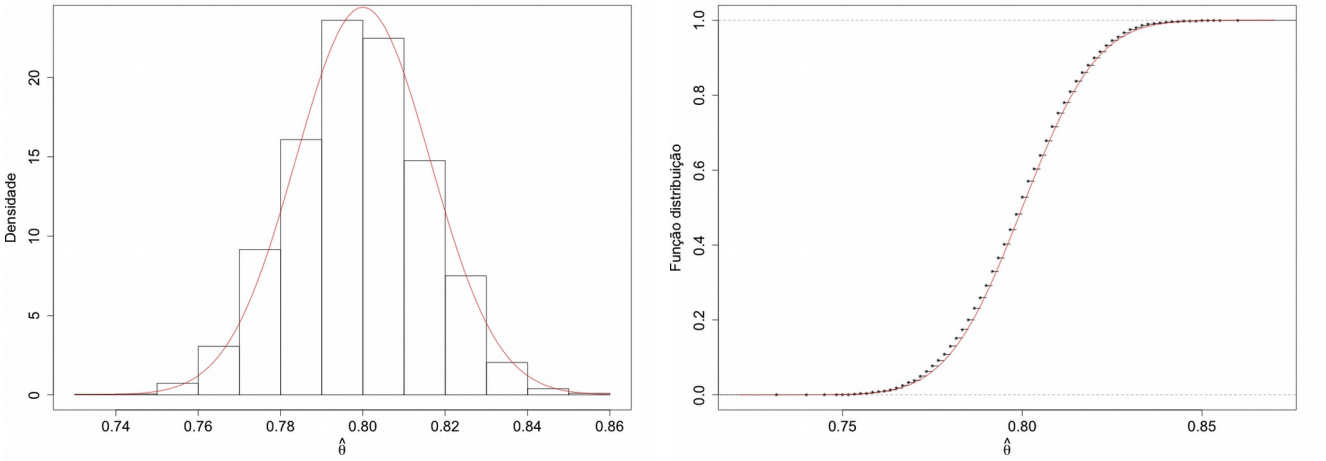

Figura 1. Esquerda: histograma e função densidade teórica. Direita: funções distribuição empírica e teórica.

Os resultados do teste com a estatística G^2 são apresentados em seguida.

```
G2 <- -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 =", M)
cat("\n estatística G2:")
cat("\n proporção de rejeição de H0 =", mean(G2 >= 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 G2:
 proporção de rejeição de H0 = 0.0494
```

Em seguida calculamos as frequências esperadas estimadas sob a hipótese (1) e realizamos o teste com a estatística X^2.

```
esp <- n * piteta
X2 <- colSums((dados - esp)^2 / esp)
cat("\n estatística X2 de Pearson:")
cat("\n proporção de rejeição de H0 =", mean(X2 >= x2crit), "\n")
```

```
estatística X2 de Pearson:
proporção de rejeição de H0 = 0.0504
```

Para este cenário (escolhas de α, θ, *n* e *nrep*) as proporções de rejeição da hipótese (1) com G^2 e X^2 são próximas entre si e também são próximas do valor nominal (α = 5%), indicando uma boa aproximação da distribuição assintótica das duas estatísticas de teste. Os gráficos de quantis abaixo 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))
```

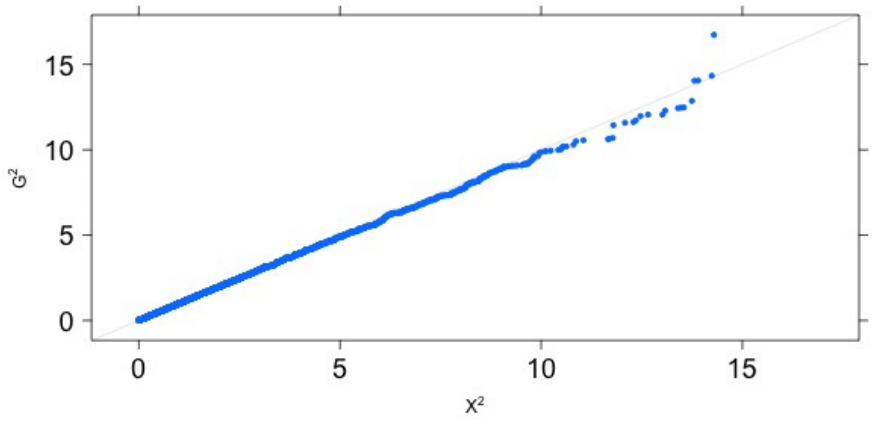

Figura 2. Gráfico de dispersão de X^2 e G^2.

```
qqmath(G2, distribution = function(p) qchisq(p, df = 2), pch = 20,
       ylab = expression(G^2),
       xlab = expression(paste("Quantis ", chi[2]^2)),
       panel = function(x, ...) {
          panel.qqmathline(x, ...)
          panel.qqmath(x, ...)
       }, scales = list(cex = 1.5))
```

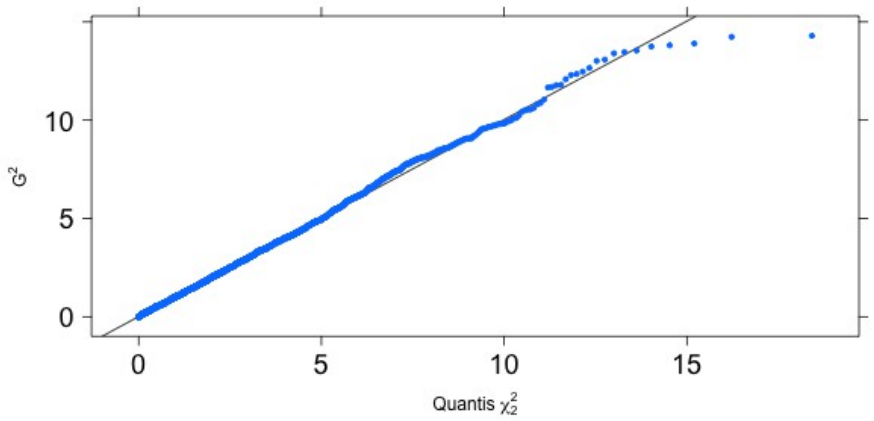

Figura 3. Gráficos de quantis de G^2.

```
qqmath(X2, distribution = function(p) qchisq(p, df = 2), pch = 20,
       ylab = expression(X^2), xlab = expression(paste("Quantis ", chi[2]^2)),
       panel = function(x, ...) {
          panel.qqmathline(x, ...)
          panel.qqmath(x, ...)
       }, scales = list(cex = 1.5))
```

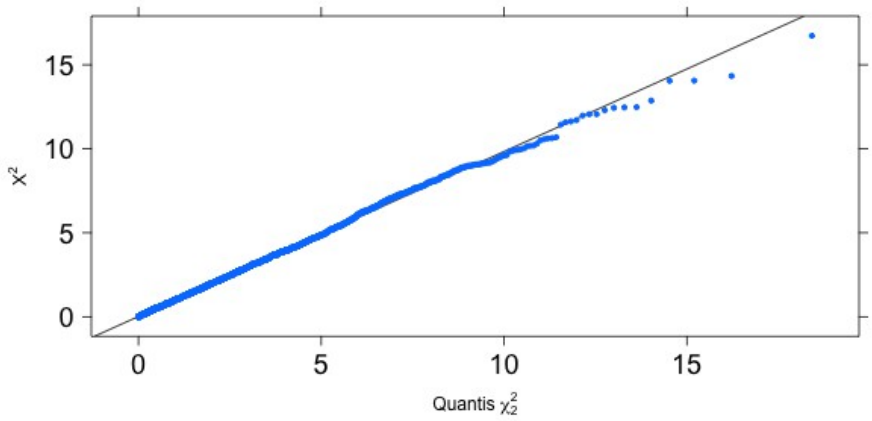

Figura 4. Gráficos de quantis de X^2.

Nota 1. Refaça as simulações com α = 0,01 e 0,10.

2. Exemplo

Em uma amostra de $n = 215$ observações as frequências observadas (contagens) são $n_1 = 19$, $n_2 = 62$, $n_3 = 90$ e $n_4 = 44$.

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

A EMV de θ 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 log-verossimilhança é mostrado na Figura 5.

```
logver <- function(theta) {
   n123 * log(theta) + n234 * log(1 - theta)
}

n123 <- 2 * dados[1] + dados[2] + dados[3]
n234 <- 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")
```

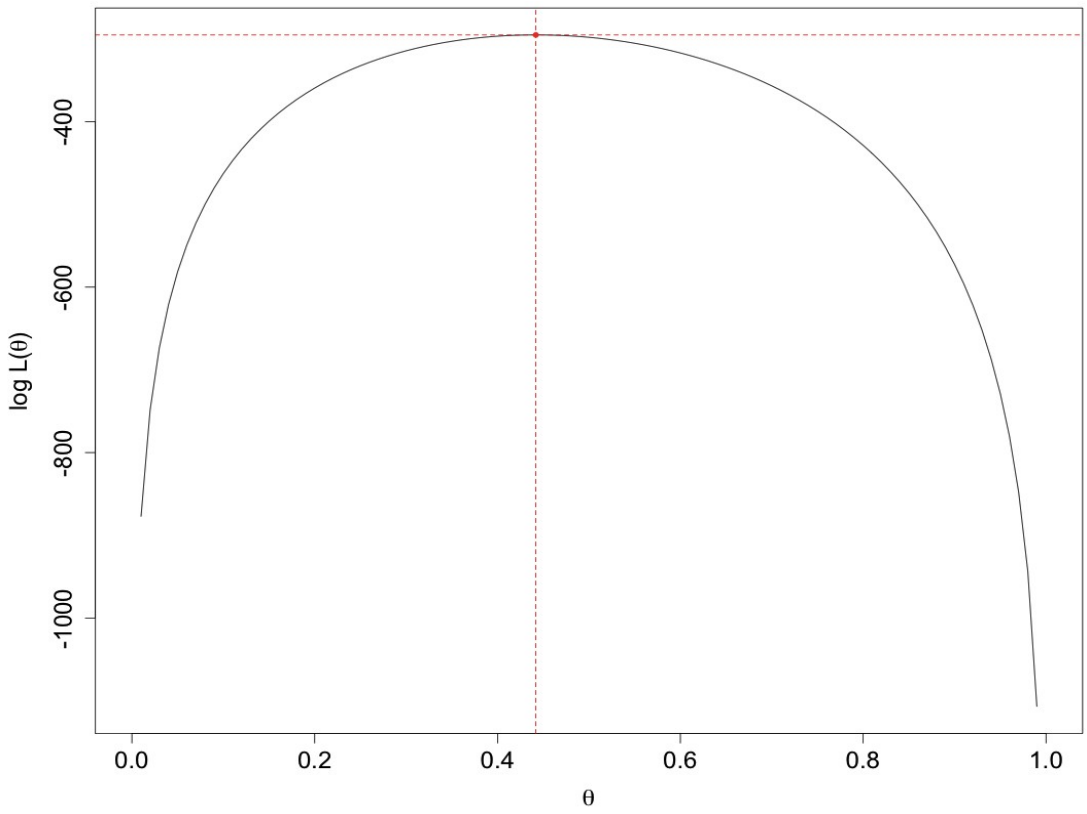

Figura 5. Função log-verossimilhança de θ.

Por último realizamos o teste da hipótese na expressão (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))
G2 = -2 * (logLpiteta - logLpi)

esp <- n * piteta
X2 <- sum((dados - esp)^2 / esp)

cat("\n Frequências observadas:", dados,
    "\n Frequências esperadas estimadas:", round(esp))
```

```
   Frequências observadas: 19 62 90 44
   Frequências esperadas estimadas: 42 53 53 67
```

```
cat("\n G2 =", G2, "(p =", pchisq(G2, 2, lower.tail = FALSE), ")")
cat("\n X2 =", X2, "(p =", pchisq(X2, 2, lower.tail = FALSE), ")")
```

```
   G2 = 47.53287 (p = 4.768353e-11 )
   X2 = 47.7652 (p = 4.24539e-11 )
```

Neste exemplo os valores de G^2 e X^2 são próximos. Ambas as estatísticas de teste indicam diferenças significativas em relação à hipótese formulada na expressão (1) ($p < 0{,}0001$).

O cálculo em si da estatística X^2 de Pearson pode ser realizado com a função `chisq.test` em R utilizando o EMV do vetor de probabilidades sob a hipótese em (1) (`piteta`). Observe que o valor-*p* refere-se ao teste em que a hipótese é simples com 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
```