Trabalho 4. Simulações e exemplo

1. Simulações

São apresentados os passos para a geração de amostras em linguagem R e, a partir destas, o teste da hipótese da questão 3(c) do terceiro trabalho. Para realizar o teste utilizamos as estatísticas G^2 (baseada na razão de verossimilhanças) e X^2 de Pearson. Os resultados são destacados em cor azul.

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.).

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

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

```
teta0 = 0.8
pi11 = teta0^2
pi12 = teta0 * (1 - teta0) # = pi21
pi22 = (1 - teta0)^2
```

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

```
n = 200
M = 5000
```

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

```
dados = rmultinom(M, size = n, prob = c(pi11, pi12, pi12, pi22))
```

As estimativas de máxima verossimilhanças (EMV) irrestritas (ou seja, sob H_1) de π e o logaritmo da função verossimilhança `logLpi` (a menos de uma constante aditiva) são calculados por meio de funções matriciais. As EMV 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 $x \log(x) \rightarrow 0$ quando $x \downarrow 0$.

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

As EMV de π sob H_0 são calculadas com a expressão $(2n_{11} + n_{12} + n_{21}) / (2n)$ aplicada às colunas de `dados`. Tendo estas estimativas podemos calcular as estimativas das probabilidades e o logaritmo da função verossimilhança `logLpiteta` 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 θ, que é normal com média θ_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")
```

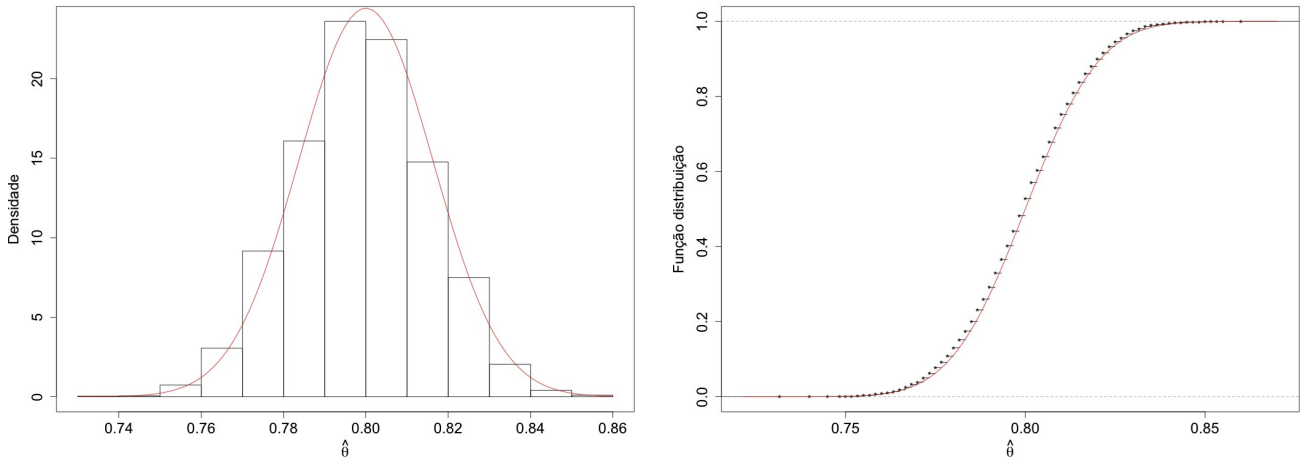

Figura 1. Esquerda: histograma e função densidade teórica. Direita: funções distribuição empírica e teórica.

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

```
G2 = 2 * (logLpi - logLpiteta)
cat("\nResultados\n nível de significância =", alfa, "\n valor crítico =", x2crit)
cat("\n teta =", teta0, "\n pi sob H0 =", c(pi11, pi12, pi12, pi22))
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
```

Calculamos as frequencias esperadas estimadas sob H_0 e realizamos o teste com a estatística X^2.

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

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

Para este cenário (escolhas de α, θ, n e M) as proporções de rejeição de H_0 com G^2 e X^2 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("G2", "X2"), each = M) ~ c(G2, X2), xlab = expression(X^2),
   ylab = expression(G^2), pch = 20, scales = list(cex = 1.5), main = "(a)")

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), main = "(b)")

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), main = "(c)")
```

2. Exemplo

Em uma amostra de $n = 215$ observações as contagens são $n_{11} = 19$, $n_{12} = 62$, $n_{21} = 90$ e $n_{22} = 44$.

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

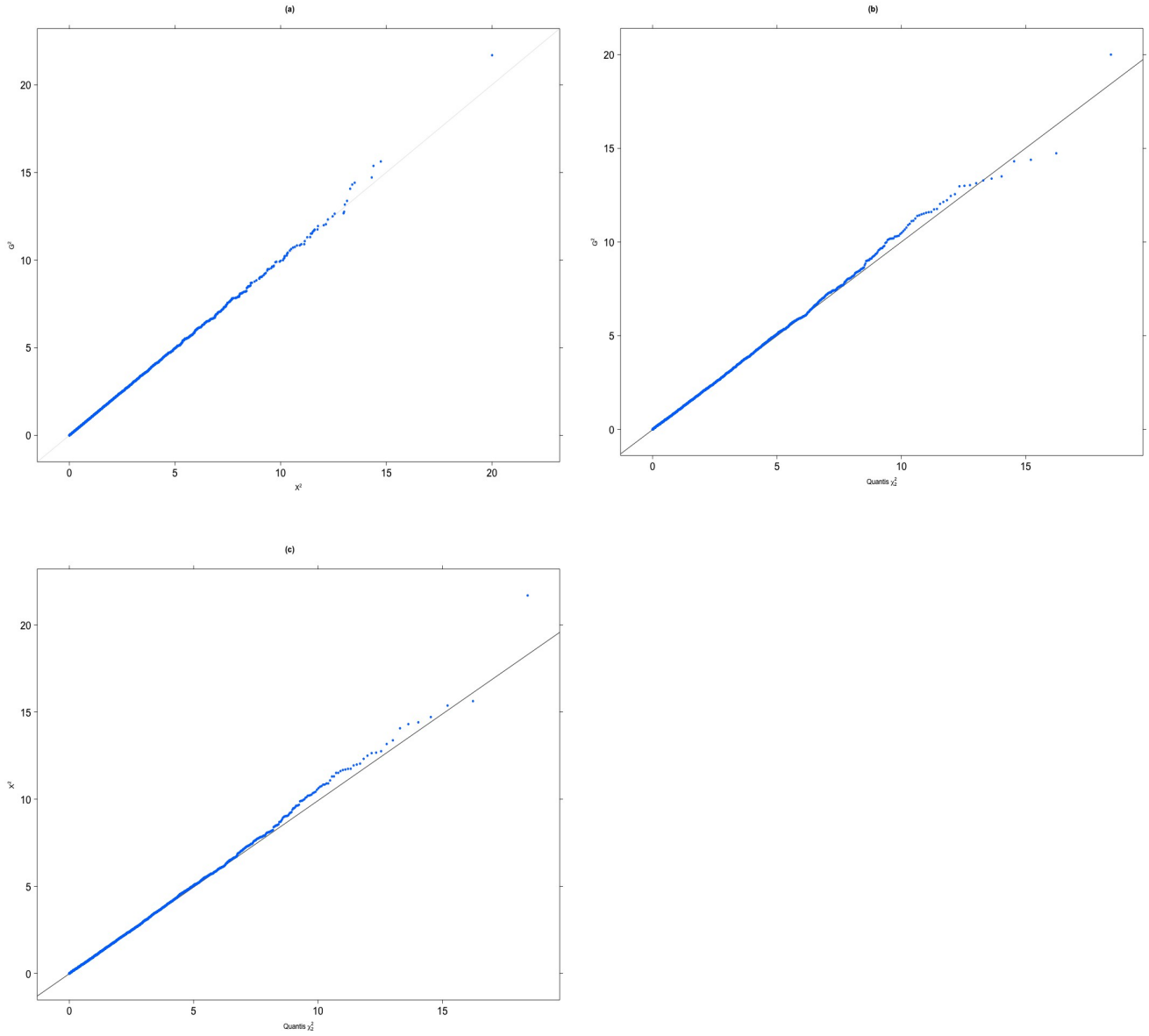

Figura 2. Gráfico de quantis. (a) G^2 e X^2 (b) G^2 e (c) X^2.

A EMV de θ é apresentada abaixo.

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

```
dados:
     [,1] [,2]
[1,]   19   62
[2,]   90   44
 n = 215
 emv teta = 0.4418605
```

O gráfico da função verossimilhança é mostrado na Figura 3.

```
logver = function(theta) {
   n11s * log(theta) + n22s * log(1 - theta)
}

n11s = 2 * dados[1] + dados[2] + dados[3]
n22s = 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")
```

Por último realizamos o teste da hipótese da questão 3(c) do terceiro trabalho.

```
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 * (logLpi - logLpiteta)

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

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 ($p < 0{,}0001$).

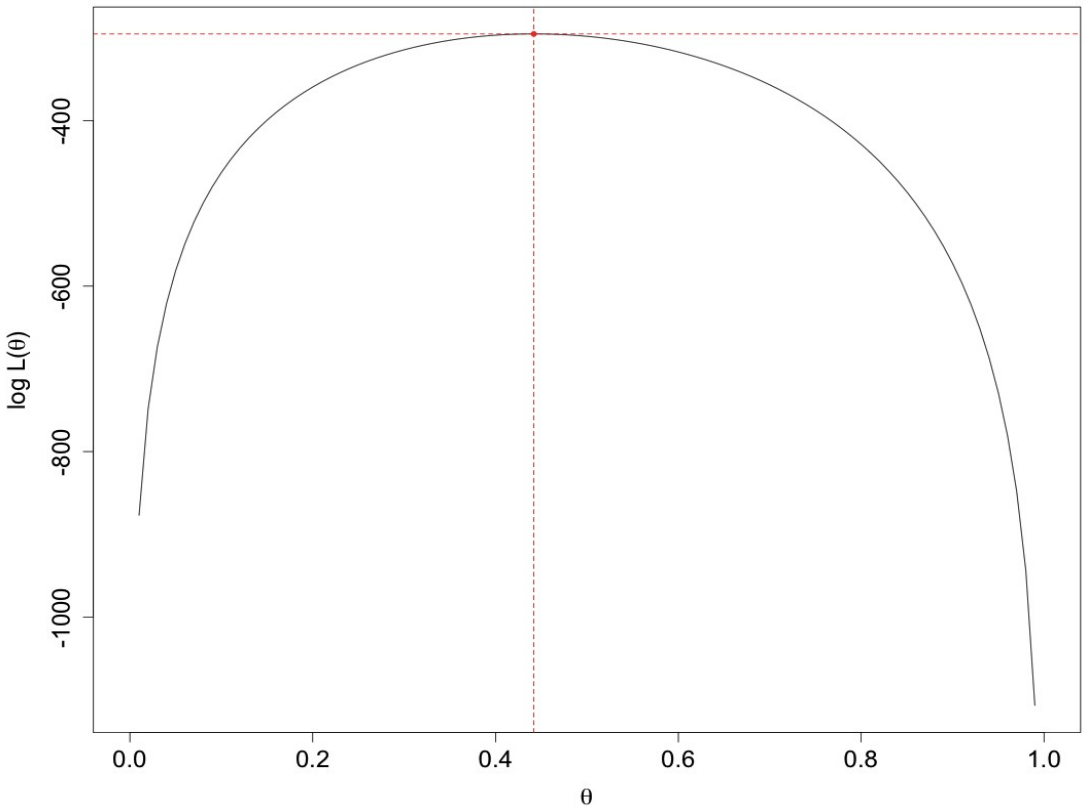

Figura 3. Função log-verossimilhança.