Verificação da suposição de normalidade

(linguagem R)

2026

Em uma amostra aleatória com observações $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, o vetor de médias e a matriz de covariâncias amostral são denotados por $\overline{\boldsymbol{x}}$ e $\boldsymbol{S}$, respectivamente, sendo que $\overline{x}_j = \sum_{i=1}^n x_{ij}/n, \quad j = 1, \ldots, p$ e $\boldsymbol{S} = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top/(n-1)$.

Se $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ é uma amostra aleatória da distribuição $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então $D_1^2, \ldots, D_n^2$ são variáveis aleatórias independentes com distribuição χ^2_p, em que

$$D_i^2 = (\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}), \ i = 1, \ldots, n \tag{1}$$

é o quadrado da distância estatística ou distância de Mahalanobis entre $\boldsymbol{X}_i$ e $\boldsymbol{\mu}$.

Com os dados da amostra calculamos

$$d_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \ i = 1, \ldots, n, \tag{2}$$

que são ordenados resultando em

$$d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2. \tag{3}$$

Os valores em (3) são comparados aos respectivos quantis da distribuição χ^2_p como forma de avaliar a suposição de normalidade. Em linguagem R é utilizada a aproximação

$$\mathrm{E}(D_{(i)}^2) \simeq \begin{cases} \text{quantil } \dfrac{i - 3/8}{n + 1/4} \text{ de } \chi^2_p, & \text{se } n \leq 10, \\ \text{quantil } \dfrac{i - 1/2}{n} \text{ de } \chi^2_p, & \text{se } n > 10, \end{cases} \quad i = 1, \ldots, n. \tag{4}$$

As probabilidades em (4) sào obtidas com a função `ppoints` e os quantis são calculados com a função `qchisq`. O padrão de referência no gráfico de dispersão dos n pares de pontos em (3) e (4) é a reta identidade.

Em seguida efetuamos o cálculo em (2) de forma matricial. Iniciamos com a decomposição de Choleski de $\boldsymbol{S}$ dada por $\boldsymbol{S} = \boldsymbol{C}^\top \boldsymbol{C}$, em que $\boldsymbol{C}$ é uma matriz triangular superior chamada de "raiz quadrada" de $\boldsymbol{S}$. Desta forma $\boldsymbol{S}^{-1} = \boldsymbol{C}^{-1}\boldsymbol{C}^{-\top}$ e

$$\begin{aligned} d_i^2 &= (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \boldsymbol{C}^{-1}\boldsymbol{C}^{-\top} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) \\ &= \{\boldsymbol{C}^{-\top}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\}^\top \boldsymbol{C}^{-\top} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \ i = 1, \ldots, n. \end{aligned} \tag{5}$$

Os vetores p-dimensionais $\boldsymbol{C}^{-\top}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ em (5), $i = 1, \ldots, n$, são organizados na matriz $\boldsymbol{B}$ $p \times n$. Finalmente, notamos que d_i^2 em (5) corresponde à soma dos quadrados dos elementos na i-ésima coluna de $\boldsymbol{B}$.

```
## Conjunto de dados USArrests
? USArrests
```

```
USArrests {datasets}    R Documentation
Violent Crime Rates by US State

Description
This data set contains statistics, in arrests per 100,000 residents for assault, murder,
```

and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Format
A data frame with 50 observations on 4 variables.

[,1] Murder numeric Murder arrests (per 100,000)
[,2] Assault numeric Assault arrests (per 100,000)
[,3] UrbanPop numeric Percent urban population
[,4] Rape numeric Rape arrests (per 100,000)

Source
World Almanac and Book of facts 1975. (Crime rates).

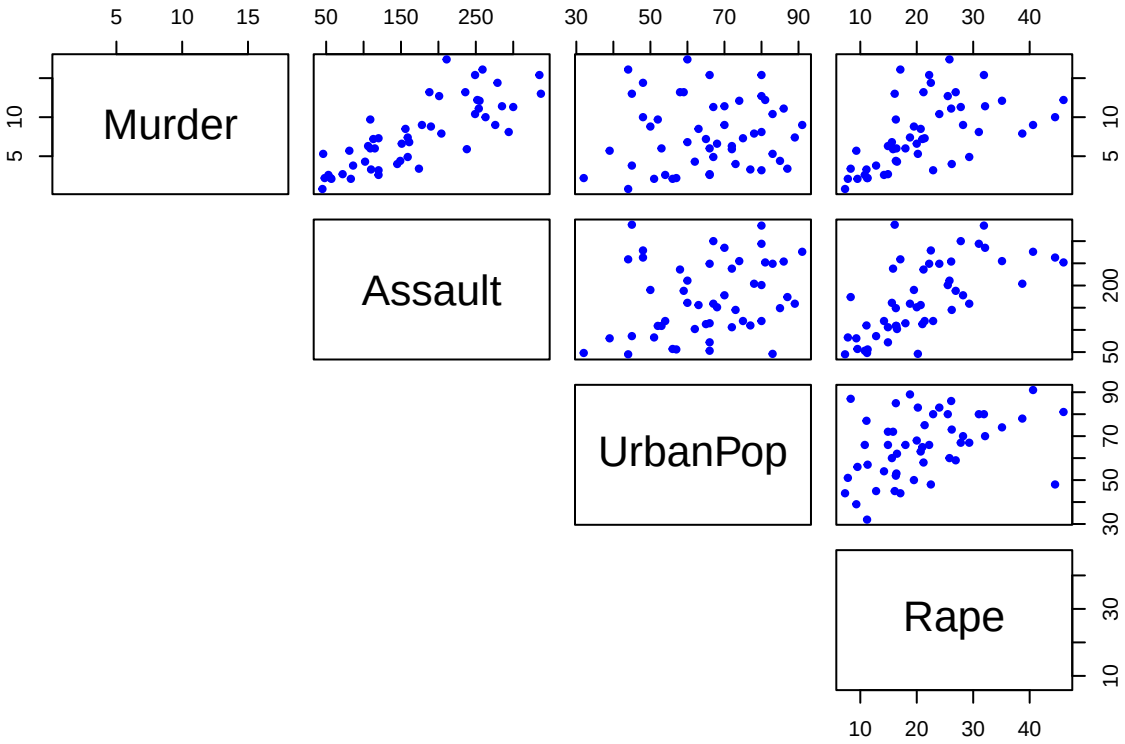
Statistical Abstracts of the United States 1975, p.20, (Urban rates), possibly available as https://books.google.ch/books?id=zl9qAAAAMAAJ&pg=PA20.

References
McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.

```
names(USArrests)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

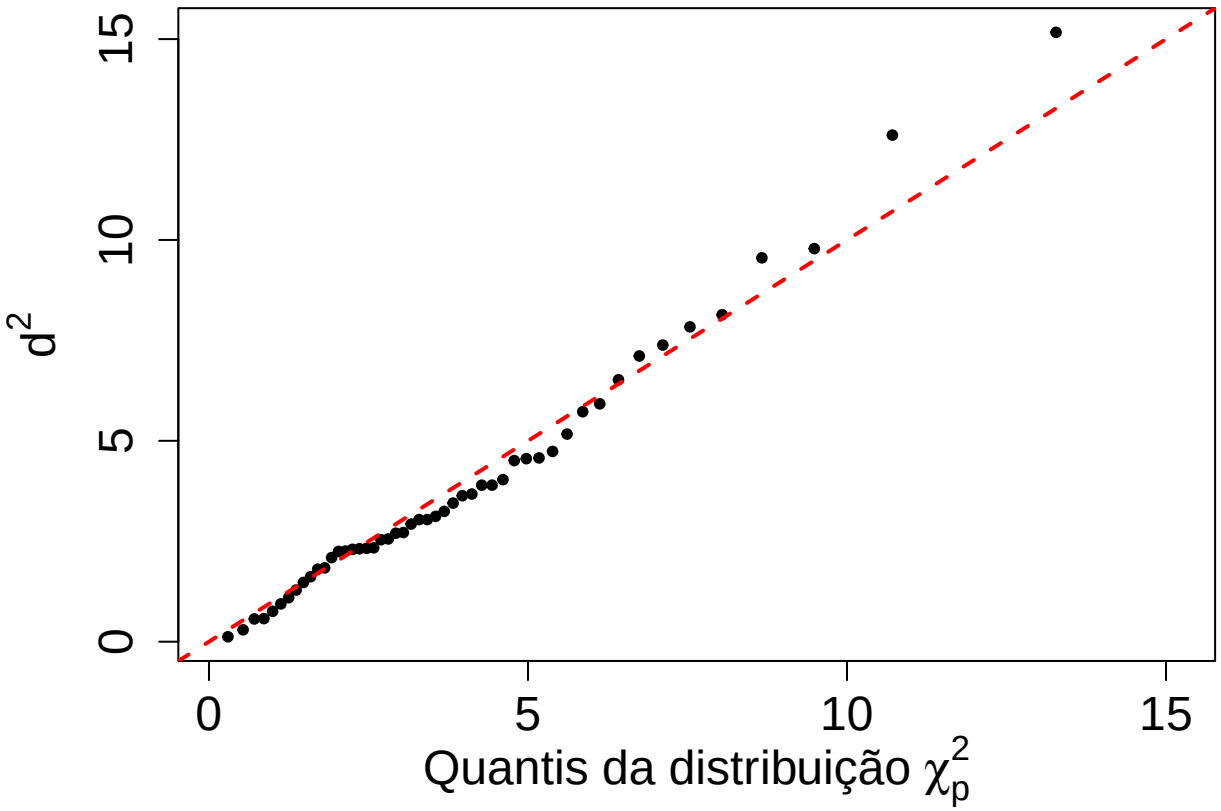
```
dados <- USArrests

pairs(dados, pch = 20, col = "blue", lower.panel = NULL)
```




```
n <- nrow(dados)
p <- ncol(dados)
cat("\n n =", n, ", p =", p, "\n")
```

```
##
##  n = 50 , p = 4
```

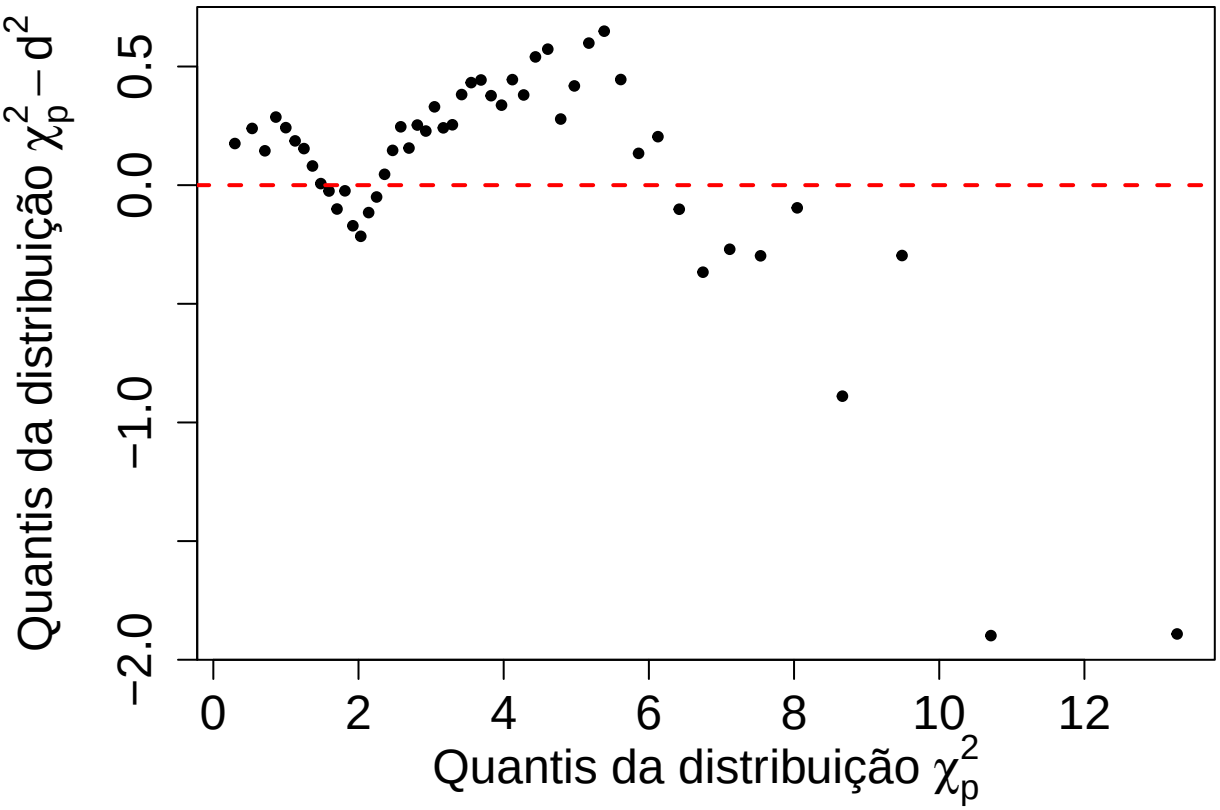
```r
## Posições de plotagem
pplot <- ppoints(n)

## Quantis teóricos
xq2 <- qchisq(pplot, df = p)

## Estimadores
Xb <- colMeans(dados)  # média
S <- cov(dados)        # matriz de covariâncias

## Quadrado da distância estatística
difmedia <- t(t(dados) - Xb) # X - Xbarra
C <- chol(S)  # "raiz quadrada" de S
B <- t(solve(C)) %*% t(difmedia)
d2 <- sort(colSums(B^2))

# Gráfico
faixa <- range(d2, xq2)
par(mai = c(1.0, 1.0, 0.1, 0.1))  # Margens (em pol)
plot(xq2, d2, xlab = expression(paste("Quantis da distribuição ",chi[p]^2)),
     xlim = faixa, ylim = faixa, ylab = expression(d^2), pch = 20,
     cex.lab = 1.5, cex.axis = 1.5)
abline(0, 1, lty = 2, lwd = 2, col = "red")
```



Os pontos no gráfico de quantis (gráfico QQ) acima estão próximos à reta identidade, com apenas dois mais afastados. Há indicação de boa aproximação pela distribuição normal.

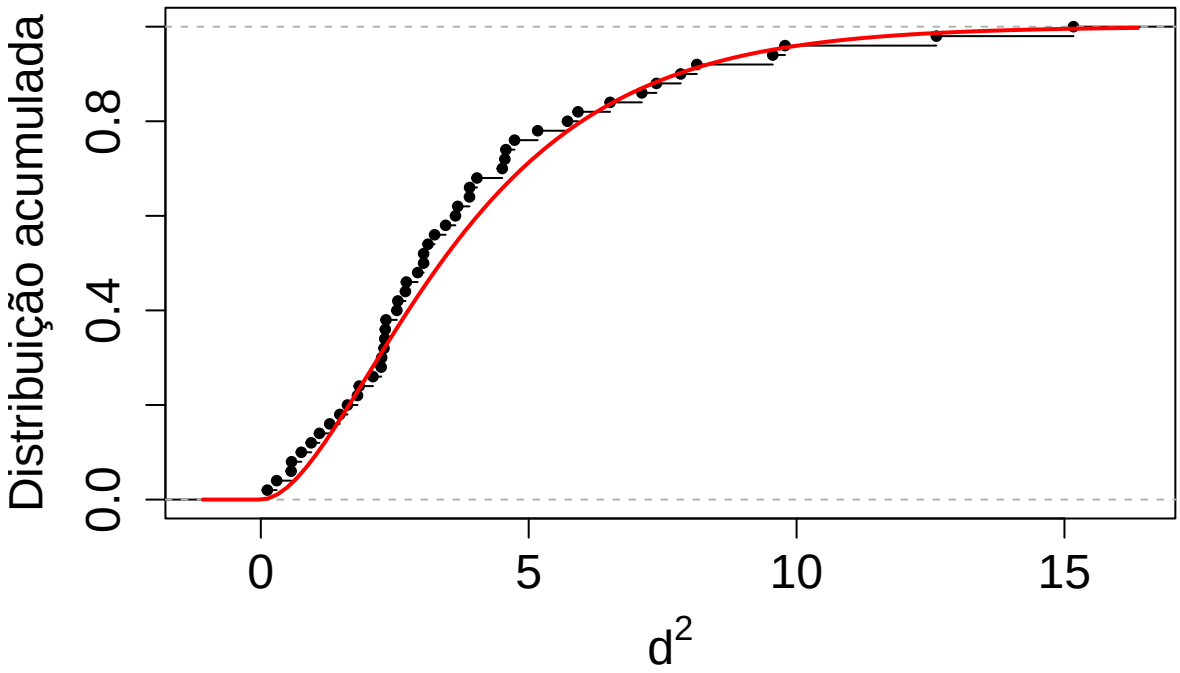
Representando no eixo vertical as diferenças entre os quantis em (4) e $d^2_{(i)}$, $i = 1, \ldots, n$, obtemos um gráfico em que a referência é a reta horizontal correspondente à diferença nula.

```
par(mai = c(1.0, 1.1, 0.1, 0.1))  # Margens (em pol)
plot(xq2, xq2 - d2, xlab = expression(paste("Quantis da distribuição ",chi[p]^2)),
    ylab = expression(paste("Quantis da distribuição ",chi[p]^2) - d^2), pch = 20,
     cex.lab = 1.5, cex.axis = 1.5)
abline(h = 0, lty = 2, lwd = 2, col = "red")
```



Em seguida comparamos a função distribuição empírica de d^2_i, $i = 1, \ldots, n$, com a função distribuição de uma variável aleatória χ^2_p (com $p = 4$).

```
plot(ecdf(d2), main = "", xlab = expression(d^2),  pch = 20,
     ylab = "Distribuição acumulada", cex.lab = 1.5, cex.axis = 1.5)
curve(pchisq(x, df = p), add = TRUE, col = "red", lwd = 2)
```



De forma aproximada, a hipótese de normalidade pode ser testada com o teste de Kolmogorov-Smirnov.

```
ks.test(d2, "pchisq", df = p)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  d2
## D = 0.08139, p-value = 0.8682
## alternative hypothesis: two-sided
```

Observação. Testes de normalidade mais formais podem ser encontrados no pacote `MVN` (vide https://cran.r-project.org/web/packages/MVN/MVN.pdf).