Inferência para a Distribuição Normal Multivariada: parte 1

Prof. Caio Azevedo

Introdução

- Estudaremos como realizar Inferência Estatística (estimação pontual, intervalar e testes de hipótese) para os parâmetros da distribuição normal multivariada.
- Veremos abordagens para cada componente univariada do vetor de médias ($\boldsymbol{\mu}$) e matriz e variâncias e covariâncias ($\boldsymbol{\Sigma}$), bem como para cada um dos parâmetros.
- Um dos elementos básicos é a matriz de dados em que as linhas ($\boldsymbol{X}_i = (X_{i1}, ..., X_{ip})'$) são consideradas vetores aleatórios independentes e identicamente distribuídos segundo uma distribuição $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Matriz de dados (MD)

Indivíduo	Variável 1	Variável 2	...	Variável p
1	X_{11}	X_{12}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	X_{n1}	X_{n2}	...	X_{np}

Temos que (cada linha da MD) $\boldsymbol{X}_i = (X_{i1}, ..., X_{ip})' \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, 2, ..., n$.

Verossimilhança

- Com a suposição de amostra aleatória (aa) da $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ temos que a verossimilhança pra uma aa de tamanho n é dada por

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-0,5\left[\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{x}_i - \boldsymbol{\mu})\right]\right\} \quad (1)$$

- Além disso, note que, para a fdp da $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, temos que

$$\begin{aligned}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) &= tr[(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})] \\ &= tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})']. \quad (2)\end{aligned}$$

Verossimilhança

- Note, agora, que (somando-se e subtraindo-se $\overline{\boldsymbol{x}}$ em cada "()" na Equação (1)), temos que (exercício)

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})' = \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})' + n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})' \quad (3)$$

em que $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \frac{1}{n}\left(\sum_{i=1}^{n}x_{i1}, ..., \sum_{i=1}^{n}x_{ip}\right)' = (\overline{x}_1, ..., \overline{x}_p)'$

(note que $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n}x_{ij}$, j=1,2,..,p)

Estatística suficiente

- Assim, de (2) e (3) em (1), vem que a verossimilhança pode ser reescrita como:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -0,5 tr \left[\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) (\boldsymbol{x}_i - \overline{\boldsymbol{x}})' \right] + 0,5n \times tr \left[\boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) (\overline{\boldsymbol{x}} - \boldsymbol{\mu})' \right] \right\}$$

- Assim, pelo critério da fatoração temos que $\boldsymbol{T} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i, \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}) (\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \right)$ é uma estatística suficiente para $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ("abuso de notação"), em que $\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i = \frac{1}{n} \left(\sum_{i=1}^{n} X_{i1}, ..., \sum_{i=1}^{n} X_{ip} \right)' = (\overline{X}_1, ..., \overline{X}_p)'$ (note que $\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}, j = 1, 2, .., p$).

Estimadores de máxima verossimilhança

- Por outro lado, temos que a log-verossimilhança pode ser escrita como (definindo $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1}$):

$$
\begin{aligned}
l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= const + \frac{n}{2} \ln |\boldsymbol{V}| - 0,5 \left[\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})' \, \boldsymbol{V} \, (\boldsymbol{x}_i - \boldsymbol{\mu}) \right] \quad (4) \\
&= const + \frac{n}{2} \ln |\boldsymbol{V}| - 0,5 tr \left[\boldsymbol{V} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}) (\boldsymbol{x}_i - \boldsymbol{\mu})' \right] \quad (5)
\end{aligned}
$$

- Derivando (4) com relação à $\boldsymbol{\mu}$ e (5) com relação à $\boldsymbol{V}$, vem que (lembrando que $\frac{\partial}{\partial \boldsymbol{V}} \ln |\boldsymbol{V}| = \boldsymbol{\Sigma}$ e $\frac{\partial}{\partial \boldsymbol{V}} tr(\boldsymbol{V}\boldsymbol{A}) = \boldsymbol{A}'$)

Estimadores de máxima verossimilhança

$$
\begin{aligned}
S(\boldsymbol{\mu}) &= \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\sum_{i=1}^{n}\left[2\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\right] = n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - n\boldsymbol{\Sigma}^{-1}\overline{x} \\
S(\boldsymbol{V}) &= \frac{n}{2}(\boldsymbol{V}^{-1})' - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)' = \frac{n}{2}(\boldsymbol{V}^{-1})' - \frac{1}{2}(n-1)\boldsymbol{S}^2
\end{aligned}
$$

- Assim, igualando cada uma das derivadas acima à $\mathbf{0}_{(p\times 1)}$ e $\mathbf{0}_{(p\times p)}$, respectivamente, e resolvendo o sistema resultante em relação aos parâmetros, temos:

$$
\widehat{\mu} = \overline{\boldsymbol{X}}; \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)'.
$$

Estimadores de máxima verossimilhança

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} X_{i1} \\ \vdots \\ \sum_{i=1}^{n} X_{ip} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)' = \frac{1}{n} \times$$

$$\begin{bmatrix} \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)^2 & \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)(X_{i2}-\overline{X}_2) & \dots & \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)(X_{ip}-\overline{X}_p) \\ \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)(X_{i2}-\overline{X}_2) & \sum_{i=1}^{n}(X_{i2}-\overline{X}_2)^2 & \dots & \sum_{i=1}^{n}(X_{i2}-\overline{X}_2)(X_{ip}-\overline{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)(X_{ip}-\overline{X}_p) & \sum_{i=1}^{n}(X_{i2}-\overline{X}_2)(X_{ip}-\overline{X}_p) & \dots & \sum_{i=1}^{n}(X_{ip}-\overline{X}_p)^2 \end{bmatrix}$$

(Cont.) Distribuiçoes amostrais dos $\widehat{\mu}$ e $\boldsymbol{S}^2$

- Contudo, em geral, utiliza-se $\boldsymbol{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)' = \frac{n}{n-1}\widehat{\boldsymbol{\Sigma}}$ para estimar $\boldsymbol{\Sigma}$ (por ser não viciado), enquanto que $\mathcal{E}(\widehat{\boldsymbol{\Sigma}}) = \frac{(n-1)}{n}\boldsymbol{\Sigma}$. Pode-se provar ainda que $\overline{\boldsymbol{X}} \perp \boldsymbol{S}^2$ (Teorema de Basu).
- Temos que $\widehat{\boldsymbol{\mu}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ (exercício).
  - Assim, $\widehat{\mu}_j \sim N(\mu_j, \sigma_j^2/n), j = 1, 2..., p$, em que $\widehat{\mu}_j = \overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$.
- Temos ainda que $(n-1)\boldsymbol{S}^2 \sim W_p(n-1, \boldsymbol{\Sigma})$ (link com demonstrações e referências a respeito).
  - Assim, $\frac{(n-1)S_j^2}{\sigma_j^2} \sim \chi^2_{(n-1)}, j = 1, 2, ..., p$, em que $S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \widehat{\mu}_j)^2$, é o j-ésimo elemento da diagonal principal de $\boldsymbol{S}^2$.

(Cont.) Distribuiçoes amostrais dos $\widehat{\mu}$ e $\boldsymbol{S}^2$

- Densidade da distribuição de Wishart ($\boldsymbol{\Sigma} \sim W_p(n, \boldsymbol{V})$), (pesquisar sobre ela):

$$f(\boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{\frac{n-p-1}{2}} e^{-\frac{\operatorname{tr}(\mathbf{V}^{-1}\boldsymbol{\Sigma})}{2}}}{2^{\frac{np}{2}} |\mathbf{V}|^{\frac{n}{2}} \Gamma_p(\frac{n}{2})}$$

  em que $\Gamma_p(k) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(k + (1-p)/2)$ é a função gama p-variada.

- Seja $\boldsymbol{X}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{V})$ e defina $\boldsymbol{\Sigma} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'$, então $\boldsymbol{\Sigma} \sim W_p(n, \boldsymbol{V})$ (obs: esta matriz $\boldsymbol{V}$ não é a mesma definida anteriormente).

Inferência para cada componente: intervalos de confiança

- $IC(\theta, \gamma)$ representa um intervalo de confiança com coeficiente de confiança γ para o parâmetro θ.
- (médias) $IC(\mu_j, \gamma) = \left[\overline{X}_j - t_{\frac{1+\gamma}{2}} \sqrt{\frac{S_j^2}{n}}, \overline{X}_j + t_{\frac{1+\gamma}{2}} \sqrt{\frac{S_j^2}{n}}\right]$, em que $P(T \leq t_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2}$, $T \sim t_{(n-1)}$.
- (variâncias) $IC(\sigma_j^2, \gamma) = \left[\frac{(n-1)S_j^2}{q_{\frac{1+\gamma}{2}}}, \frac{(n-1)S_j^2}{q_{\frac{1-\gamma}{2}}}\right]$, em que $P(Q \leq q_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2}$, e $P(Q \leq q_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2}$ $Q \sim \chi^2_{(n-1)}$.

Inferência para cada componente: intervalos de confiança

- (correlações, $i \neq j$) $IC(\rho_{ij}, \gamma) = [\tanh(z_1'); \tanh(z_2')]$, em que

$$z_1' = z' - \frac{z_{\frac{1+\gamma}{2}}}{\sqrt{n-3}} \text{ e } z_2' = z' + \frac{z_{\frac{1+\gamma}{2}}}{\sqrt{n-3}},$$

$z' = 0,5 \ln\left(\frac{1+\widehat{\rho}_{ij}}{1-\widehat{\rho}_{ij}}\right)$,

$$\widehat{\rho}_{ij} = \frac{\frac{1}{n}\sum_{k=1}^{n} X_{ki}X_{kj} - \overline{X}_i\overline{X}_j}{\sqrt{S_i^2 S_j^2}}$$

é a correlação amostral de Pearson e
$P(Z \leq z_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2}, Z \sim N(0,1)$.

Testes de hipótese (TH) para cada componente de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$

- Para as componente μ_j e σ_j^2, $j = 1, 2, ..., p$ temos diversos resultados aqui apresentados:
  - Aulas teóricas (ME414): TH1, TH2 e TH3.
  - Aulas práticas (ME414): TH1, TH2.
- Para testar $H_0 : \rho_{ij} = 0$ vs $H_1 : \rho_{ij} \neq 0$, podemos usar a estatística do teste $Z_t = \frac{\widehat{\rho}_{ij}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}_{ij}^2}}$, rejeitando H_0 se $|z_t| > z_c$ ou se $p - valor = P(Z > z_t|H_0) > \alpha$, em que z_t é o valor calculado da estatística do teste, α é o nível de significância, $Z \sim N(0,1)$ e $P(Z < z_c|H_0) = \frac{1-\alpha}{2}$.

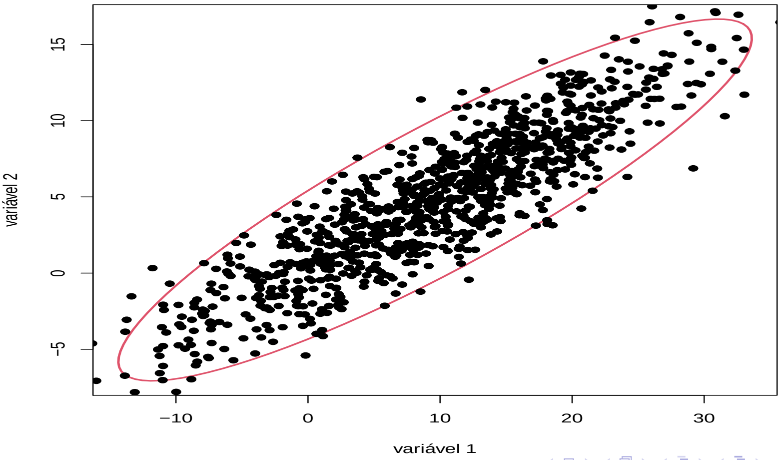
Inferência para combinações lineares do vetor de médias

- Defina $\theta = \boldsymbol{R}_{(1\times p)}\boldsymbol{\mu}$ em que $\boldsymbol{R}$ é um vetor não aleatório.
- Temos que $\widehat{\theta} = \boldsymbol{R}\widehat{\boldsymbol{\mu}} \sim N(\boldsymbol{R}\boldsymbol{\mu}, \boldsymbol{R}\Sigma\boldsymbol{R}'/n)$ (exercício).
- Assim, analogamente ao caso anterior, temos que $IC(\theta,\gamma) = \left[\boldsymbol{R}\overline{\boldsymbol{X}} - t_{\frac{1+\gamma}{2}}\sqrt{\frac{\boldsymbol{R}\boldsymbol{S}^2\boldsymbol{R}'}{n}}, \boldsymbol{R}\overline{\boldsymbol{X}} + t_{\frac{1+\gamma}{2}}\sqrt{\frac{\boldsymbol{R}\boldsymbol{S}^2\boldsymbol{R}'}{n}}\right]$, em que $P(T \leq t_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2}$, $T \sim t_{(n-1)}$.

Região de confiança (RC) para $\boldsymbol{\mu}$

- Pode-se provar que :
  - $(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_{(p)}$ e $Q = n(\overline{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\left(\overline{\boldsymbol{X}}-\boldsymbol{\mu}\right) \sim \chi^2_{(p)}$ (prova:fgm).
  - $T = n(\overline{\boldsymbol{X}}-\boldsymbol{\mu})'\left(\boldsymbol{S}^2\right)^{-1}\left(\overline{\boldsymbol{X}}-\boldsymbol{\mu}\right) \sim T^2_{(p,n-1)}$ de Hotelling e $F = \frac{n-p}{(n-1)p}T \sim F_{(p,n-p)}$.
  - Definição $\boldsymbol{X} \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ e $\boldsymbol{W} \sim W_p(k,\boldsymbol{\Sigma})$, $\boldsymbol{X} \perp \boldsymbol{W}$. Assim $Y = (\boldsymbol{X}-\boldsymbol{\mu})'(\boldsymbol{W}/k)^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim T^2_{(p,k)}$ de Hotelling e $F = \frac{k-p+1}{pk}Y \sim F_{(p,k-p+1)}$ (veja site do curso e referências).
- Portanto, uma região de confiança é um conjunto, digamos $R(\boldsymbol{\mu},\gamma)$, contido em $\mathcal{R}^p$, tal que $P(R(\boldsymbol{\mu},\gamma) \leq q_\gamma) = \gamma$, em que q_γ é um quantil apropriado da distribuição associada à $R(\boldsymbol{\mu},\boldsymbol{\gamma})$.

Exemplo de RC para μ com p = 2



Região de confiança para $\boldsymbol{\mu}$

- Então, se f_γ for um quantil da distribuição $F_{(p,n-p)}$, tal que $P(F \leq f_\gamma) = \gamma$, uma região de confiança γ para $\boldsymbol{\mu}$ é o conjunto de todas as $p-$uplas, digamos $\boldsymbol{\mu}^*$, tais que

$$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}^*)' \left(\boldsymbol{s}^2\right)^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}^*) \leq \frac{p(n-1)}{(n-p)} f_\gamma$$

- Para uma dada amostra (matriz de dados observada) basta calcular a forma quadrática acima (para todas as $p-$uplas) e ver quais satisfazem à inequação.
- Mais detalhes (incluindo $p \geq 3$): Johson and Wichern (2018). No R: função "ellipse" do pacote "ellipse": página, manual.

Simulação de um vetor aleatório $N_p(\boldsymbol{\mu}, \Sigma)$

- Seja $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\mu} \in \mathcal{R}^p, \boldsymbol{\Sigma}$ positiva definida.
- Sabemos que $\boldsymbol{X} = \boldsymbol{\Psi}\boldsymbol{Z} + \boldsymbol{\mu}$, $\boldsymbol{\Psi}$ = Cholesky$(\boldsymbol{\Sigma})$,
  $\boldsymbol{Z} = (Z_1, ..., Z_p), Z_i \overset{i.i.d.}{\sim} N(0,1), i = 1, ..., p$ (exercício).
- Ou seja, $\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{F}_Z^{-1}(\boldsymbol{u}) + \boldsymbol{\mu}, \boldsymbol{u} = (u_1, ..., u_p)'$,
  $u_i \overset{iid}{\sim} U(0,1), i = 1, 2, ..., p$, em que
  $\boldsymbol{F}^{-1}(\boldsymbol{u}) = (F_Z^{-1}(u_1), ..., F_Z^{-1}(u_p))'$ (vetor coluna), $Z \sim N(0,1)$.

No pacote R

- Código genérico:

  ```
  m.u = cbind(runif(n)); m.z = qnorm(m.u)
  for (i in 2:p){m.u = cbind(runif(n))
  m.z = cbind(m.z,qnorm(m.u))};
  m.x = t(chol(m.sigma)%*%t(m.z) + matrix(v.mu,p,n))
  ```

- Também no R, de forma mais simples: função "mvrnorm " do pacote "MASS":

  mvrnorm(n="tamanho da amostra", mu="vetor de médias", Sigma="matriz de covariâncias")

Verificação da normalidade multivariada de uma matriz de dados

- $X_{ij} \sim N(\mu_j, \sigma_j^2), j = 1, 2, ..., p, i = 1, ..., n$, ou seja: $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- $n\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)' \left(\boldsymbol{S}^2\right)^{-1} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right) \approx \chi_p^2$, para n suficientemente grande.
- Construir histogramas e/ou gráficos de quantis-quantis com envelope para cada variável e para a forma quadrática, bem como a utilização de testes não paramétricos (p.e., Kolmogorov-Smirnov).

Procedimento para se gerar o gráfico de envelopes (quantil-quantil)

1) Simule n variáveis aleatórias independentes de interesse ($N(0,1)$ ou $\chi^2_{(p)}$, por exemplo). Repita este processo m vezes.
2) Ao final teremos uma matriz com valores simulados dessas variáveis aleatórias, digamos V_{ij}, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{v2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \dots & v_{nm} \end{bmatrix}$$

Cont.

3) Dentro de cada amostra (coluna), ordena-se, de modo crescente, os valores simulados, obtendo-se $v^*_{(i)j}$ (i-ésima estatística de ordem da j-ésima amostra):

$$\boldsymbol{V}^* = \begin{bmatrix} v_{(1)1} & v_{(1)2} & \cdots & v_{(1)m} \\ v_{(2)1} & v_{(2)2} & \cdots & v_{(2)m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{(n)1} & v_{(n)2} & \cdots & v_{(n)m} \end{bmatrix}$$

4) Pode-se obter os limites $v_{(i)I} = \min_{1 \leq j \leq m} v_{(i)j}$ e $v_{(i)S} = \max_{1 \leq j \leq m} v_{(i)j}$.

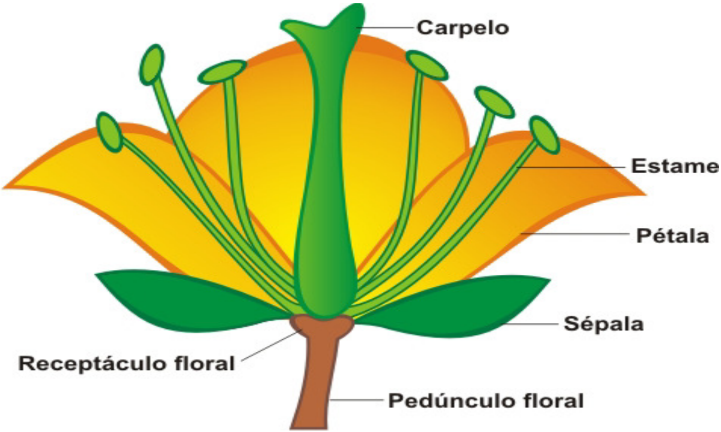
Cont.

5) Porém, na prática considera-se $v_{(i)I} = \frac{v_{(i)(2)} + v_{(i)(3)}}{2}$ e $v_{(i)S} = \frac{v_{(i)(m-2)} + v_{i(m-1)}}{2}$ (para se gerar limites de confiança), em que $v_{(i)(r)}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ..., n$.

- Além disso, consideramos como a linha de referência $v_{(i)} = \frac{1}{m}\sum_{j=1}^{m} v_{(i)j}, i = 1, 2, ..., n$.

Dados da íris “de Fisher”

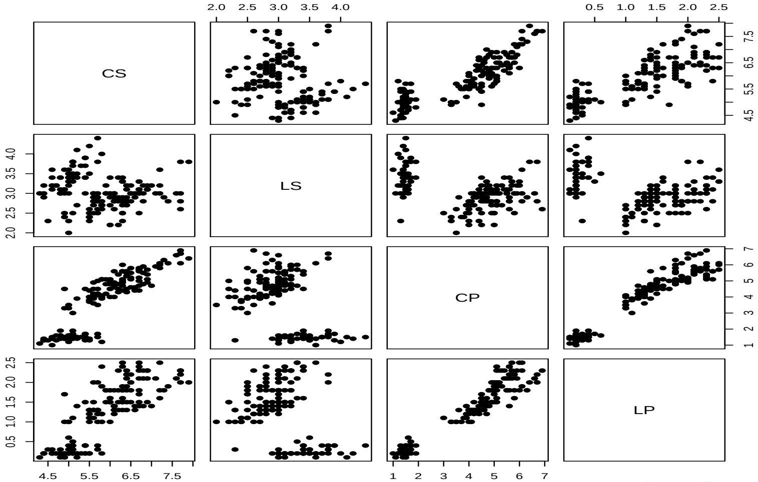
- Os dados consistem de 50 unidades amostrais de três espécies (setosa, virginica, versicolor) de íris (uma espécie de planta), ou seja, temos um total de 150 unidades amostrais (Fisher (1936) e Anderson (1935)).
- De cada uma delas mediu-se quatro variáveis: comprimento e largura da sépala (CS, LS) e comprimento e largura da pétala (CP,LP).
- Objetivo original: quantificar a variação morfológica em relação à essas espécies com bases nas quatro variáveis de interesse.
- Banco de dados default no R, sob o nome de “iris”.


Carpelo
Estame
Pétala
Sépala
Receptáculo floral
Pedúnculo floral

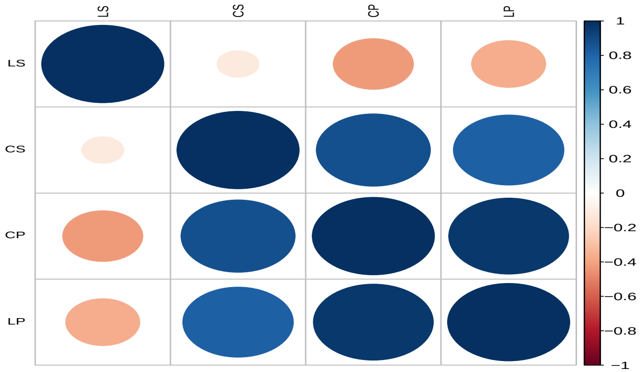
Cont.

- Seja Y_{ijk} : o valor da k-ésima variável (k=1,2,3,4), para a j-ésima flor (j=1,...,50) do i-ésimo grupo ($i = 1, 2, 3$).
- Por enquanto, vamos desconsiderar as espécies de Íris, ou seja, vamos considerar que $\boldsymbol{Y}_j = (Y_{j1}, Y_{j2}, Y_{i3}, Y_{j4})' \stackrel{ind.}{\sim} N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, na seguinte ordem: comprimento e largura da sépala (CS, LS) e comprimento e largura da pétala (CP,LP).

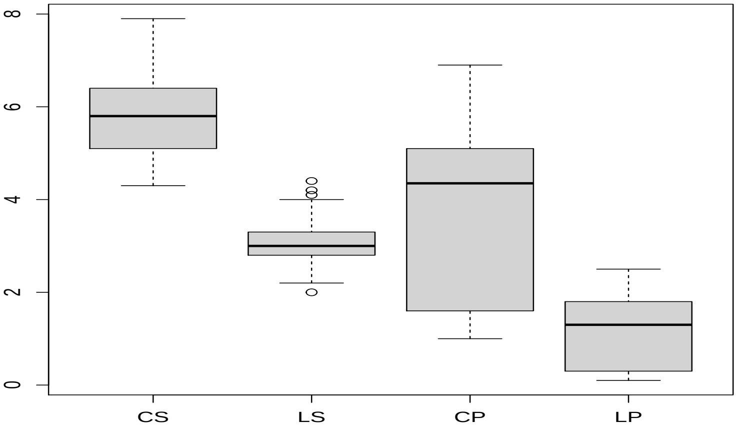
Gráficos de dispersão múltipla



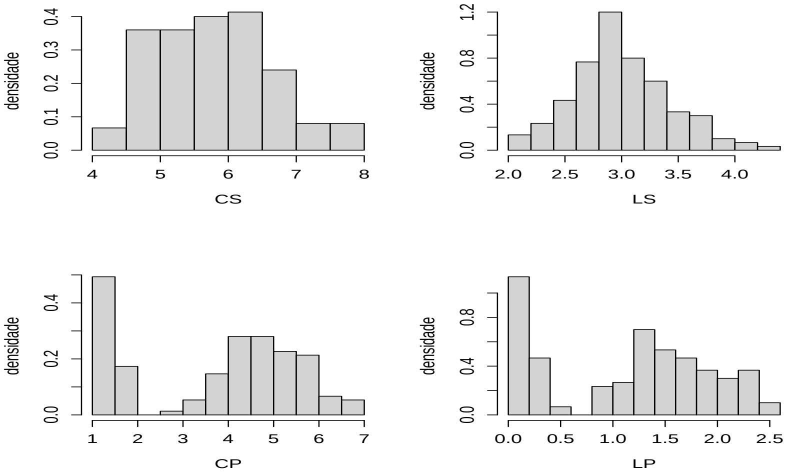
Matriz de correlações



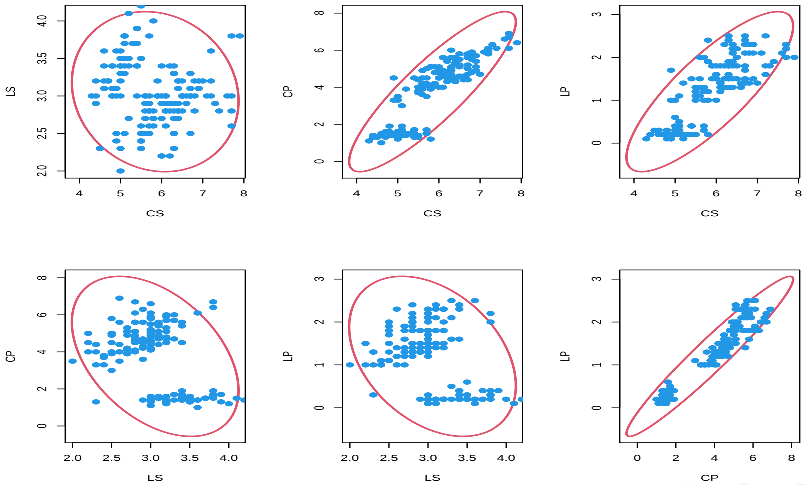
Box-plot das variáveis



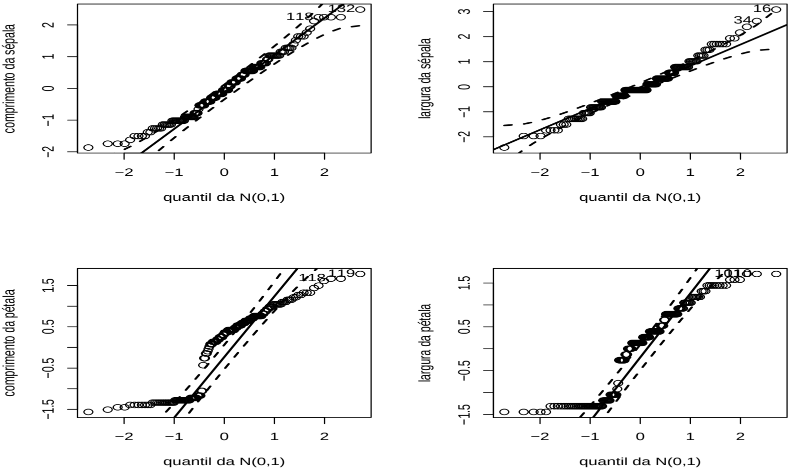
Histogramas das variáveis



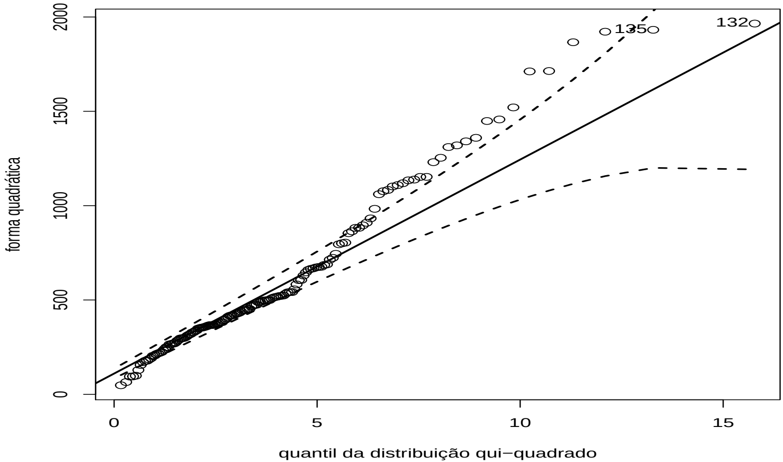
Elipsóides de 95% confiança para cada par de variáveis



Gráficos QQ plot para cada variável



Gráficos QQ plot para a forma quadrática



- Testes de Kolmogorov-Smirnov (p-valor entre parênteses):
  - CS: D= 0,089 ($p = 0{,}1891$).
  - LS: D= 0,106 ($p = 0{,}0723$).
  - CP: D= 0,198 ($p < 0{,}0001$).
  - LP: D= 0,173 ($p < 0{,}0003$).
- Teste de Kolmogorov-Smirnov (p-valor entre parênteses) para a forma quadrática:
  - D= 1 ($p < 0{,}0001$).