Análise de dados sob normalidade

Prof. Caio Azevedo

Resultados probabilísticos importantes

- Se $Y|x \sim N(x, \sigma^2)$ e $X \sim N(a, b)$, então $Y \sim N(a, \sigma^2 + b)$.
- Se $X|y \sim N(\mu, y/\nu)$ e $y \sim IG(a, b)$, então

$$(X, Y) \sim NIG(\mu, \nu, a, b)$$

e

$$X \sim t_{(2a)}\left(\mu, \sqrt{\frac{b}{\nu a}}\right).$$

Variância conhecida

- Seja $X_1|\boldsymbol{\theta}, ..., X_n|\boldsymbol{\theta}, \boldsymbol{\theta} = (\mu, \sigma^2)$ uma amostra aleatória de $X|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$.
- Se σ^2 conhecido, e $\mu \sim N(a, b)$, (família conjugada) então $\mu|\mathbf{x} \sim N(\lambda, \psi)$, em que

$$\psi = \frac{\sigma^2 b}{nb + \sigma^2}; \lambda = \psi\left(\frac{a}{b} + \frac{n\overline{x}}{\sigma^2}\right)$$

- Distribuição preditiva à posteriori de uma única observação $X_{n+1}|\mathbf{x} \sim N(\lambda, \psi + \sigma^2)$.

Média conhecida

- Seja $X_1|\boldsymbol{\theta}, ..., X_n|\boldsymbol{\theta}, \boldsymbol{\theta} = (\mu, \sigma^2)$ uma amostra aleatória de $X|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$.
- Se μ conhecido, e $\sigma^2 \sim IG(a, b)$, (família conjugada) então $\sigma^2|\mathbf{x} \sim IG(a^*, b^*)$, em que

$$a^* = \frac{n}{2} + a; b^* = \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + b$$

- Distribuição preditiva à posteriori para uma única observação $X_{n+1}|\mathbf{x} \sim t_{(2a^*)}\left(\mu, \sqrt{\frac{b^*}{a^*}}\right)$

Ambos os parâmetros desconhecidos

- Família conjugada (normal inversa gama)

$$\begin{aligned} \mu|\sigma^2 &\sim N(\lambda, \sigma^2/\nu) \\ \sigma^2 &\sim \text{IG}(a, b) \end{aligned}$$

- Posteriori conjunta

$$\begin{aligned} \mu|\mathbf{x}, \sigma^2 &\sim N(c, \sigma^2/\nu^*) \\ \sigma^2|\mathbf{x} &\sim \text{IG}(a^*, b^*) \end{aligned}$$

em que $c = \frac{n\bar{x}+\nu\lambda}{n+\nu}$, $\nu^* = \nu + n$,
$b^* = \frac{1}{2}\left[\frac{n\nu}{n+\nu}(\bar{x}-\lambda)^2 + (n-1)s^2\right] + b$, $a^* = \frac{n}{2} + a$ e
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Cont.

- Além disso, $\mu|\mathbf{x} \sim t_{(2a^*)}\left(c, \sqrt{\frac{b^*}{\nu^* a^*}}\right)$.
- Distribuição preditiva à posteriori para uma única observação $X_{n+1}|\mathbf{x} \sim t_{(2a^*)}\left(c, \sqrt{\frac{b^*}{a^* \nu^{**}}}\right)$, em que $\nu^{**} = \frac{\nu^*}{1+\nu^*}$.

Exemplo 11: Distribuições normais com variâncias desconhecidas porém iguais

- $X_i|\boldsymbol{\theta} \sim N(\mu_1, \sigma^2), i = 1, 2, ..., n.$
- $Y_j|\boldsymbol{\theta} \sim N(\mu_2, \sigma^2), i = 1, 2, ..., m.$
- $X_i|\boldsymbol{\theta} \perp Y_j|\boldsymbol{\theta}, \forall i, j$ e $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$.
- (Exercício) Priori de Jeffreys sob independência:

  $p(\theta) = (\sigma^2)^{-1} \mathbb{1}_{\Theta}(\boldsymbol{\theta})$, em que

  $\mathbb{1}_{\Theta}(\boldsymbol{\theta}) = \mathbb{1}_{\mathcal{R}}(\mu_1)\mathbb{1}_{\mathcal{R}}(\mu_2)\mathbb{1}_{\mathcal{R}^+}(\sigma^2)$

Exemplo 11 (cont.)

- Pode-se provar que

$$\begin{aligned} p(\theta|\mathbf{x},\mathbf{y}) &\propto e^{\left\{-\frac{n}{\sigma^2}(\mu_1-\overline{x})^2\right\}}(\sigma^2)^{-1/2} e^{\left\{-\frac{m}{\sigma^2}(\mu_2-\overline{y})^2\right\}}(\sigma^2)^{-1/2} \\ &\times (\sigma^2)^{-(k/2+1)} e^{-\frac{ks^2}{2\sigma^2}} \mathbb{1}_{\Theta}(\boldsymbol{\theta}) \end{aligned}$$

em que $k = n + m - 2$, $s^2 = \frac{1}{k}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{j=1}^{m}(y_j - \overline{y})^2\right]$

- Ou seja,

$\mu_1|(\sigma^2, \mathbf{x}, \mathbf{y}) \sim N(\overline{x}, \sigma^2/n)$, $\mu_2|(\sigma^2, \mathbf{x}, \mathbf{y}) \sim N(\overline{y}, \sigma^2/m)$ e

$\sigma^2|(\mathbf{x}, \mathbf{y}) \sim IG(k/2, ks^2/2)$

Exemplo 11 (cont.)

- Assim, se $\lambda = \mu_1 - \mu_2$, então $\lambda|(\sigma^2, \mathbf{x}, \mathbf{y}) \sim N\left(\overline{x} - \overline{y}, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$.
- Logo, $\lambda|(\mathbf{x}, \mathbf{y}) \sim t_{(k)}\left(\overline{x} - \overline{y}, s\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}\right)$
- Logo, podemos utilizar a distribuição a posteriori acima para verificar se $\mu_1 = \mu_2$.
- Se $X \sim t_\nu(0, 1)$ então $Y = \delta X + \mu \sim t_\nu(\mu, \delta)$, logo $p_Y(y|\mu, \delta, \nu) = \frac{1}{\delta} p_X((x - \mu)/\delta|\nu)$
- Além disso, $\frac{\lambda - (\overline{x} - \overline{y})}{s\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}}|(\mathbf{x}, \mathbf{y}) \sim t_{(k)}(0, 1)$

Exemplo 12: Distribuições normais com variâncias desconhecidas e diferentes

- $X_i|\boldsymbol{\theta} \sim N(\mu_1, \sigma_1^2), i = 1, 2, ..., n.$
- $Y_j|\boldsymbol{\theta} \sim N(\mu_2, \sigma_2^2), i = 1, 2, ..., m.$
- $X_i|\boldsymbol{\theta} \perp Y_j|\boldsymbol{\theta}, \forall i, j$ e $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2).$
- (Exercício) Priori de Jeffreys sob independência:

  $p(\theta) = (\sigma_1^2)^{-1}(\sigma_2^2)^{-1}\mathbb{1}_\Theta(\boldsymbol{\theta})$, em que

  $\mathbb{1}_\Theta(\boldsymbol{\theta}) = \mathbb{1}_\mathcal{R}(\mu_1)\mathbb{1}_\mathcal{R}(\mu_2)\mathbb{1}_{\mathcal{R}^+}(\sigma_1^2)\mathbb{1}_{\mathcal{R}^+}(\sigma_2^2)$

Exemplo 12 (cont.)

- Pode-se provar que:

$$\mu_1|(\mathbf{x},\mathbf{y}) \sim t_{k_1}(\overline{x}, s_1/\sqrt{n})$$

$$\mu_2|(\mathbf{x},\mathbf{y}) \sim t_{k_2}(\overline{y}, s_2/\sqrt{m})$$

$$\sigma_1^2|(\mathbf{x},\mathbf{y}) \sim IG(k_1/2, k_1 s_1^2/2)$$

$$\sigma_2^2|(\mathbf{x},\mathbf{y}) \sim IG(k_2/2, k_2 s_2^2/2)$$

em que $k_1 = n-1$, $k_2 = m-1$, $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, $s_2^2 = \frac{1}{m-1}\sum_{j=1}^{n}(y_j - \overline{y})^2$, $s_i = \sqrt{s_i^2}, i = 1, 2.$

Exemplo 12 (cont.)

- Definindo-se $\lambda = \mu_1 - \mu_2$ e $\tau = \frac{\lambda - (\overline{x} - \overline{y})}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$, podemos provar que

$$\tau | (\mathbf{x}, \mathbf{y}) \approx t_{(b)}(0, a)$$

em que $a = \sqrt{(b-2)c_1/b}$, $b = 4 + c_1^2/c_2$,

$c_1 = \frac{k_1}{k_1 - 2} sin^2 u + \frac{k_2}{k_2 - 2} cos^2 u$,

$c_2 = \frac{k_1^2}{(k_1 - 2)^2 (k_1 - 4)} sin^4 u + \frac{k_2^2}{(k_2 - 2)^2 (k_2 - 4)} cos^4 u$

- Por outro lado podemos, simplesmente, obter uma aproximação numérica para a distribuição de $\lambda | (\mathbf{x}, \mathbf{y})$.

Exemplo 12 (cont.)

- Uma vez que $\mu_1|(\mathbf{x},\mathbf{y}) \sim t_{k_1}(\overline{x}, s_1/\sqrt{n}) \perp \mu_2|(\mathbf{x},\mathbf{y}) \sim t_{k_2}(\overline{y}, s_2/\sqrt{m})$, podemos simular R variáveis aleatórias, mutuamente independentes, com distribuições t específicas, e obter λ para cada par, ou seja
- Simular $(\mu_1^{(r)}, \mu_2^{(r)}), r = 1, ..., R$ (das respectivas distribuições) e calcular $\lambda^{(r)} = \mu_1^{(r)} - \mu_2^{(r)}$.

Exemplo 12 (cont.)

- Para comparar as variâncias, basta notar que $\frac{k_1 s_1^2}{\sigma_1^2}|(\mathbf{x},\mathbf{y}) \sim \chi^2_{(k_1)} \perp \frac{k_2 s_2^2}{\sigma_2^2}|(\mathbf{x},\mathbf{y}) \sim \chi^2_{(k_2)}$ e, assim

$$\frac{s_2^2}{s_1^2}\psi|(\mathbf{x},\mathbf{y}) \sim F_{(k_2,k_1)}$$

em que $\psi = \frac{\sigma_1^2}{\sigma_2^2}$.

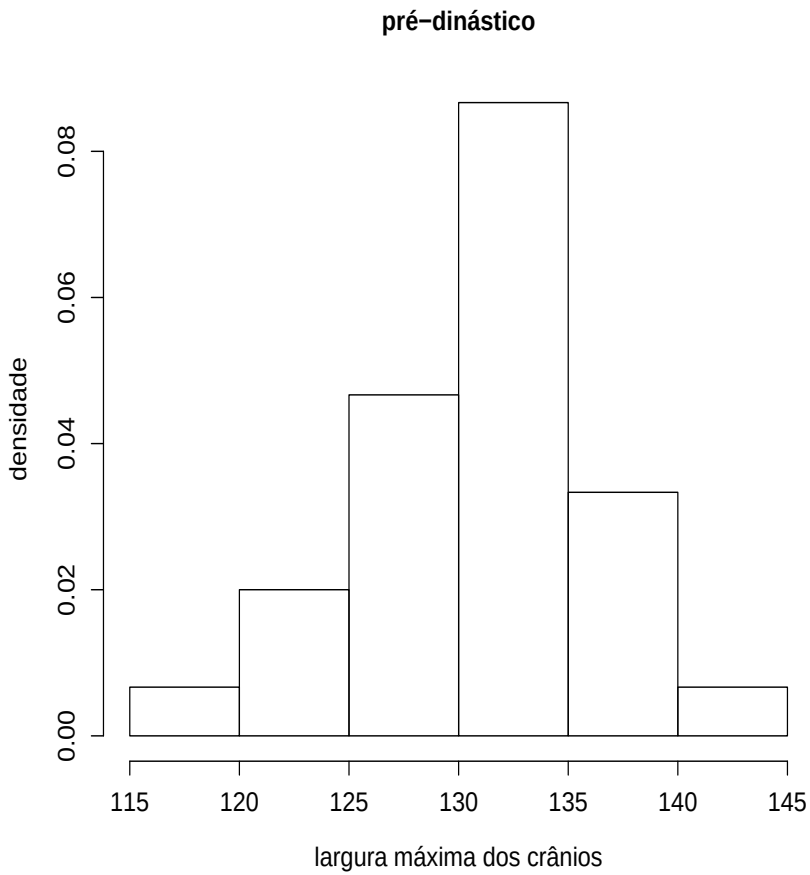
Dados

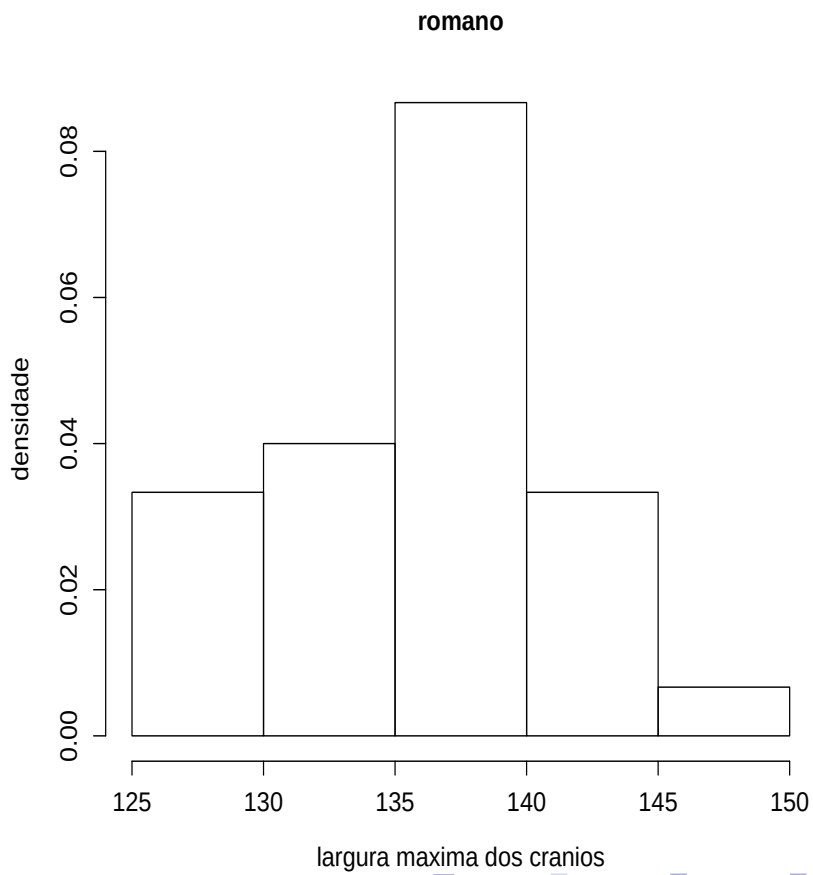
- O conjunto de dados se refere à n = m = 30 observações correspondentes às larguras máximas de crânios humanos, datadas do período pré-dinástico (grupo 1) e romano (grupo 2), respectivamente.
- Objetivo principal: comparar as médias populacionais das larguras máximas entre os tipos de crânios (origem).

Medidas resumo

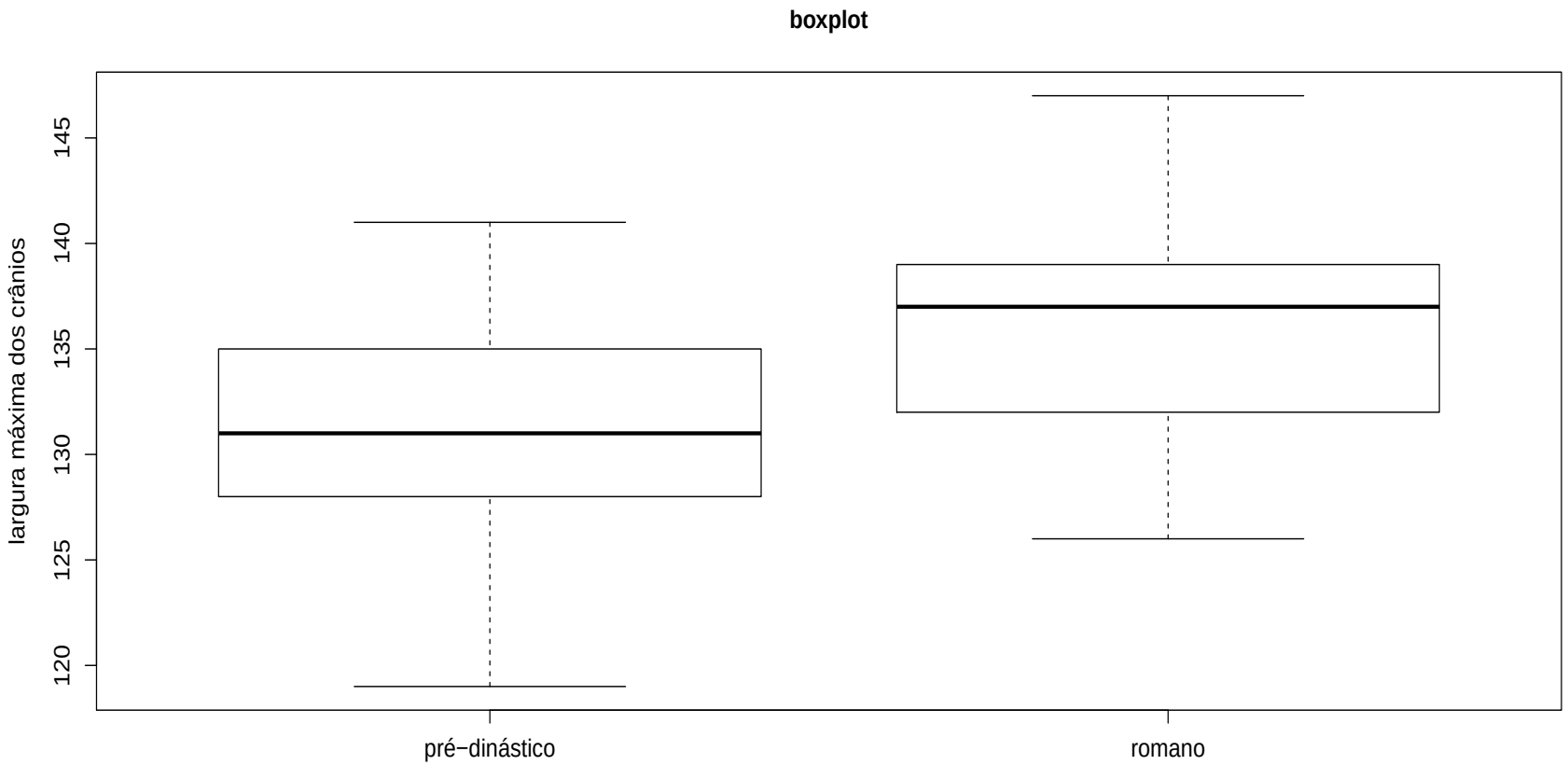
Origem	média	var.	dp	cv(%)	min.	med.	máx.
pré-dinástico	131,37	26,31	5,13	3,90	119,00	131,00	141,00
romano	136,17	28,63	5,35	3,93	126,00	137,00	147,00

Histogramas

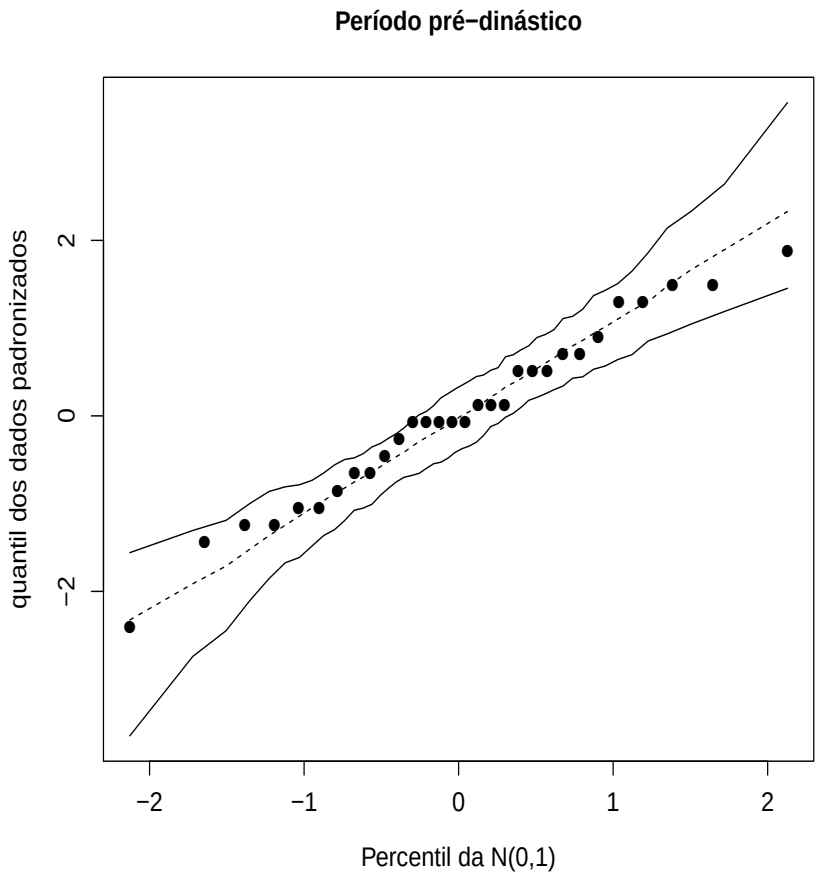


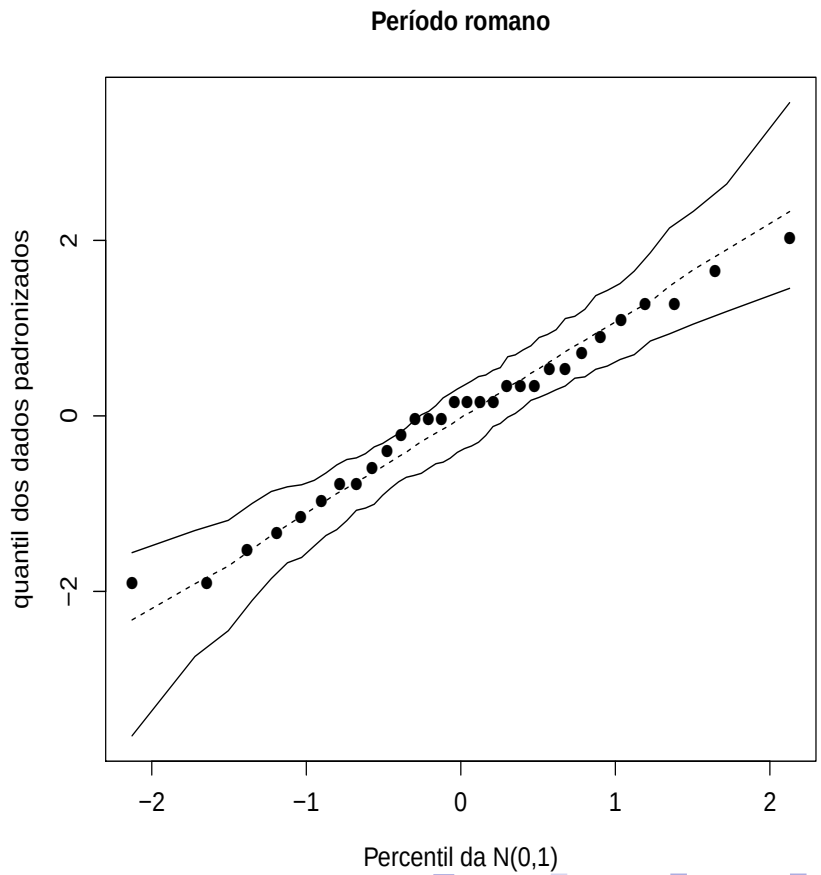


Boxplots

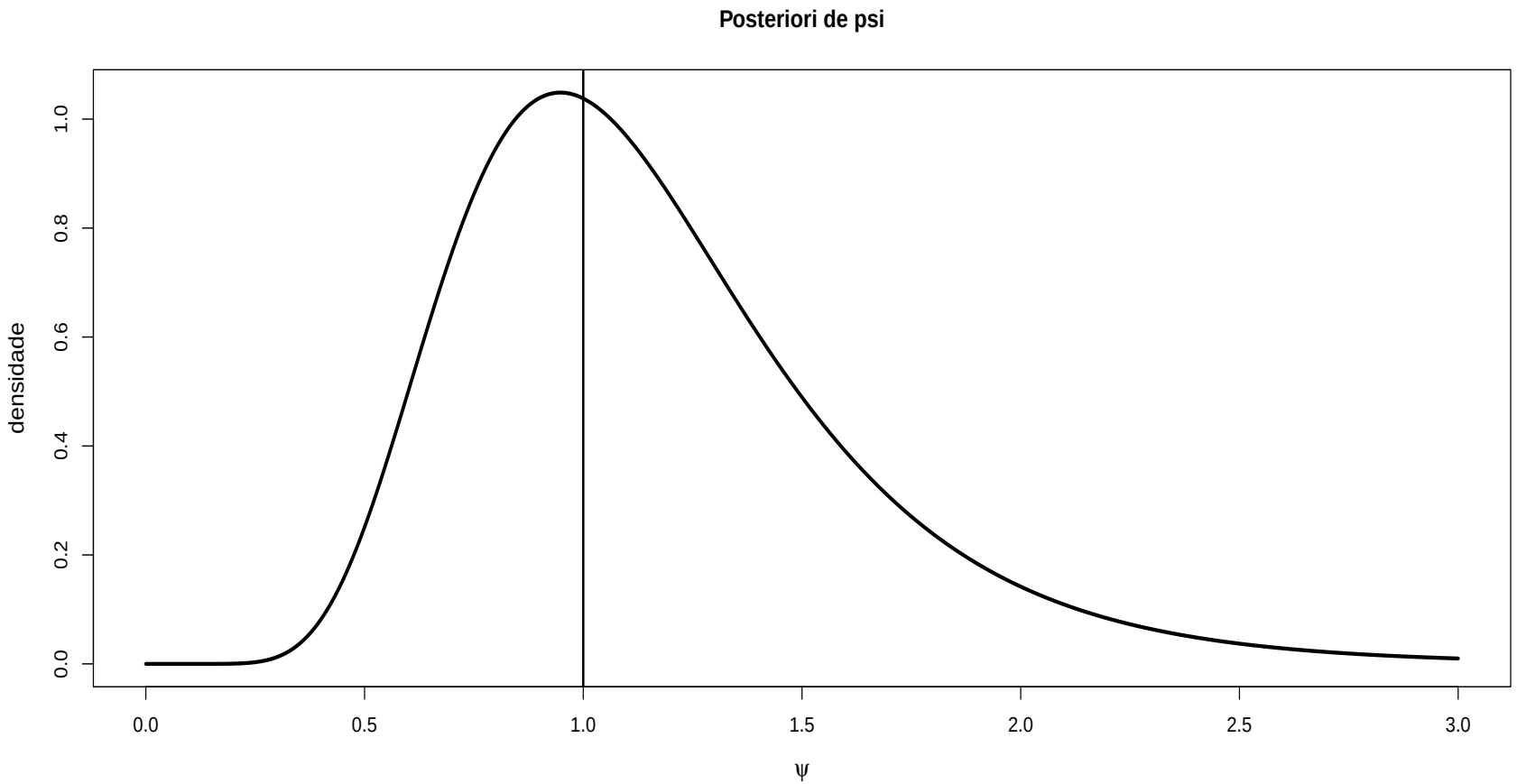

boxplot
largura máxima dos crânios
120
125
130
135
140
145
pré−dinástico
romano

Gráficos de Quantis-quantis N(0,1)





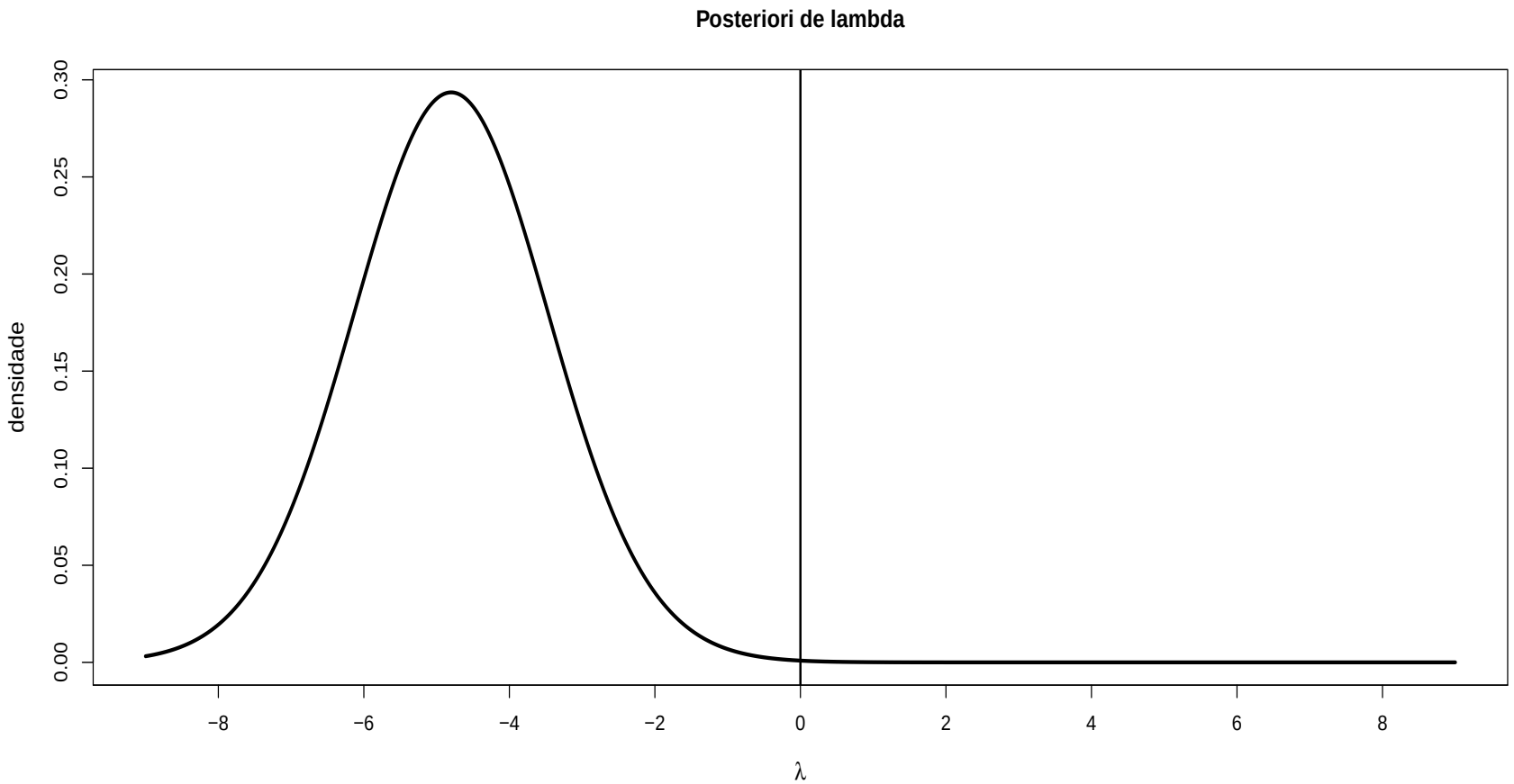
Posteriori de ψ



Comparação das variâncias

- $IC_B(\psi; 0,95) = [0,437; 1,931]$.
- $HPD(\psi; 0,95) = [0,360; 1,757]$.
- Neste caso, como a transformação que associa ψ à distribuição F é linear, podemos obter o intervalo HPD para transformação e depois para ψ, através da transformação inversa.
- Os resultados acima nos levam à concluir que $\sigma_1^2 = \sigma_2^2$ com uma credibilidade de 95%.

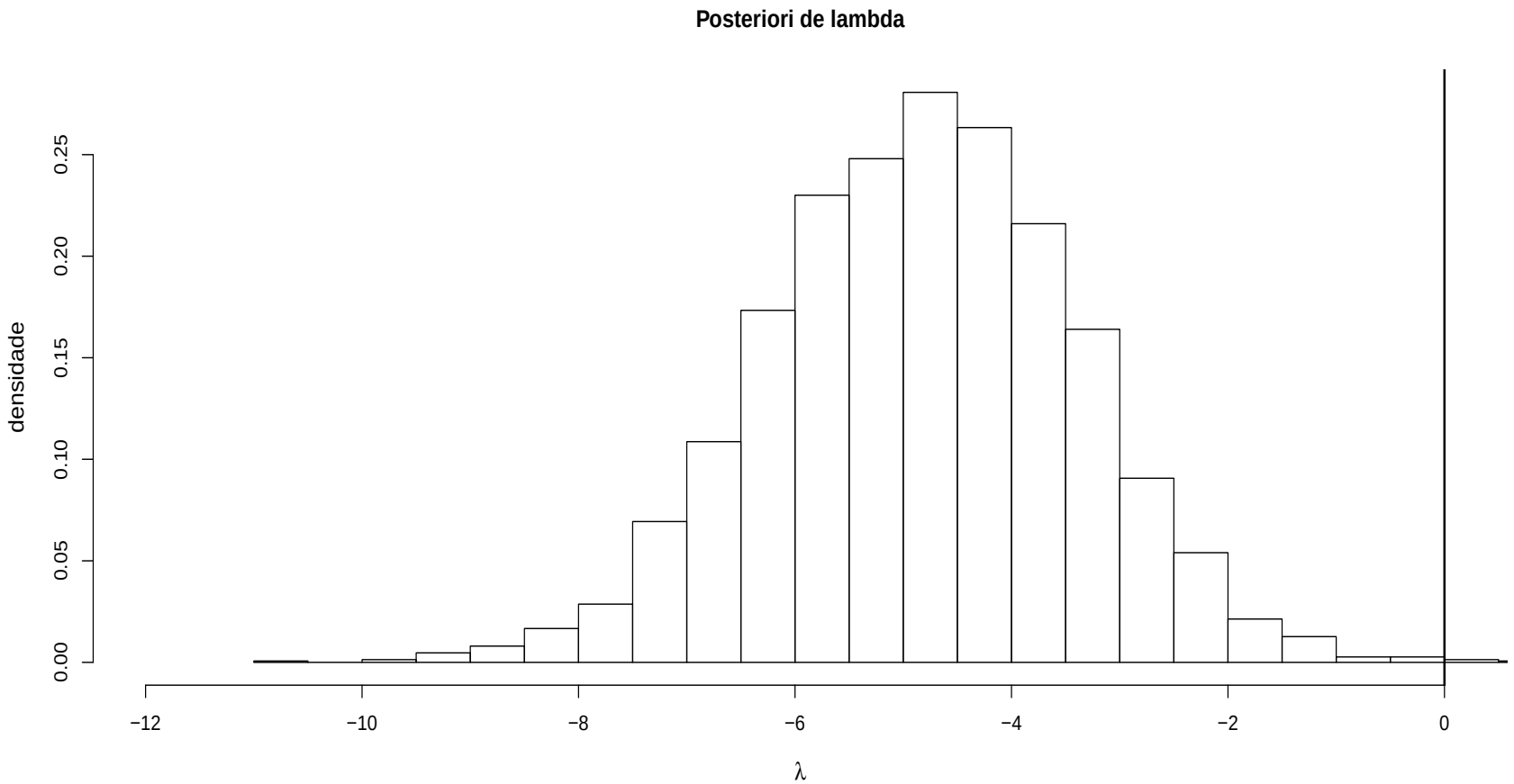
Posteriori de λ, $\sigma_1^2 = \sigma_2^2$



Comparação das médias, supondo iguais as variâncias

- $IC_B(\psi; 0,95) = HPD(\psi; 0,95) = [-7,509; -2,091]$ (pois a posteriori é simétrica e unimodal).
- $P(\lambda < 0|\mathbf{x}) = 0,9996$.
- Os resultados acima nos levam à concluir que $\mu_1 < \mu_2$ com uma credibilidade de 95%.

Posteriori de λ, $\sigma_1^2 \neq \sigma_2^2$, $R = 3000$



Comparação das médias, supondo as variâncias diferentes

- $IC_B(\psi; 0,95) = HPD(\psi; 0,95) = [7,592; -2.106]$ (pois a posteriori é simétrica e unimodal, neste caso, os resultados são aproximados).
- $P(\lambda < 0|\mathbf{x}) = 0,9987$.
- Os resultados acima nos levam à concluir que $\mu_1 < \mu_2$ com uma credibilidade de 95%.