Análise de dados sob normalidade

Prof. Caio Azevedo

Introdução

- Desenvolveremos alguns resultados importantes para análise de dados, sob normalidade.
- Assim como na inferência frequentista, a distribuição normal desempenha um papel bastante relevante, embora a suposição de normalidade seja raramente válida.
- A importância acima descrita se deve tanto ao fato da normal apresentar propriedades bastante importantes/úteis, bem como ser base para o desenvolvimento de modelos mais flexíveis (normal assimétrica, t de Student assimétrica, escala normal, escala normal assimétrica).

Resultados probabilísticos importantes

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

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

e

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

- Normal inversa gama (NIG): aqui

Distribuição t de Student

- Se $X|(\mu, \sigma^2, \nu) \sim t_\nu(\mu, \sigma)$, $\boldsymbol{\theta} = (\mu, \sigma^2, \nu)'$, $\sigma = \sqrt{\sigma^2}$, então:

$$p(x \mid \nu, \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} \mathbb{1}_{(-\infty,\infty)}(x)$$

- Neste caso, $\mathcal{E}(X|\boldsymbol{\theta}) = \mu$, $\mathcal{V}(X|\theta) = \sigma^2 \frac{\nu}{\nu-2}$, se $\nu > 2$, $\mu \in \mathcal{R}$, $\sigma^2, \nu \in \mathcal{R}^+$ e $\sigma = \sqrt{\sigma^2}$.

Variância conhecida

- Seja $X_1|\boldsymbol{\theta}, ..., X_n|\boldsymbol{\theta}, \boldsymbol{\theta} = (\mu, \sigma^2)'$ uma amostra aleatória de $X|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$, então (exercício):
  - Se σ^2 é conhecido, e $\mu \sim N(a, b)$, (família conjugada) então $\mu|\boldsymbol{x} \sim N(\lambda, \psi)$, 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 $X_{n+1}|\boldsymbol{x} \sim N(\lambda, \psi + \sigma^2)$.

Média conhecida

- Seja $X_1|\boldsymbol{\theta}, ..., X_n|\boldsymbol{\theta}, \boldsymbol{\theta} = (\mu, \sigma^2)$ uma amostra aleatória de $X|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$ (exercício):
  - Se μ conhecido, e $\sigma^2 \sim IG(a, b)$, (família conjugada) então $\sigma^2|\boldsymbol{x} \sim IG(a^*, b^*)$, 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 $X_{n+1}|\boldsymbol{x} \sim t_{(2a^*)}\left(\mu, \sqrt{\frac{b^*}{a^*}}\right)$

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

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

- continua no próximo slide.

Cont.

- (Cont.) em que
$$c = \frac{n\overline{x} + \nu\lambda}{n+\nu},\ b^* = \frac{1}{2}\left[\frac{n\nu}{n+\nu}(\overline{x}-\lambda)^2 + (n-1)s^2\right] + b,$$
$$\nu^* = \nu + n,\ a^* = \tfrac{n}{2} + a \text{ e } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$
- Além disso, $\mu|\boldsymbol{x} \sim t_{(2a^*)}\left(c, \sqrt{\frac{b^*}{\nu^* a^*}}\right)$.
- Distribuição preditiva à posteriori para uma única observação
$X_{n+1}|\boldsymbol{x} \sim t_{(2a^*)}\left(c, \sqrt{\frac{b^*}{a^*\nu^{**}}}\right)$, 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

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

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,

$$\begin{aligned} \mu_1|(\sigma^2,\boldsymbol{x},\boldsymbol{y}) &\sim N(\overline{x},\sigma^2/n) \\ \mu_2|(\sigma^2,\boldsymbol{x},\boldsymbol{y}) &\sim N(\overline{y},\sigma^2/m) \\ \sigma^2|(\boldsymbol{x},\boldsymbol{y}) &\sim IG(k/2,ks^2/2) \end{aligned}$$

Exemplo 11 (cont.)

- Assim, se $\lambda = \mu_1 - \mu_2$, então
  $\lambda|(\sigma^2, \boldsymbol{x}, \boldsymbol{y}) \sim N\left(\overline{x} - \overline{y}, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$.
- Logo, $\lambda|(\boldsymbol{x}, \boldsymbol{y}) \sim t_{(k)}\left(\overline{x} - \overline{y}, s\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}\right)$, $s = \sqrt{s^2}$.
- Portanto, podemos utilizar a distribuição a posteriori acima para verificar se $\mu_1 = \mu_2$.

Exemplo 11 (cont.)

- 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,

$$\tau = \frac{\lambda - (\overline{x} - \overline{y})}{s\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}} | (\boldsymbol{x}, \boldsymbol{y}) \sim t_{(k)}(0,1)$$

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:

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

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

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

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

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

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

em que

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

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

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

$u = arctan\left[\left(s_1/\sqrt{n}\right)\left(s_2/\sqrt{m}\right)\right]$

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

Exemplo 12 (cont.)

- Uma vez que $\mu_1|(\boldsymbol{x},\boldsymbol{y}) \sim t_{k_1}(\overline{x}, s_1/\sqrt{n}) \perp \mu_2|(\boldsymbol{x},\boldsymbol{y}) \sim t_{k_2}(\overline{y}, s_2/\sqrt{m})$, podemos simular R variáveis aleatórias, mutuamente independentes, com distribuições t de Student, específicas, e obter λ para cada par, ou seja:
  - Simular $(\mu_1^{(r)}, \mu_2^{(r)})', r = 1, ..., R$ (das respectivas distribuições) e calcular $\lambda^{(r)} = \mu_1^{(r)} - \mu_2^{(r)}$ ou $\tau^{(r)} = \dfrac{\mu_1^{(r)} - \mu_2^{(r)} - (\overline{x} - \overline{y})}{\sqrt{\dfrac{s_1^2}{n} + \dfrac{s_2^2}{m}}}$.

Exemplo 12 (cont.)

- No caso de termos uma amostra numérica das posteriori (ao invés da própria densidade), podemos obter o HPD numericamente (ou seja, resolve numericamente o sistema de equações necessário para se obter o HPD, veja aqui, slide 31), através da função "emp.hpd" do pacote "TeachingDemos".

Exemplo 12 (cont.)

- Para comparar as variâncias, basta notar que

$$\frac{k_1 s_1^2}{\sigma_1^2}|(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2_{(k_1)} \perp \frac{k_2 s_2^2}{\sigma_2^2}|(\boldsymbol{x}, \boldsymbol{y}) \sim \chi^2_{(k_2)}$$

e, assim

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

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

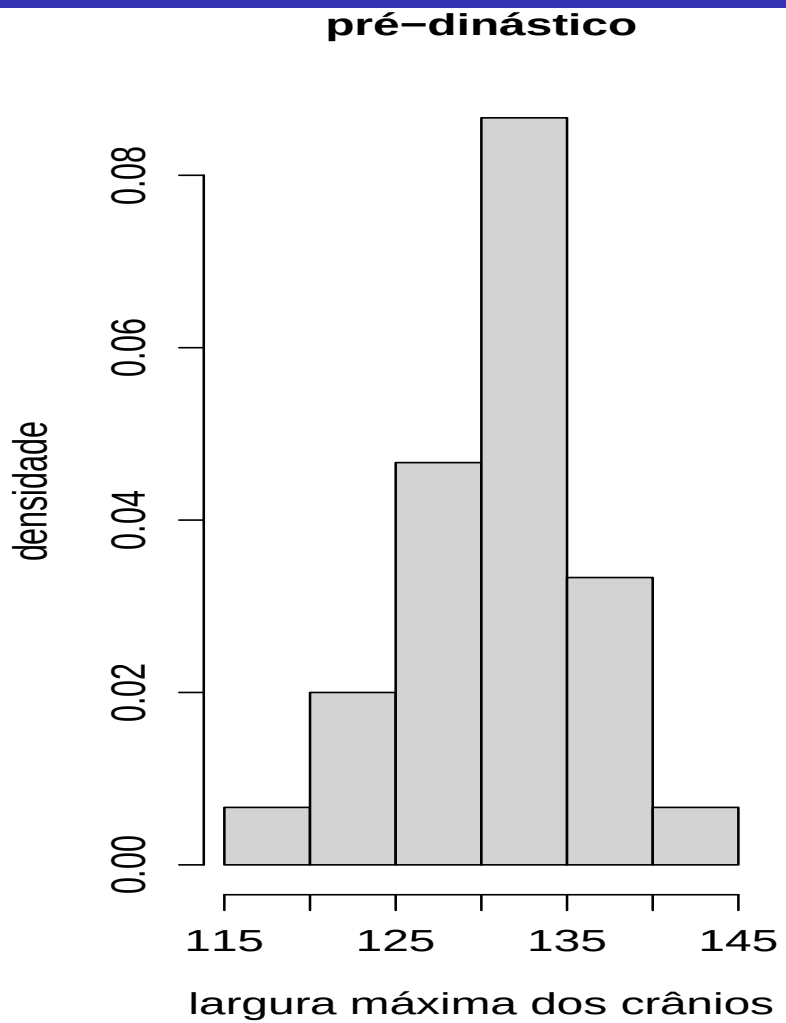
Dados

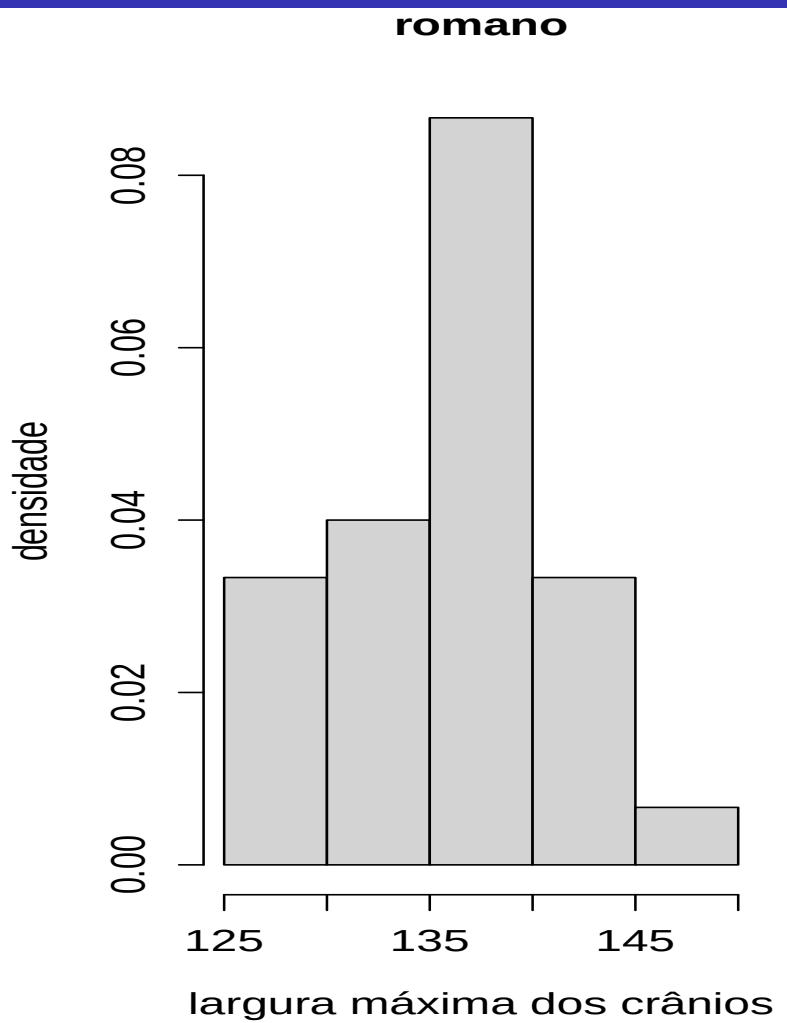
- O conjunto de dados se refere à n = m = 30 observações correspondentes às larguras máximas de crânios humanos, datadas do período pré-dinástico (grupo 1, PD) e romano (grupo 2, R), respectivamente.
- Objetivo principal: comparar as médias populacionais das larguras máximas entre os tipos de crânios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

Medidas resumo

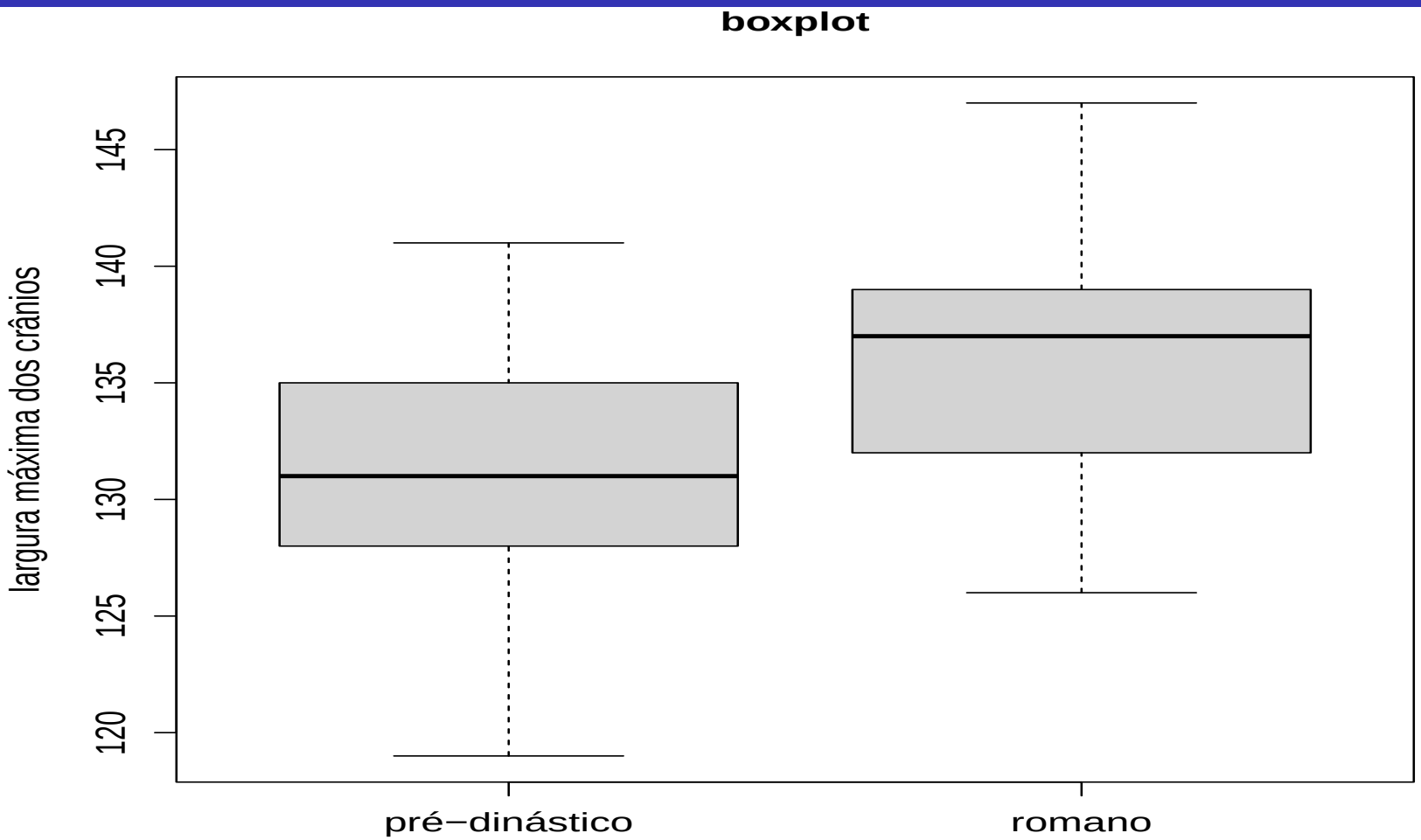
MR	Pré-dinástico	Romano
média	131,37	136,17
variância	26,31	28,63
dp	5,13	5,35
cv(%)	3,90	3,93
mediana	119,00	126,00
mínimo	131,00	137,00
máximo	141,00	147,00
ca	-0,18	-0,12
curtose	2,69	2,51

Histogramas

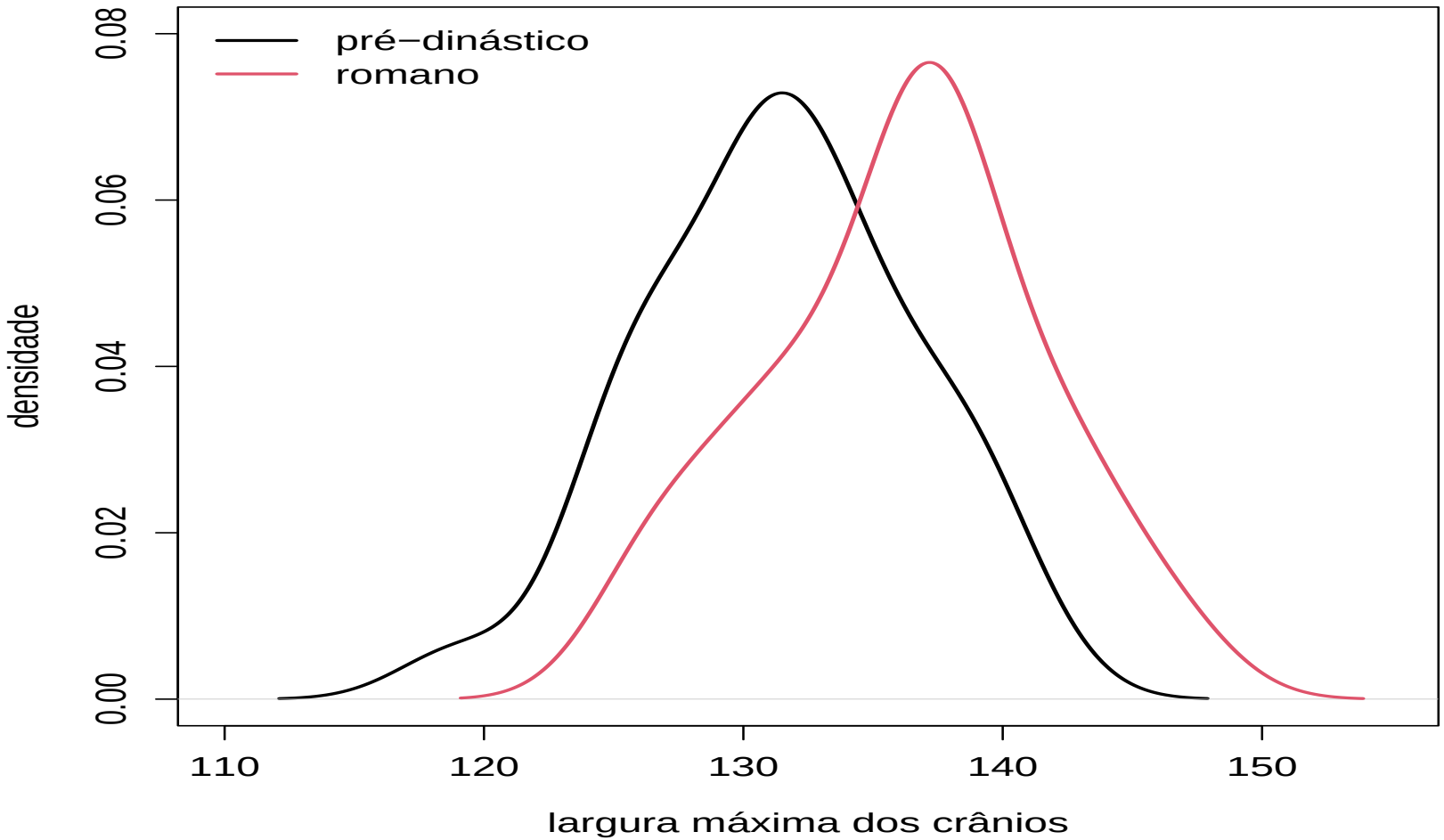




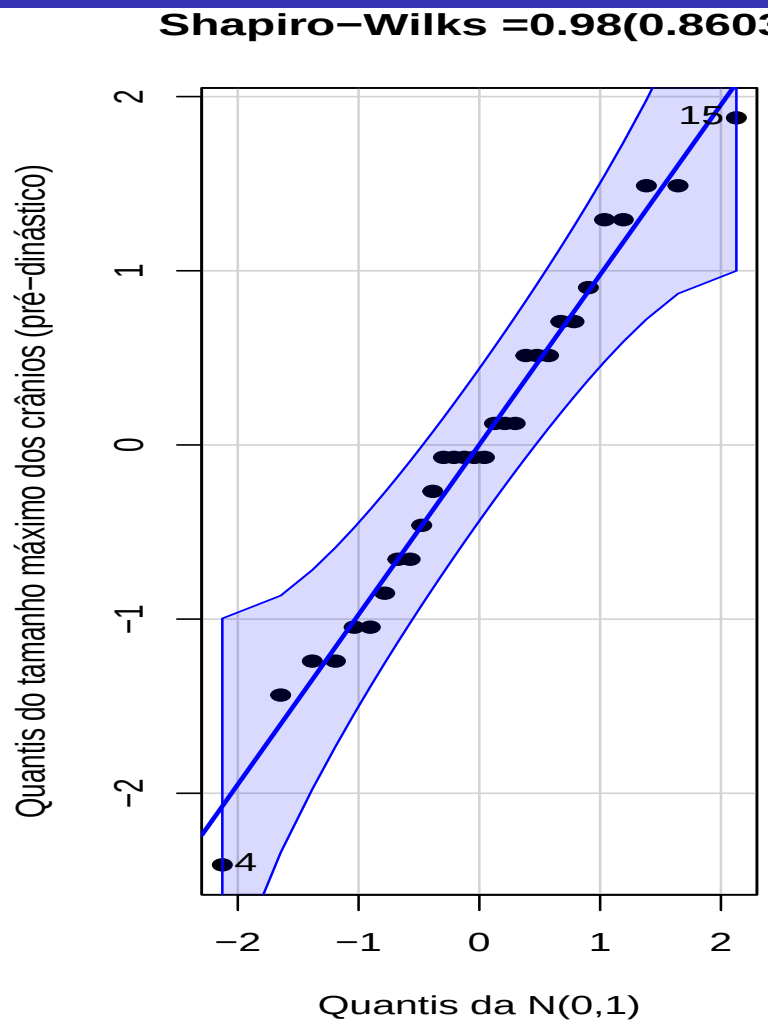
Boxplots

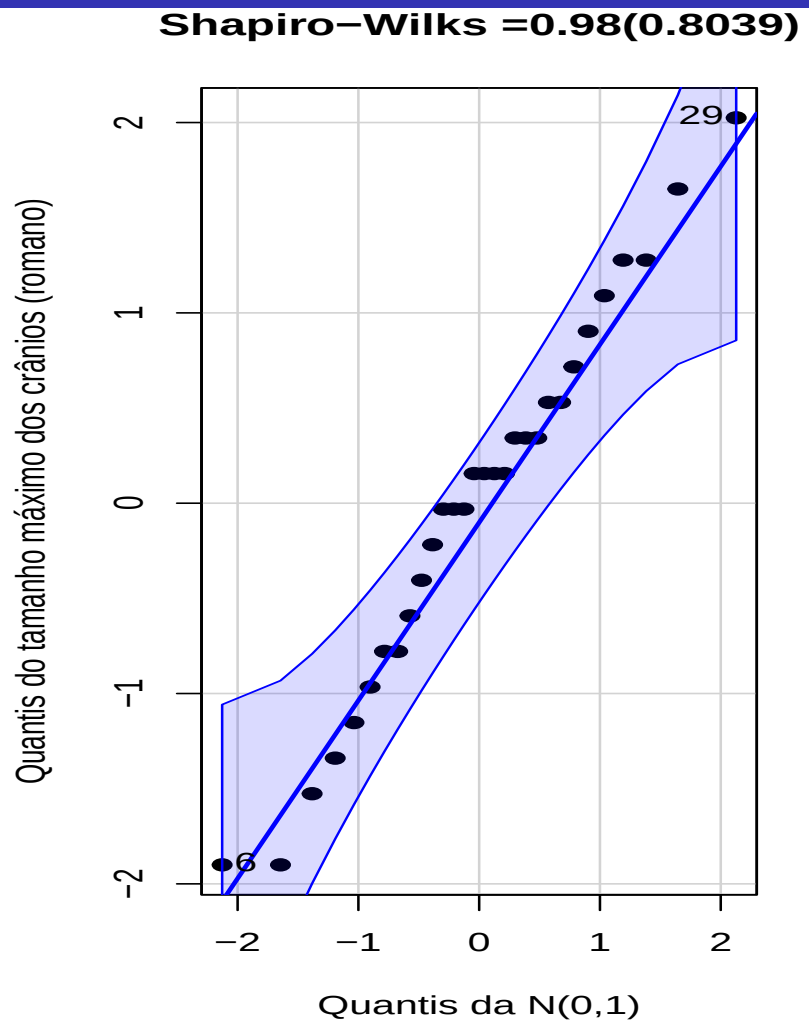

boxplot
largura máxima dos crânios
120
125
130
135
140
145
pré−dinástico
romano

Densidades

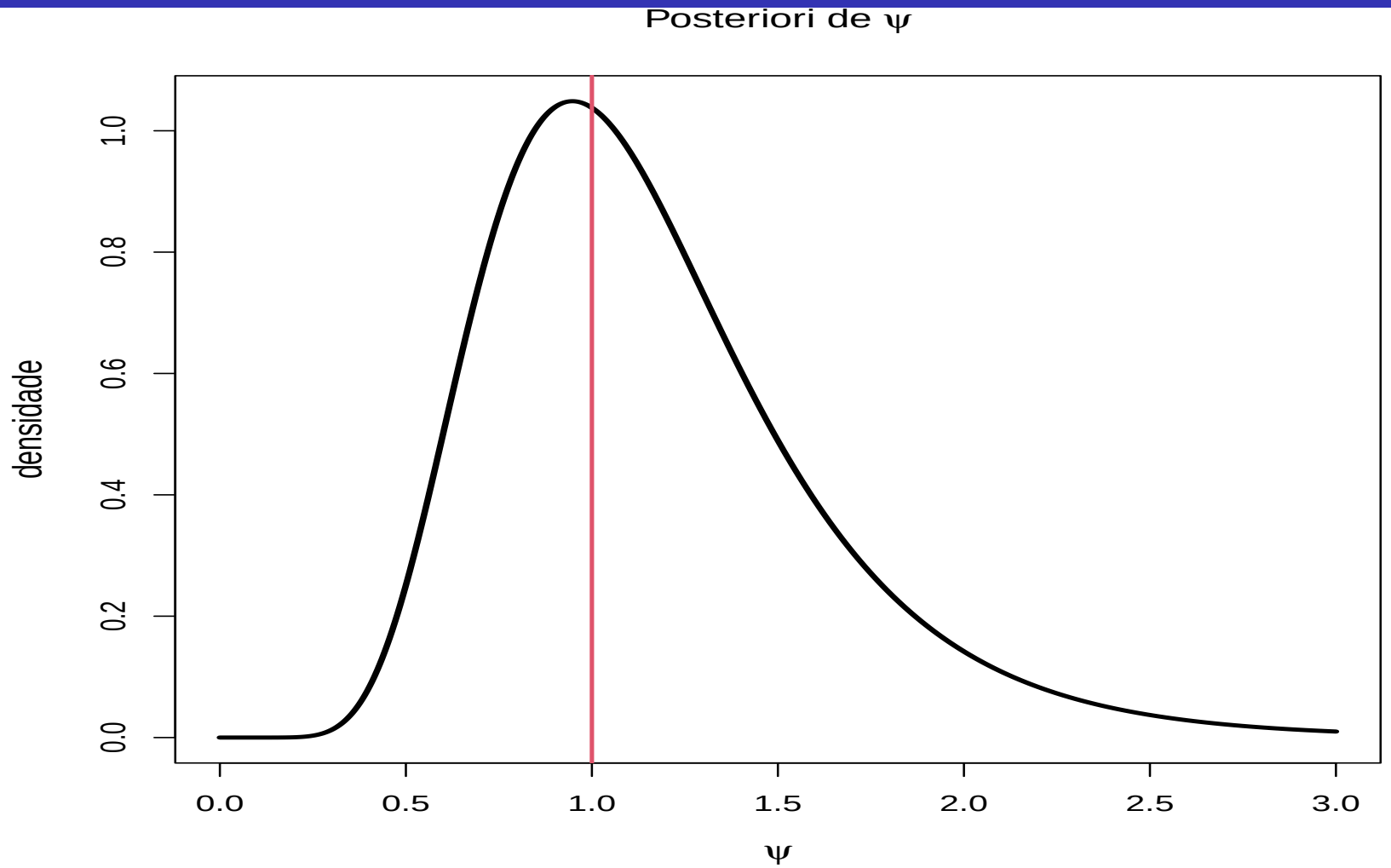

pré−dinástico
romano
densidade
0.00
0.02
0.04
0.06
0.08
110
120
130
140
150
largura máxima dos crânios

Gráficos de Quantis-quantis N(0,1)





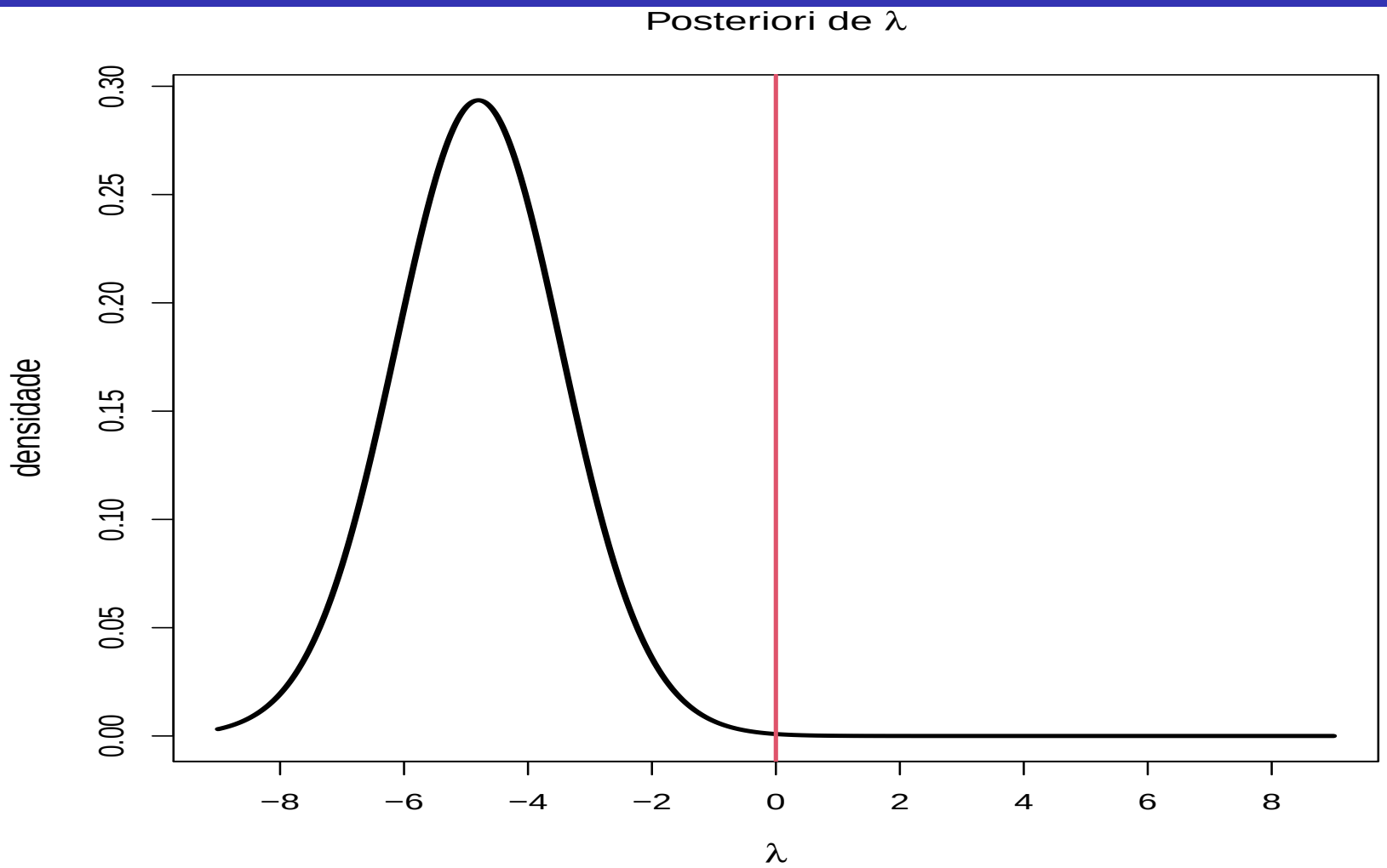
Posteriori de ψ



Comparação das variâncias

- $IC_B(\psi; 0,95) = [0,437; 1,931]$, $(CIC = 1,493)$.
- $HPD(\psi; 0,95) = [0,360; 1,757]$, $(CIC = 1,397)$.
- 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%.

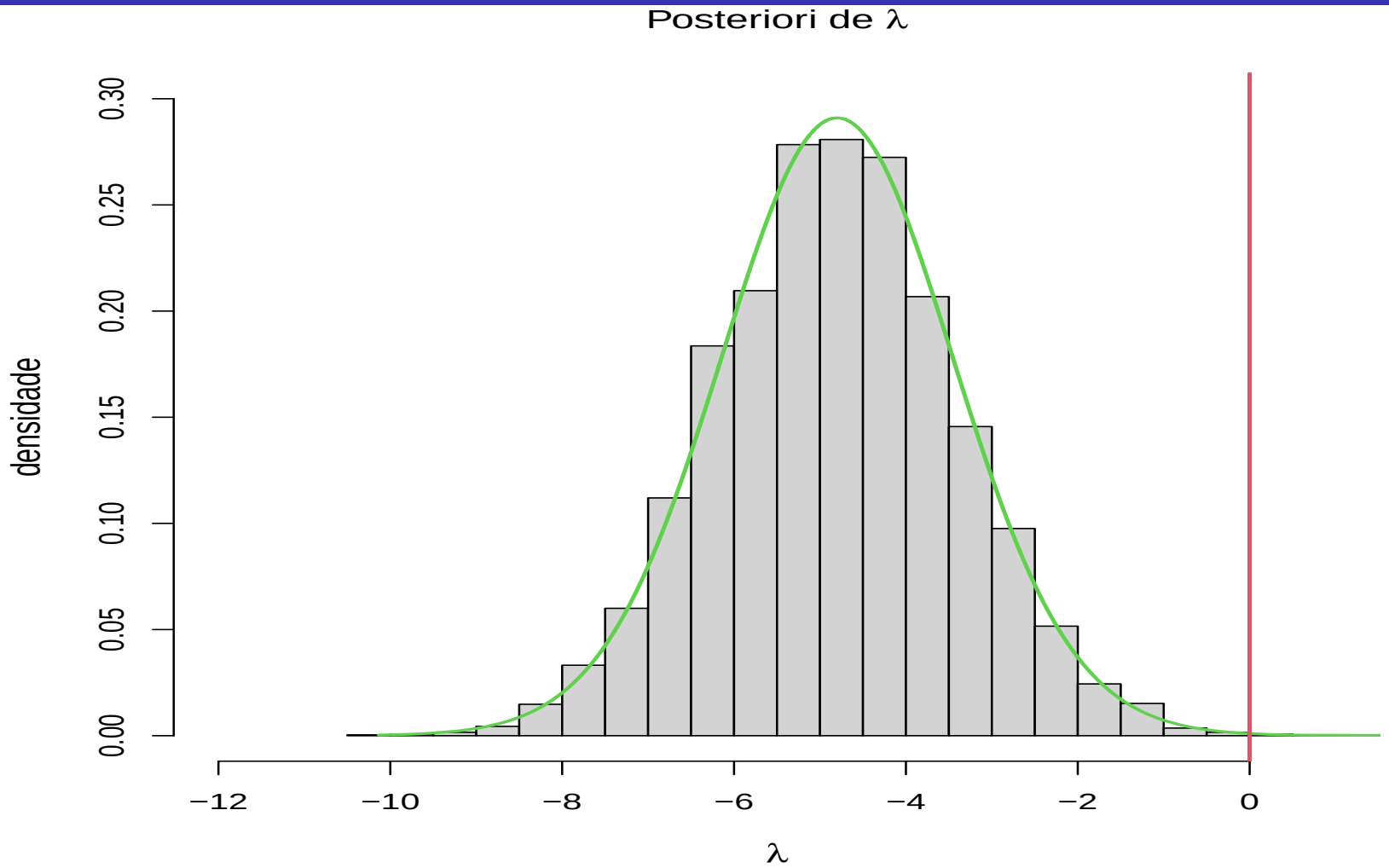
Posteriori de λ, $\sigma_1^2 = \sigma_2^2$



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).
- $P(\lambda < 0|\boldsymbol{x}) = 0,9996$.
- Os resultados acima nos levam à concluir que $\mu_1 < \mu_2$ com uma credibilidade de 95%. Ou seja, que a média do tamanho máximo dos crânios do período pré-dinástico é menor do que aquela para o período romano.

Posterioris de λ, $\sigma_1^2 \neq \sigma_2^2$, $R = 5.000$



Comparação das médias, supondo as variâncias diferentes

- Aproximação analítica:
  - $IC_B(\psi; 0,95) = HPD(\psi; 0,95) = [-7,536; -2,064]$, $(CIC = 5,473)$ (pois a posteriori é simétrica e unimodal, neste caso, os resultados são aproximados).
  - $P(\lambda < 0|\boldsymbol{x}) = 0,9996$.
- Aproximação numérica:
  - $IC_B(\psi; 0,95) = HPD(\psi; 0,95) = [-7,660; -2,174]$, $(CIC = 5,417)$ (pois a posteriori é simétrica e unimodal, neste caso, os resultados são aproximados).
  - $P(\lambda < 0|\boldsymbol{x}) = 0,9992$.

Comparação das médias, supondo as variâncias diferentes

- Os resultados acima nos levam à concluir que $\mu_1 < \mu_2$ com uma credibilidade de 95%. Ou seja, que a média do tamanho máximo dos crânios do período pré-dinástico é menor do que aquela para o período romano.
- (Exercício/Pesquisar) Como resposta final, além do fato de reportar que as médias são diferentes, devemos encontrar as posteriores marginais $\mu_1|\boldsymbol{x}, \boldsymbol{y}$ e $\mu_2|\boldsymbol{x}, \boldsymbol{y}$ e prover estimativas pontuais e intervalares par cada média.

Comparação das médias, supondo as variâncias diferentes

- Considerando as variâncias conhecidas, teríamos, usando as posterioris do slide 14:
  - Pré-dinástico: 131,37 (0,97); [129,45;133,28].
  - Romano: 136,17 (1,01); [134,17;138,16].