Métodos de reamostragem

Prof. Caio Azevedo

Notações

- Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra iid de $X \sim F(, \theta)$, $\theta \in \Theta$ pode ser vetor.
- Seja $\mathbf{x} = (x_1, ..., x_n)$ os valores observados da amostra definida acima.
- Estimador : $\widehat{\theta} = f(\mathbf{X})$, estimativa :$\widetilde{\theta} = f(\mathbf{x})$.
- Variância: $Var(\widehat{\theta})$ e erro-padrão: $EP(\widehat{\theta}) = \sqrt{Var(\widehat{\theta})}$ associados ao estimador.

Notações cont.

- Estimadores para a variância $\widehat{Var}(\widehat{\theta})$ e o erro-padrão $\widehat{EP}(\widehat{\theta}) = \sqrt{\widehat{Var}(\widehat{\theta})}$ associados ao estimador.
- Estimativas para a variância $\widetilde{Var}(\widehat{\theta})$ e o erro-padrão $\widetilde{EP}(\widehat{\theta}) = \sqrt{\widetilde{Var}(\widehat{\theta})}$ do associados ao estimador.

Exemplo $N(\mu, \sigma^2)$

- Seja $X_1, ..., X_n$ uma amostra iid de $X \sim N(\mu, \sigma^2)$, (μ, σ^2) desconhecidos e $x_1, ..., x_n$ valores observados.
- Estimador para μ, $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. Estimativa $\widetilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$.
- Variância do estimador: $Var(\widehat{\mu}) = \frac{\sigma^2}{n}$. Estimador para a variância do estimador $\widehat{Var}(\widehat{\theta}) = \frac{\widehat{\sigma}^2}{n}$.
- Estimativa para a variância do estimador $\widetilde{Var}(\widehat{\theta}) = \frac{\widetilde{\sigma}^2}{n}$.

Exemplo $N(\mu, \sigma^2)$

- Seja $X_1, ..., X_n$ uma amostra iid de $X \sim N(\mu, \sigma^2)$, (μ, σ^2) desconhecidos.
- Dispomos de valores observados relativos à uma amostra selecionada a partir de uma população de interesse.
- Estimadores ótimos (consistentes e não-viciados),

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = S^2.$$

Cont.

- Sabemos que $\widehat{\mu} \sim N(\mu, \sigma^2/n)$ e $\frac{(n-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-1)}$.
- Podemos constuir intervalos de confiança (IC) e testar hipóteses de interesse, com base no resultado acima.
- Mesmo que $X_1, ..., X_n$ não tenham distribuição normal (aproximadamente) se a sequência a qual elas pertencem satisfizer o TCL, os resultados acima são válidos, assintoticamente.
- Questão: suponha que queremos fazer inferência com respeito à $\theta = \frac{\sigma}{\mu}, \sigma = \sqrt{\sigma^2}$, sem usar resultados assintóticos e sob suposição de normalidade.

Cont.

- Estimador natural de θ, $\widehat{\theta} = \frac{\widehat{\sigma}}{\widehat{\mu}}, \widehat{\sigma} = \sqrt{\widehat{\sigma^2}}$.
- Qual a distribuição exata de $\widehat{\theta}$? Qual a expressão para o erro-padrão associado à este estimador?
- Podemos obter uma aproximação empírica (não assintótica) de $\widehat{\theta}$ através da suposição de normalidade.
- Replicas de Monte Carlo.

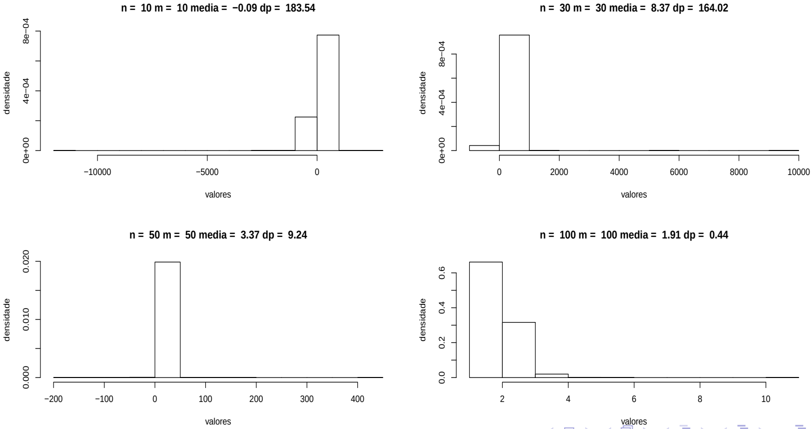
Procedimento

- Com base na amostra observada $\mathbf{x}$, estima-se μ e σ^2 através do estimadores propostos.
- Com as estimativas $\widetilde{\mu}$ e $\widetilde{\sigma}^2$, gera-se R vezes amostras de tamanho m, ou seja temos para $r = 1, ..., R$, m valores simulados de uma $N(\widetilde{\mu}, \widetilde{\sigma}^2)$, $X_{r1}, ..., X_{rm}$.
- Para cada conjunto r de valores simulados, calcula-se $\widetilde{\mu}_r^* = \frac{1}{m}\sum_{i=1}^m x_{ri}$, $(\widetilde{\sigma}^*)_r^2 = \frac{1}{m-1}\sum_{i=1}^m (x_{ri} - \widetilde{\mu}_r^*)^2$ e $\widetilde{\theta}_r^* = \frac{\sqrt{(\widetilde{\sigma}^*)_r^2}}{\widetilde{\mu^*}_r}$.
- Assim, os valores $\widetilde{\theta}_1^*, ..., \widetilde{\theta}_R^*$ serão uma amostra aleatória de $\widehat{\theta}$.

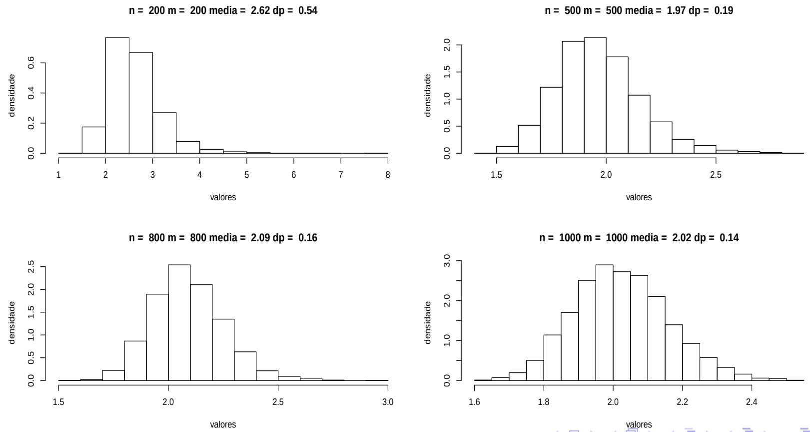
Resultados

- Uma estimativa (refinada) de θ é $\widetilde{\theta}^* = \frac{1}{R}\sum_{r=1}^{R}\widetilde{\theta}_r^*$ (podemos utilizar, também, a estimativa pontual baseada na amostra observada).
- Uma estimativa da variância do estimador proposto é $var(\widehat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\widetilde{\theta}^* - \widetilde{\theta}_r^*\right)^2$. Estimativa do vício $B(\widehat{\theta}) = (\widetilde{\theta}^* - \widetilde{\theta})$
- Sejam $\widetilde{\theta}_{(1)}^*, \widetilde{\theta}_{(2)}^*, ..., \widetilde{\theta}_{(R)}^*$ as estatísticas de ordem da amostra gerada.
- Quantis empíricos podem ser obtidos através de um dos algoritmos propostos por Hyndman and Fan (1996).
- Função "quantile" no R.

Amostras Bootstrap 1



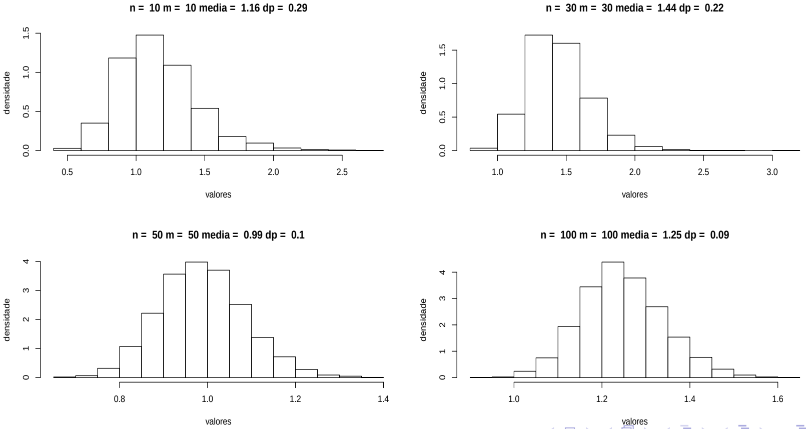
Amostras Bootstrap 2



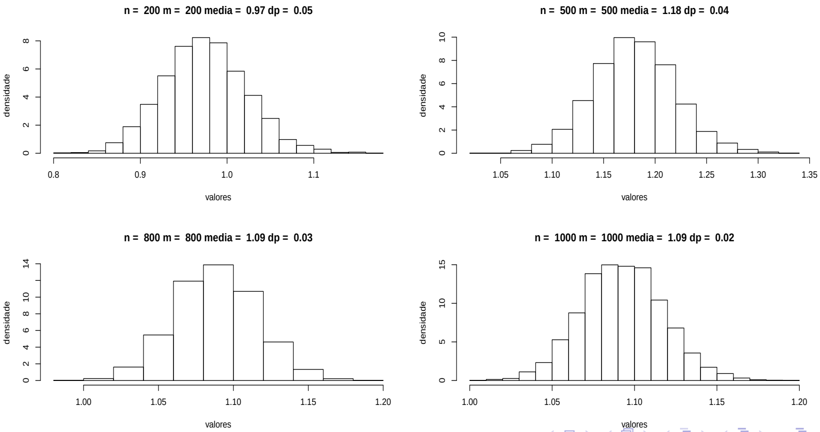
Exemplo $beta(a, b)$

- Suponha uma amostra aleatória de uma distribuição beta(a,b).
- Com base na amostra observada $\mathbf{x}$, estima-se a e b através do estimadores do método dos momentos.
- Com as estimativas $\widetilde{a}$ e $\widetilde{b}$, gera-se R vezes amostras de tamanho m, ou seja temos para $r = 1, ..., R$, m valores simulados de uma $beta(\widetilde{a}, \widetilde{b})$, $X_{r1}, ..., X_{rm}$.
- Para cada conjunto r de valores simulados, calcula-se $\widetilde{\mu}_r^* = \frac{1}{m}\sum_{i=1}^{m} x_{ri}$, $(\widetilde{\sigma}^*)_r^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_{ri} - \widetilde{\mu}_r^*)^2$ e $\widetilde{\theta}_r^* = \frac{\sqrt{(\widetilde{\sigma}^*)_r^2}}{\widetilde{\mu}_r^*}$.
- Assim, os valores $\widetilde{\theta}_1^*, ..., \widetilde{\theta}_R^*$ serão uma amostra aleatória de $\widehat{\theta}$.

Amostras Bootstrap 1



Amostras Bootstrap 2



Bootstrap paramétrico

- Sob a suposição de um modelo paramétrico $F(, \boldsymbol{\theta})$ e com a amostra selecionada, estima-se $\boldsymbol{\theta}$ através de algum método apropriado.
- Com a estimativa $\widetilde{\boldsymbol{\theta}}$, gera-se R vezes uma amostra aleatória de tamanho m $(m \leq n)$ da amostra $\mathbf{x} = (x_1, ..., x_n)$.
- Para cada amostra r $r = 1, ..., R$ obtem-se a estimativa desejada $\widetilde{\theta}_r^*$.
- Estimativa Bootstrap $\widetilde{\theta}^* = \frac{1}{R} \sum_{r=1}^{R} \widetilde{\theta}_r^*$

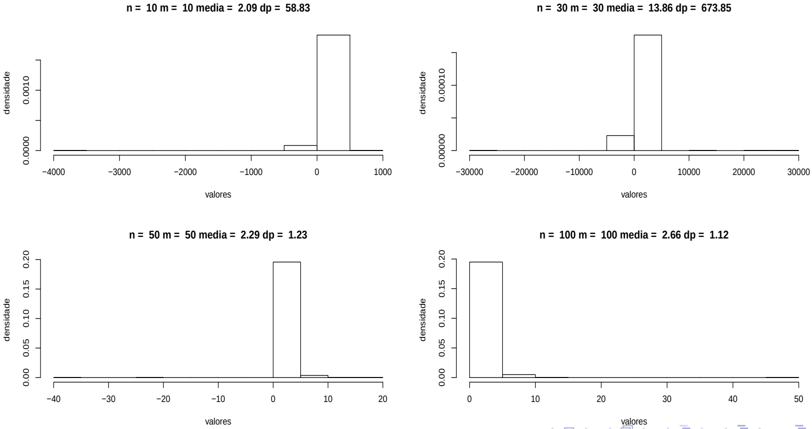
Caso geral

- Suponha agora que não conhecemos a distribuição (mesmo uma boa aproximação) geradora da amostra
- Não queremos fazer nenhuma suposição a respeito da distribuição.
- Caso conheçamos ou queiramos fazer alguma suposição, por algum motivo, é difícil de gerar da variável aleatória em questão.

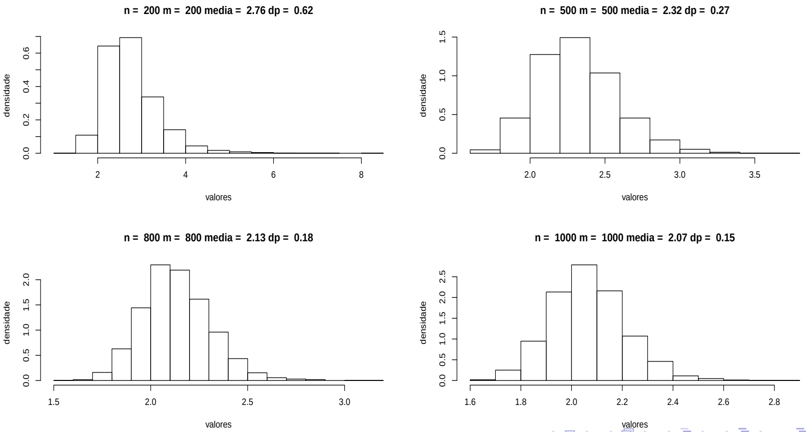
Bootstrap não-paramétrico

- Seleciona-se R vezes uma amostra aleatória (com reposição) de tamanho m ($m \leq n$) da amostra $\mathbf{x} = (x_1, ..., x_n)$.
- Para cada amostra r $r = 1, ..., R$ obtem-se a estimativa desejada $\widetilde{\theta}_r^*$.
- Estimativa Bootstrap $\widetilde{\theta}^* = \frac{1}{R}\sum_{r=1}^{R}\widetilde{\theta}_r^*$
- Bootstrap paramétrico: conhecemos a distribuição dos dados F e simulamos valores a partir dessa distribuição usando estimativas dos parâmetros relacionados à ela.
- Bootstrap não-paramétrico: não conhecemos a distribuição dos dados F e simulamos valores a partir da fda empírica.

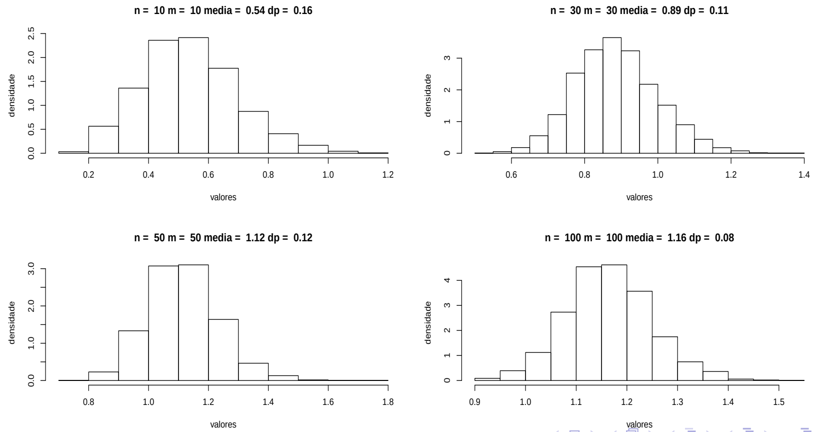
Amostras Bootstrap Não-paramétrico Normal (1)



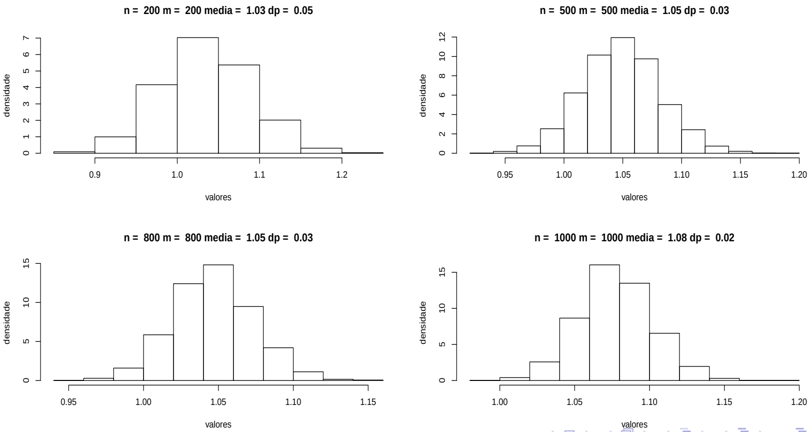
Amostras Bootstrap Não-paramétrico Normal (2)



Amostras Bootstrap Não-paramétrico Beta (1)



Amostras Bootstrap Não-paramétrico Beta (2)



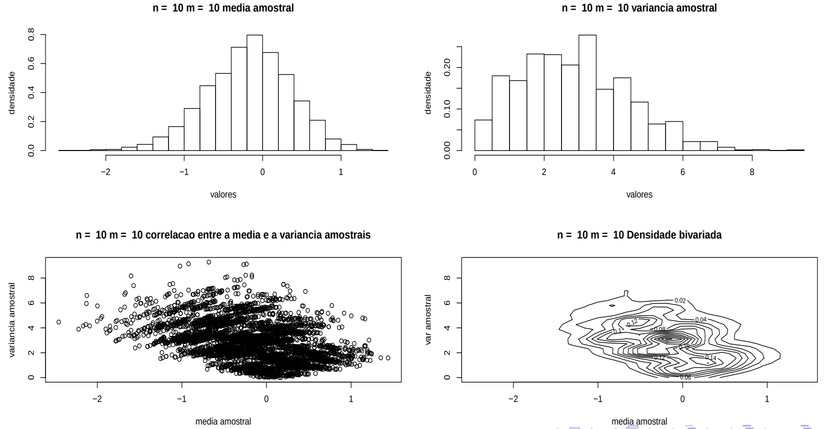
Comparações

- Bootstrap paramétrico: conhecemos a distribuição dos dados F e simulamos valores a partir dessa distribuição usando estimativas dos parâmetros relacionados à ela. Dependente: qualidade das estimativas e do método de simulação escolhido.
- Bootstrap não-paramétrico: não conhecemos a distribuição dos dados F e simulamos valores a partir da fda empírica. Com ou sem reposição.

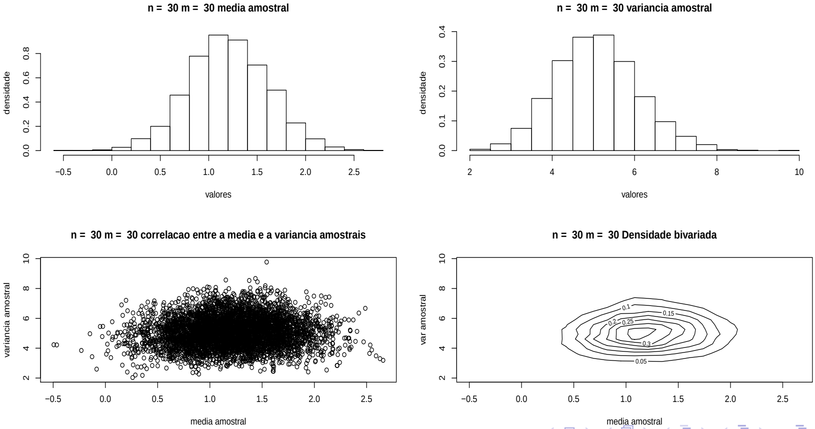
Exemplo: correlação entre média e variância amostrais

- Sabemos que a corre$(\overline{X}, \widehat{\sigma}^2) = 0$, se $X \sim N(\mu, \sigma^2)$.
- Contudo, isso não esperado para outras variáveis aleatórias.
- Ademais, o TCL sugere que a distrbuição conjunta e marginais de $\overline{X}$ e $\widehat{\sigma}^2$, sejam normais.

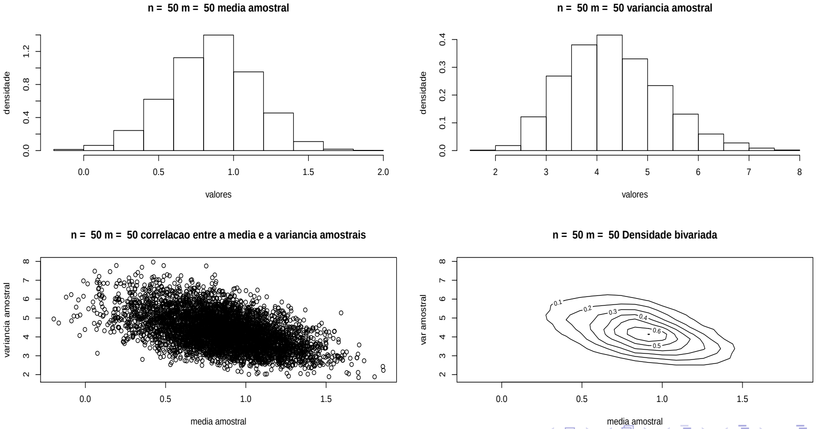
Amostras Bootstrap de uma normal, n=10



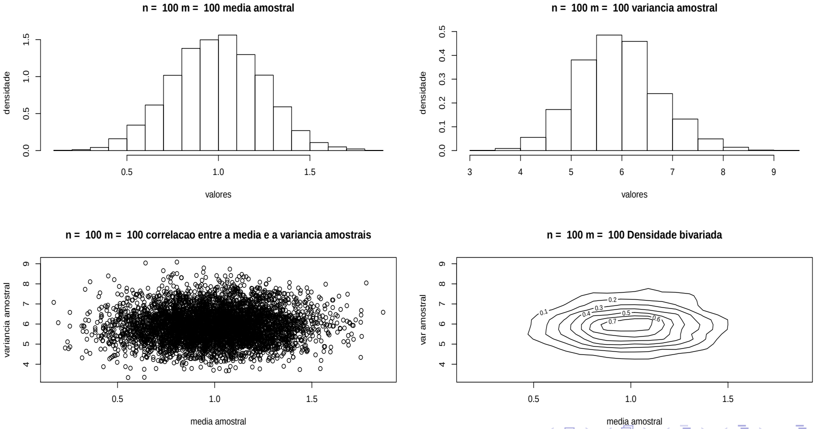
Amostras Bootstrap de uma normal, n=30



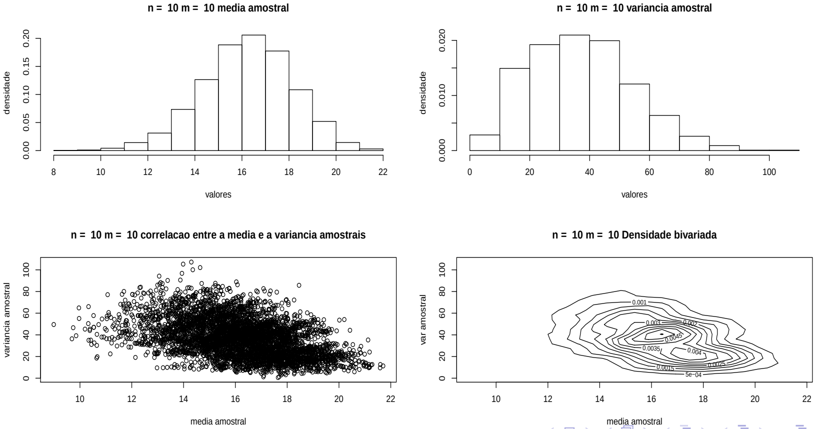
Amostras Bootstrap de uma normal, n=50



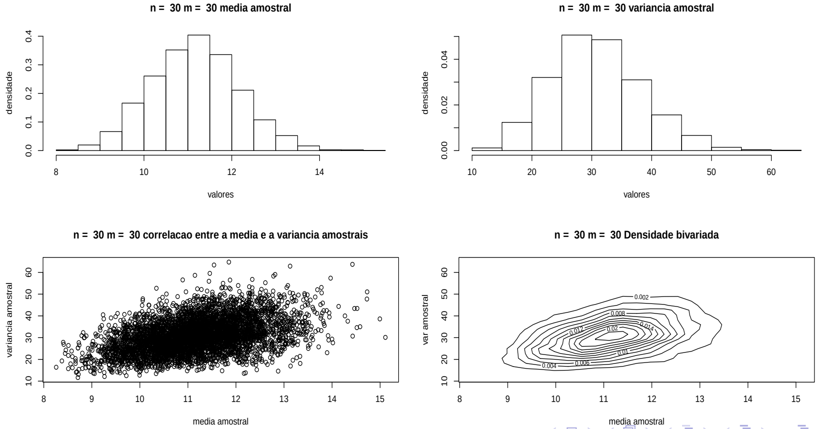
Amostras Bootstrap de uma normal, n=100



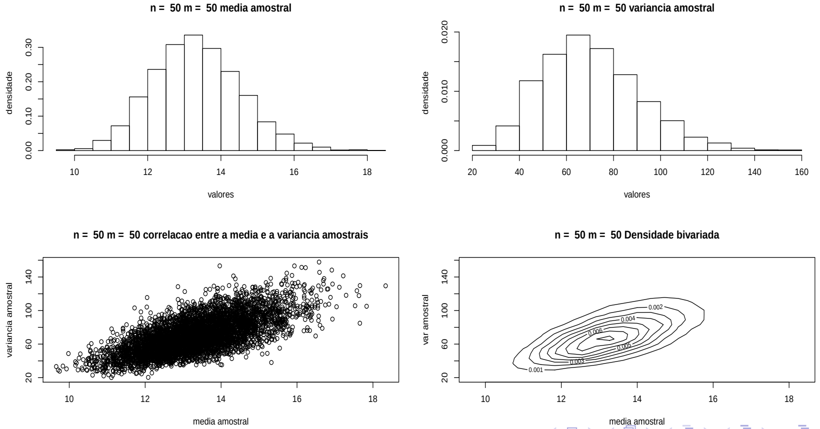
Amostras Bootstrap de uma gama, n=10



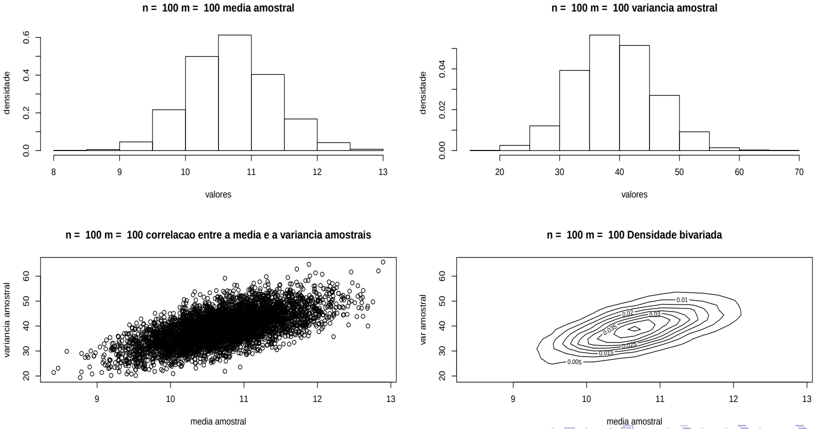
Amostras Bootstrap de uma gama, n=30



Amostras Bootstrap de uma gama, n=50



Amostras Bootstrap de uma gama, n=100



Resumo

- Queremos estimar θ relativo ou não à $F(,\theta)$. Para isso, dispõe-se de uma amostra observada de $F(;\theta)$.
- Define-se um estimador $\widehat{\theta}$, eg, máxima verossimilhança, momentos, mínimos quadrados. Sejam $\mathrm{Var}(\widehat{\theta})$ e $\mathrm{EP}(\widehat{\theta})$ a variância e o erro-padrão desse estimador (se conhecidas).
- Paramétrica (usando-se $F(;\widetilde{\theta})$) ou não parametricamente (usando a fda empírica da amostra observada), simula-se R vezes valoes, $x_{r1},...,x_{rm}$.
- Em cada uma das r amostras, calcula-se $\widetilde{\theta}_1^*,...,\widetilde{\theta}_R^*$, estimativas bootstrap.

Resumo cont.

- Estimativa (refinada) bootstrap $\widetilde{\theta}^* = \frac{1}{R}\sum_{r=1}^{R}\widetilde{\theta}_r^*$
- Definimos $\widetilde{Var}_B(\widehat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}(\widetilde{\theta}_r^* - \widetilde{\theta}^*)^2$ e $\widetilde{EP}_B(\widehat{\theta}) = \sqrt{\widetilde{Var}_B(\widehat{\theta})}$, estimativas bootstrap para variância e o erro-padrã do estimador em questão.

Intervalos de confiança

- Usar o erro-padrão bootstrap, sob a suposição de normalidade (mesmo assintótica) do estimador utilizado.

$$(\widehat{\theta} - z_{\frac{1+\gamma}{2}} EP_B(\widehat{\theta}), \widehat{\theta} + z_{\frac{1+\gamma}{2}} EP_B(\widehat{\theta}))$$

$P(Z < z_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2}, Z \sim N(0,1)$, $EP_B()$ erro-padrão bootstrap (estimativa).

- Usar quantis empíricos das estimativas bootstrap.

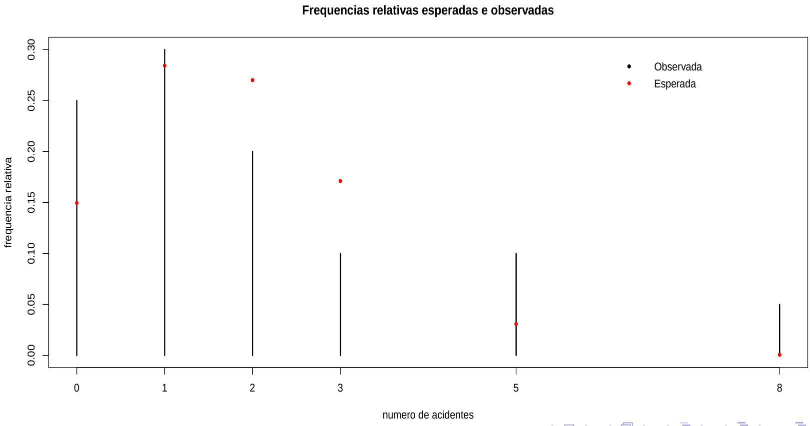
$$(\kappa_{\frac{1-\gamma}{2}}, \kappa_{\frac{1+\gamma}{2}}),$$

κ_α quantil de ordem α das estimativas bootstrap.

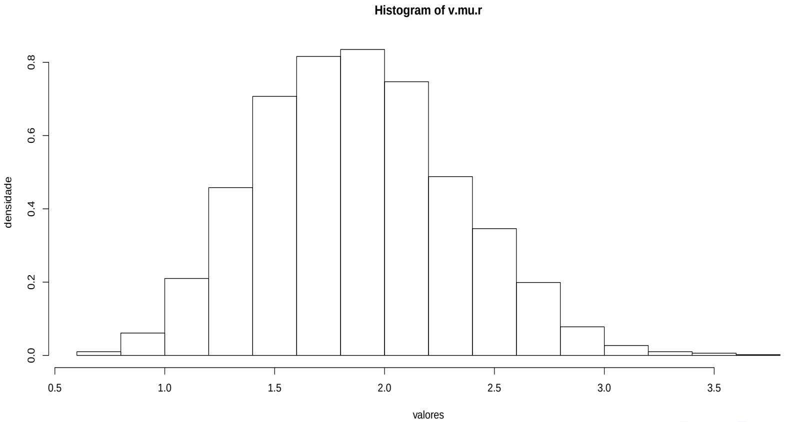
Exemplo: Poisson

- Dados relativos ao número de acidentes em um ano em 20 autoestradas estadunidenses.
- $\mathbf{x} = (1, 2, 5, 0, 3, 1, 0, 1, 1, 2, 0, 1, 8, 0, 5, 0, 2, 1, 2, 3)$.
- Objetivos, fazer inferências sobre μ e $\theta = \frac{\sigma}{\mu}$.
- Considere $\widehat{\mu} = \overline{X}$, $\widehat{\sigma} = \sqrt{S^2} = S$ e $\widehat{\theta} = \frac{S}{\overline{X}}$.
- Há indícios de superdispersão $\overline{x} = 1,90 < \widetilde{s}^2 = 4,31$.

Frequências relativas e observadas



Distribuição empírica do estimador



Resultados

- Estimativas intervalares e pontuais

Método	Estimativa	EP	IC(95%)
Assintótico	1,90	0,31	[1,30;2,50]
Assintótico com EP bootstrap	1,90	0,46	[1,01;2,79]
Bootstrap	1,91	0,46	[1,10;2,80]

Teste de Hipótese

- Estabeleça as hipóteses de interesse H_0, H_1.
- Escolha uma estatística do teste (digamos T) apropriada.

  Observação: Não é necessário conhecer a distribuição da estatistica sob H_0.
- Calcule o valor observado da estatística t_{obs}.
- Gere R amostras da distribuição sob H_0.
- Para cada amostras, calcule $t_r^*, r = 1, ..., R$.
- Calcule a proporção de vezes que os valores t_r^* foram "mais extremos" que t_{obs}.
- Rejeite H_0 de acordo com o nível de signficância α.

Observação

- "Mais extremo", depende de H_1.
- Para $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$, mais extremo signfica menor.
- Para $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$, mais extremo signifca maior.
- Para $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, mais extremo significa $|bootstrap| > |observado|$.

Cont. Exemplo: Poisson

- Suponha que queiramos testar $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, θ é a média da Poisson.
- Definse a estatística $T = \frac{\overline{X} - \theta_0}{\sqrt{\overline{X}/n}}$. Valor observado $t = -0,324$ (para $\theta_0 = 2$).
- Gera-se R vezes amostras (bootstrap não paramétrico) e calcula-se T para cada amostra $t_1^*, ..., t_R^*$.
- Dois p-valores $p_1 = \frac{\sum_{r=1}^{R} \mathbb{1}(|t_r^*| > |t|)}{R}$ e $p_2 = \frac{\sum_{r=1}^{R} \mathbb{1}(|t_r^*| > |t|) + 1}{R+1}$.
- $p_1 = 0,7642$ e $p_2 = 0,7642$.

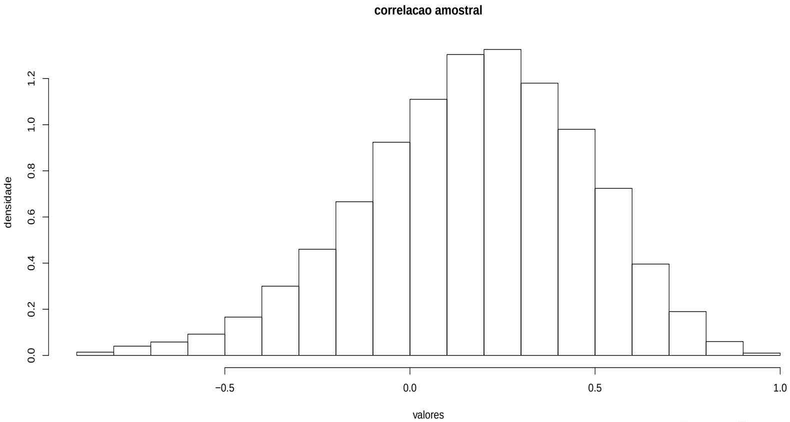
Teste de ajuste à uma distribuição

- Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra iid, supostamente de $X \sim F()$. Seja $F_1()$ uma outra FDA.
- Para $H_0 : F = F_1$ vs $H_1 : F \neq F_1$.
- Repetir o processo acima usando a estatística $D = \max(\widehat{F}_n(x) - F(x))$, em que $\widehat{F}_n()$ é fda empírica.
- Outas estatísticas podem ser usadas.
- No exemplo da Poisson, $H_0 : F \sim \text{Poisson}(1,9)$, $p_1 = 0,5526$ e $p_2 = 0,5527$

Exemplo: Correlações entre notas, Matemática e Música

- Dados relativos as notas de n $= 12$ crianças em testes de Matemática e música.
- Matemática: $\mathbf{x} = (60, 58, 73, 51, 54, 75, 48, 72, 75, 83, 62, 52)$.
- Música: $\mathbf{y} = (80, 62, 70, 83, 62, 92, 79, 88, 54, 82, 64, 69)$.
- Objetivos, fazer inferências sobre $\rho =$ correlação(X, Y).
- Testes paramétricos baseados na correlação de Pearson amostral e testes não-paramétricos baseados nas correlações de Kendall e Spearman.

Histograma das amostras bootstraps



Resultados

- Estimativas intervalares e pontuais: $\widetilde{\rho} = 0,169(0,300)$.

Método	IC(95%)
Normal	[-0,422; 0,754]
Básico	[-0,365; 0,808]
Percentil	[-0,470;0,703]
BCa	[-0,572;0,652]

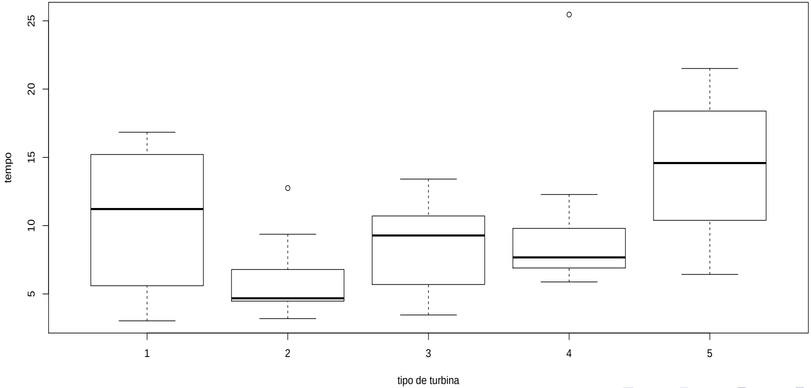
Teste de nulidade $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

- p-valores: 0,5992(Pearson) ; 0,6793 (Kendall) ; 0,6635 (Spearman).
- p-valor bootstrap via estatística $\frac{\widetilde{\rho}\sqrt{n-2}}{\sqrt{1-\widetilde{\rho}}}$: 0,7250.

Exemplo: Modelo linear generalizado

- Seja $Y_i \sim \text{gama}(\mu_i, \phi), i = 1, ..., n$, independentes, $\mathcal{E}(Y_i) = \mu_i = \exp(\mathbf{X}_i'\beta), \mathcal{V}(Y_i) = \frac{\mu_i}{\phi}$.
- Estimar $(\boldsymbol{\beta}, \phi)$.
- Hipóteses de interesse $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$; $H_0 : \phi = 1$ vs $H_1 : \phi \neq 1$.
- Conjuntos de dados relativos ao desempenho de 5 tipos de turbina de avião. Tempo (em unidade de milhões de cilcos) até a perda de velocidade.

Boxplot dos dados das turbinas



Análise descritiva

- Medidas descritivas.

Tipo	Média	DP	CV(%)	Min	Max
1	10,69	3,27	30,58	3,03	16,84
2	6,05	2,46	40,66	3,19	12,75
3	8,64	2,94	34,03	3,46	13,41
4	9,80	3,13	31,95	5,88	25,46
5	14,71	3,83	26,08	6,43	21,51

Análise Inferencial

- Modelo: $Y_{ij} \sim \text{gama}(\mu_i, \phi), i = 1, 2, 3, 4, 5, j = 1, 2, ..., 10$

$$\mu_i = \mu + \alpha_i, \alpha_1 = 0$$

Análise Inferencial

- Estimativas de máxima verossimilhança.

Parâmetro	Est.	EP	Wald	p-valor (t)
μ	2,37	0,14	16,42	$<$ 0,0001
α_2	-0,57	0,20	-2,79	0,0076
α_3	-0,22	0,20	-1,05	0,3006
α_4	-0,09	0,20	-0,43	0,6704
α_5	0,32	0,20	1,56	0,1254

- $\widetilde{\phi} = 5,81(1,13),\ [3,59;8,01]$

Análise Inferencial Bootstrap

- Estimativas de máxima verossimilhança (bootstrap).

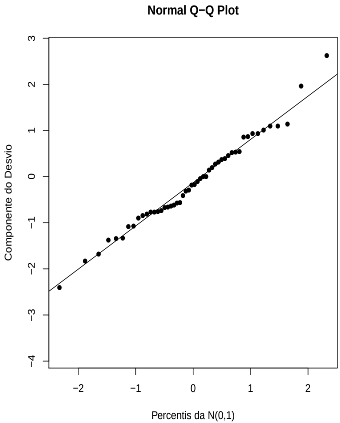
Parâmetro	Est.	EP	IC(%95)
μ	2,36	0,15	[2,04;2,62]
α_2	-0,58	0,21	[-0,96;-0,16]
α_3	-0,22	0,19	[-0,58;0,18]
α_4	-0,10	0,23	[-0,51;0,38]
α_5	0,32	0,18	[-0,02;0,68]

- $\widetilde{\phi} = 6,85(1,73)$, $[4,52;11,19]$

Verificação do ajuste do modelo

- Resíduos componente do desvio: sob as validades do modelo devem apresentar distribuição aproximadamente normal padrão.
- Comparar os quantis dos resíduos, com os resíduos teóricos da N(0,1).
- Visualmente, muitas vezes, é complicado avaliar a proximidade dos quantis.
- Solução: criar bandas de confiança empíricas.

Gráfico QQplot



Procedimento para a construção do gráfico de envelope

1. Ajustar o modelo de regressão, obtendo as estimativas $\widetilde{\beta}, \widetilde{\phi}$ e os resíduos componente do desvio, $r_1, ..., r_n$.
2. Com as estimativas originais, obtidas no passo 1, gerar n valores, $Y_1^*, ..., Y_n^*$, ajustar novamento o modelo de regressão e calcular os resíduos componenente do desvio.
3. Repetir o passo 2, m vezes, obtendo-se assim, $r_{ij}^*, i = 1, ..., n; j = 1, ..., m$
4. Colocamos cada grupo de n resíduos em ordem crescente, $r_{(i)j}^*, i = 1, ..., n; j = 1, ..., m$

Continuação

1. Para cada um dos m grupos, obtemos os limites $r^*_{(i)I} = min_j r^*_{(i)j}$ e $r^*_{(i)S} = max_j r^*_{(i)j}$. Assim, os limites correspondentes ao i-ésimo resíduo serão dados por $r^*_{(i)I}$ e $r^*_{(i)S}$.

Gráfico QQplot com bandas de confiança