BOOTSTRAP

INTRODUÇÃO

- IDEIA BÁSICA: reamostrar de um conjunto de dados, diretamente ou via um modelo ajustado, a fim de criar replicas dos dados, a partir das quais podemos avaliar a variabilidade de quantidades de interesse, sem usar cálculos analíticos.

- APLICAÇÃO DO MB: podem ser aplicados quando existe, ou não, um modelo probabilístico bem definido para os dados.

- METODO: COMPUTER-INTENSIVE

- CONCEITOS BASICOS

DADOS: $y_1, y_2, ..., y_n$ ~Y com fdp f e fda F

θ: característica populacional

T: estatística; t: valor de T na amostra

- INTERESSE: obter a distribuição de probabilidade de T; viés de T, dp(T); quatis, intervalo de confiança para θ, testes.

- SITUAÇÕES: PARAMETRICA E NÃO-PARAMETRICA

- FUNÇÃO DE DISTRIBUIÇÃO EMPIRICA(FDE)

$\hat{F}$: estimativa de F, a partir da distribuição empírica, que coloca probabilidade 1/n em cada y_j.

$$\hat{F}(y) = \frac{\#\{y_j \leq y\}}{n}$$

- FUNÇÃO ESTATÍSTICA

Estatística de interesse: t=f(y(1), ..., y(n))

$t = t(\hat{F})$: função estatística

$\theta = t(F)$

$\hat{F} \to F \quad \Rightarrow \quad T = t(\hat{F}) \to \theta = t(F)$ em probabilidade(consistência)

- PRECISÃO DA MEDIA AMOSTRAL

Amostra: x_1, ..., x_n: $\bar{x} = \frac{\sum x_i}{n}$

Erro padrão de $\bar{x} = \frac{s}{\sqrt{n}}, \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

- ERRO PADRÃO DE T: estimador de θ

$$ep(T) = \sqrt{\text{var}(T)}$$

Em geral, var(T) depende de θ, portanto

$$\hat{ep}(T) = \sqrt{\hat{\text{var}}(T)}$$

Para a maioria dos estimadores, não há formulas para calcular o ep.

BOOTSTRAP

$\underline{x} = (x_1, \cdots, x_n)$: dados independentes → $s(\underline{x})$: estatística de interesse

Amostra bootstrap: $\underline{x}^* = (x_1^*, \cdots, x_n^*)$, amostramos, com reposição, n vezes de $\underline{x}$

- ALGORITMO BOOTSTRAP:

gera um grande número de amostras bootstrap independentes: $\underline{x}^{*1}, \underline{x}^{*2}, \cdots, \underline{x}^{*B}$

Cada uma de tamanho n. B ≈ 200.

- OBJETIVO: estimar ep dos estimadores

- RÉPLICA BOOTSTRAP:

Amostra bootstrap $\underline{x}^* \to s(\underline{x}^*)$: réplica bootstrap

- ESTIMADOR BOOTSTRAP DO ERRO PADRÃO: desvio padrão das réplicas bootstrap

$$\hat{ep}_{boot} = \left[\sum_{b=1}^{B}\left[s(\underline{x}^{*b}) - s(\bullet)\right]^2 /(B-1)\right]^{\frac{1}{2}}$$

$$\text{Com } s(\bullet) = \frac{\sum_{b=1}^{B} s(\underline{x}^{\bullet b})}{B}$$

- ESTIMADOR BOOTSTRAP DE $\hat{ep}_F(\hat{\theta})$: usa $\hat{F}$ no lugar de F, isto é, o estimador bootstrap é

$ep_{\hat{F}}(\hat{\theta})$: estimador bootstrap **ideal** do ep de $\hat{\theta}$

Não há formula que permite calcular o estimador bootstrap ideal exatamente.

- ALGORITMO BOOTSTRAP: forma computacional de obter uma boa aproximação do valor numérico de $ep_{\hat{F}}(\hat{\theta})$

Para implementar num computador:

(1) um mecanismo aleatório seleciona inteiros i_1, i_2, ..., i_n, entre 1 e n, com probabilidade 1/n;

(2) a amostra bootstrap consiste nos números

$$x_1^* = x_{i_1}, \cdots, x_{i_n}$$

ALGORITMO BOOTSTRAP PARA ESTIMAR ERROS PADRÕES

[1] sele cione B amostras independentes, $\underline{x}^{*1},\cdots,\underline{x}^{*B}$, cada uma consistindo de n valores selecionados com reposição de $\underline{x}$. Tome B ≈ 25 - 200.

[2] calcule a réplica bootstrap para cada amostra bootstrap:

$$\hat{\theta}^*(b) = s(\underline{x}^{*b}), \quad b = 1,\cdots B.$$

[3] estime o erro padrão $ep_F(\hat{\theta})$ pelo desvio padrão amostral das B réplicas:

$e\hat{p}_B = \left[\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*(b) - \hat{\theta}^*(\bullet))^2\right]^{\frac{1}{2}}$, estimador bootstrap não paramétrico, onde

$$\hat{\theta}^*(\bullet) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*(b)$$

$\lim_{B\to\infty} e\hat{p}_B = ep_{\hat{F}} = ep_{\hat{F}}(\hat{\theta}^*)$: desvio padrão empírico se aproxima do desvio padrão populacional quando B →∞.

Neste caso, a "populacional" é a população dos valores $\hat{\theta}^* = s(x^*), \quad onde \quad \hat{F} \to (x_1^*,\cdots,x_n^*) = x^*$

BOOTSTRAP PARAMÉTRICA

Útil em problemas para os quais dispomos de alguns conhecimentos sobre a **forma** da população e para comparar com análises não paramétricas.

X → F, F: FDA

$(x_1, x_2, \ldots, x_n) \sim F$

Considere um modelo paramétrico para os dados.

$\hat{F}_{par}$: estimador de F obtido deste modelo.

$\hat{\theta}$: estimador do parâmetro θ.

A estimativa bootstrap paramétrica do $ep(\hat{\theta})$ é definida por

$$ep_{\hat{F}_{par}}(\hat{\theta}^*)$$

Exemplo: escola de direito

Suponha F ~ $N_2(\underline{\mu},\underline{\nu})$, n = 15
onde

$\underline{\mu} = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}$, $\underline{\nu} = \begin{bmatrix} \sigma_y^2 & \sigma_{yz}^2 \\ \sigma_{zy}^2 & \sigma_z^2 \end{bmatrix}$ simétrica

Estimamos $\underline{\mu}$ por $\hat{\underline{\mu}} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix}$ e V por

$\hat{V} = \frac{1}{14}\begin{bmatrix} \sum (y_i - \bar{y})^2 & \sum (y_i - \bar{y})(z_i - \bar{z}) \\ \bullet & \sum (z_i - \bar{z})^2 \end{bmatrix}$

$\hat{F}_{norm} \to N_2(\hat{\underline{\mu}},\hat{V})$: estimador paramétrico de F

$ep_{\hat{F}_{norm}}(\hat{\theta}^*)$: estimador bootstrap paramétrico de $ep(\hat{\theta})$, onde $\hat{\theta} = cor\hat{r}(y,Z)$.

Retiramos B amostras de tamanho n com reposição da população

$$\hat{F}_{par} = \hat{F}_{norm} \quad \to \quad (\underline{x}^{*1},\cdots,\underline{x}^{*B})$$

Procede-se, depois, como em (2) e (3) do AB.

B amostra de tamanho 15 de $\hat{F}_{norm}$ e calculamos o coeficiente de correlação para cada amostra.

$\hat{ep}_B = 0{,}124$ (0,131: estimador não paramétrico)

ESTIMADOR BOOTSTRAP DO VIÉS

$(x_1, x_2, ..., x_n) \sim F$; $\theta = t(F)$. $\tilde{\theta} = s(x)$

CONSIDERE O ESTIMADOR $\hat{\theta} = t(\hat{F})$.

O viés de $\tilde{\theta} = s(x)$ é definido por

$$viés_F = E_F(s(x)) - t(F)$$

Ou seja, esperança do estimador – θ

Estimadores não viciados são importantes na teoria e na pratica estatística.

Podemos usar bootstrap para avaliar o viés de um estimador.

O estimador bootstrap do viés é definido por

$viés_{\hat{F}} = E_{\hat{F}}\{s(x^*)\} - t(\hat{F})$: estimador ideal do viés.

Exemplo:

a) t(F) = μ, $s(x) = \bar{x}$, $viés_{\hat{F}} = 0$

b) $s(x) = \frac{\sum (x_i - \bar{x})^2}{n}$; $viés[s(x)] = -\frac{1}{n}\sigma^2$; *neste caso,* $viés_{\hat{F}} = -\frac{1}{n^2}\sum (x_i - \bar{x})^2$

Para a maioria dos estimadores utilizados na prática, o estimador bootstrap do viés deve ser aproximado por simulação:

[1] geramos amostras bootstrap $(\underline{x}^{*1}, \cdots, \underline{x}^{*B})$ e calculamos as replicas bootstrap

$$\tilde{\theta}^*(b) = s(x^{*b}), \quad b = 1, \cdots, B.$$

[2] aproximamos as esperanças bootstrap $E_{\hat{F}}\{s^*(x)\}$ pela média

$$\tilde{\theta}^*(\bullet) = \frac{\sum_{b=1}^{B} \tilde{\theta}^*(b)}{B} = \frac{\sum_{b=1}^{B} s(x^{*b})}{B}$$

[3] o estimador bootstrap do viés é

$$\hat{vies}_B = \tilde{\theta}^*(\bullet) - t(\hat{F})$$

Exemplo: dados hormônio (bio-equivalência)

As concentrações:

	placebo	oldpatch	newpatch	Old-placebo	New-old
subject				z	y
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719
Mean				6342	-4520,3

FDA: $\frac{|E(novo) - E(antigo)|}{E(antigo) - E(placebo)} \leq 0{,}20$ critério

Parâmetro: $\theta = \frac{E(novo) - E(antigo)}{E(antigo) - E(placebo)}$

Objetivo: calcular o viés e o erro padrão de $\hat{\theta}$

Considere:

z_i = medidas com o "antigo" – medidas com o "placebo"
y_i = medidas com o "novo" – medidas com o "antigo"

$x_i=(z_i , y_i)$, i=1, 2, ..., 8
$X=(x_1, x_i ,..., x_8)$, ~ F: desconhecida

$$\theta = t(F) = \frac{E_F(y)}{E_F(z)}$$

$$\hat{\theta} = t(\hat{F}) = \frac{\bar{y}}{\bar{Z}} = \frac{\sum_{i=1}^{8} y_i / 8}{\sum_{i=1}^{8} z_i / 8} = -0{,}0713$$

Nota: Z e Y são dependentes.

$\left|\hat{\theta}\right| << 0{,}20$, portanto aparentemente a condição do FDA está satisfeita e os dois hormônios são bioequivalêntes.

B=400 amostras bootstrap: $x^{*i} = (x_1^{*i}, \cdots, x_8^{*i})$

→ 400 réplicas bootstrap

$$\hat{\theta}^* = \frac{\bar{y}^*}{\bar{z}^*}$$

As 400 réplicas tem um desvio padrão amostral, $e\hat{p}_{400} = 0{,}105$
Média amostral: $\hat{\theta}^*(\bullet) = -0{,}0670.$

Estimador bootstrap do viés:

$$vie\hat{s}_{400} = -0{,}0670 - (-0{,}0713) = 0{,}0043$$

$\Rightarrow \frac{vie\hat{s}_{400}}{e\hat{p}_{400}} = \frac{0{,}0043}{0{,}105} = 0{,}041$, portanto viés sob controle.

Regra: $vie\hat{s} < 0{,}25 * e\hat{p}$ $\Rightarrow$ podemos ignorar o viés

$$RMSE = \sqrt{E_F(\hat{\theta}-\theta)^2} = \sqrt{ep_F^2(\hat{\theta}) + vies_F^2(\hat{\theta},\theta)} = ep_F(\hat{\theta})\sqrt{1+\left(\frac{vies_F}{ep_F}\right)^2} \approx ep_F(\hat{\theta})\left[1+\frac{1}{2}\left(\frac{vies_F}{ep_F}\right)^2\right]$$

CORREÇÃO DE VIÉS

$\hat{V}$: estimador do $vies_F(\hat{\theta},\theta) \Rightarrow \overline{\theta} = \hat{\theta} - \hat{V}$: estimador corrigido para o viés.

Tomando

$$\hat{V} = vies_B = \hat{\theta}^*(\bullet) - \hat{\theta}, obtemos\, \overline{\theta} = 2\hat{\theta} - \hat{\theta}^*(\bullet)$$

Exemplo(hormônio):

$\hat{V}_{400} = 0{,}0043 \quad e \quad \hat{\theta} = -0{,}0713 \quad \overline{\theta} = -0{,}0713 - 0{,}0043 = -0{,}0756$

Observações:

1) a correção do viés pode ser perigosa na prática. Mesmo que $\overline{\theta}$ seja menos viesado do que $\hat{\theta}$, ele pode ter erro padrão substancialmente maior.
2) O viés é mais difícil de estimar do que o ep, → B maior para estimar o viés.
3) Se $\hat{V} << ep,$ melhor usar $\hat{\theta}$ do que $\overline{\theta}$.

INTERVALO DE CONFIANÇA

Dado o estimador $\hat{\theta}$ de θ, seu ep estimado, $e\hat{p}(\hat{\theta})$, o intervalo de confiança(IC) usual, com coeficiente de confiança(C.C.) 90%, para θ é

$$\hat{\theta} \pm 1.645\, e\hat{p}(\hat{\theta})$$

$$\underline{x} = (x_1, \cdots, x_n) \sim F \quad e \quad \hat{\theta} = t(\hat{F})$$

$e\hat{p}(\hat{\theta})$: algum estimador do $ep(\hat{\theta})$, baseado por ex, em réplicas jackknife ou bootstrap.

Então, sob determinadas condições,

$$\hat{\theta} \xrightarrow{D} N(\theta, e\hat{p}(\hat{\theta})), \quad n \to \infty$$

Ou

$$\frac{\hat{\theta} - \theta}{e\hat{p}(\hat{\theta})} \xrightarrow{\cdot} N(0,1) \qquad (8)$$

Assim, $\left[\hat{\theta} - Z^{(1-\alpha)}\ e\hat{p},\ \ \hat{\theta} + Z^{(1-\alpha)}\ e\hat{p}\right]$ é o IC padrão com C.C. igual a 1-2α.

APROXIMAÇÃO PARA AMOSTRAS FINITAS:

Para $\hat{\theta} = \bar{x}$, temos seguinte resultado:

$$Z \ = \ \frac{\hat{\theta} - \theta}{e\hat{p}} \ \sim \ t_{n-1} \qquad (9)$$

E o IC fica

$$\left[\hat{\theta} - t_{n-1}^{(1-\alpha)}\ e\hat{p},\ \ \hat{\theta} + t_{n-1}^{(1-\alpha)}\ e\hat{p}\right]$$

Se $\hat{\theta} = \bar{x}$ e x ~normal, a aproximação é exata e o IC é mais largo, refletindo o fato que o ep não é conhecido. Se n ≥ 100, $t_{n-1}^{(1-\alpha)} \ \cong \ Z^{(1-\alpha)}$.

INTERVALO BOOTSTRAP-t

Com o uso de bootstrap podemos obter IC acurado aem utilizar a expressão (8).

A distribuição de Z em (9) será estimada diretamente dos dados, ou seja, obtemos uma tabela apropriada para o particular conjunto de dados.

PROCEDIMENTO:

[1] geramos B amostras bootstrap $\underline{x}^{*1},\cdots,\underline{x}^{*B}$

[2] para cada amostra construímos

$$Z^*(b)=\frac{\hat{\theta}^*(b)-\hat{\theta}}{e\hat{p}^*(b)}$$

Com $\hat{\theta}^*(b)=s(\underline{x}^{*b})$ valor de $\hat{\theta}$ para a amostra $\underline{x}^{*b}$

$e\hat{p}^*(b)$: erro padrão estimado de $\hat{\theta}^*(b)$ para a amostra $\underline{x}^{*b}$

[3] o α-percentil de $Z^*(b)$ é estimado pelo $\hat{t}^{(\alpha)}$ tal que

$$\#\{Z^*(b)\le\hat{t}^{(\alpha)}\}/B=\alpha$$

[4] O IC bootstrap-t é dado por

$$\left[\hat{\theta}-\hat{t}^{(\alpha)}\ e\hat{p},\quad \hat{\theta}+\hat{t}^{(1-\alpha)}\ e\hat{p}\right]$$

Ex: se B=1000, a estimativa do 5%-percentil ($\hat{t}^{(5\%)}$) é o 50º. maior valor dos $Z^*(b)$.

Intervalo percentil

$\underline{x}^*$: dados bootstrap

$\hat{\theta}^*=s(\underline{x}^*)$: réplicas bootstrap

$\hat{G}^*$: FDA de $\hat{\theta}^*$

O intervalo percentil, com C.C. 1-2α, é definido pelos percentis α e 1-α de $\hat{G}^*$:

$$\left[\hat{\theta}_{\text{inf}},\quad \hat{\theta}_{\text{sup}}\right]=\left[\hat{G}^{-1}(\alpha),\quad \hat{G}^{-1}(1-\alpha)\right]$$

OBSERVAÇÕES:

1) Bα não inteiro, $\alpha \leq 0{,}05$. Considere k=[(B+1) α]. Os quantis α e 1-α são dados pela k-ésima maior. E (B+1-k) maior observação, respectivamente.

Ex.: B=50, α=0,05, Bα=2,5, k=[51*0,05]=[2,55]=2, portanto α-percentil é a 2ª. Observação e o (1-α)-percentil é a 49ª. Observação.

2) em amostras grandes, a cobertura do IC bootstrap-t tende a ser mais próxima do CC desejado do que o IC padrão e t.

Ex. Ratos:

16 ratos(7: tratamento; 9: controle)

Dados: tempo de sobrevivência (em dias) após o tratamento

Questão: tratamento prolonga sobrevida após a cirurgia?

Tabela 1: dados

Group	Data	Sample Size	mean	Estimated standar error
Treatment	94, 197, 16, 38, 99, 141 23	7	86,86	25,24
Control	52, 104, 146, 10, 51, 30, 40, 27, 46	9	56,22	14,14
		difference	30,63	28,93

Tabela 2: bootstrap estimates of standard error for the mean and median: treatment group. The median is less accurate (has larger standard error) than the mean for this data set.

B	50	100	250	500	1000	∞
mean	19,72	23,63	22,32	23,76	23,02	23,36
median	32,21	36,35	34,46	36,72	36,48	37,83

$$\bar{x} - \bar{y} = 30{,}63 \quad e \quad \frac{\bar{x} - \bar{y}}{dp(\bar{x} - \bar{y})} = \frac{30{,}63}{28{,}93} = 1{,}05 \quad (no)$$

$$m_1 = med(x) = 94, \quad m_2 = med(y) = 46 \quad \rightarrow \quad m_1 - m_2 = 48$$

$$B = 100 \quad \rightarrow \quad e\hat{p}(m_1) = 11{,}54, \quad e\hat{p}(m_2) = 36{,}36 \quad \rightarrow \quad e\hat{p}_{boot} = \sqrt{(36{,}35)^2 + (11{,}54)^2} = 38{,}14$$

Estatística para teste: $\frac{48}{38{,}14} = 1{,}26$

IC:
Media dos ratos tratados: $\hat{\theta} = 86{,}86 \quad e \quad e\hat{p} = 25{,}24$

IC padrão(γ=0,90):

[86,86-1,65*25,24; 86,86+1,65*25,24] =

[45; 128,4]

B=1000 réplicas: $\hat{\theta}^*$=?

Tabela 3: percentiles of $\hat{\theta}^*$ based on 1000 bootstrap replications, where $\hat{\theta}$ equais the mean of the treated mice.

2,5%	5%	10%	16%	50%	84%	90%	95%	97,5%
45,9	49,7	56,4	62,7	86,9	112,3	118,7	126,7	135,4

Percentile 5% = 49,7
Percentile 95% = 126,7

Intervalo percentil com C.C. 90% = [49,7; 126,7]

Utilizar os percentis do histograma para definir limites de confiança.

PROCEDIMENTO:

[1] geramos B amostras bootstrap $\underline{x}^{*1},\cdots,\underline{x}^{*B} \rightarrow \hat{\theta}^*(b) = s(\underline{x}^{*b})$

[2] $\hat{\theta}_B^{*(\alpha)}$: α-percentil dos valores $\hat{\theta}^*(b) = s(\underline{x}^{*b})$

[3] IC percentil aproximado com 1-2α:

$$\left[\hat{\theta}_{\%,\inf};\ \ \hat{\theta}_{\%,\sup}\right] = \left[\hat{\theta}_B^{*(\alpha)};\ \ \hat{\theta}_B^{*(1-\alpha)}\right]$$

Exemplo: $x_1, x_2, \ldots, x_{10} \sim N(0,1)$

$\theta = e^{\mu}$, μ = *média populacional* $\Rightarrow$ $\theta = e^0 = 1$

$\hat{\theta} = e^{\bar{x}} = 1{,}25$ (exemplo artificial)

IC padrão: $1{,}25 \pm 1{,}96 * e\hat{p}_{1000} = 1{,}25 \pm 1{,}96 * 0{,}34 = [0{,}59;\ \ 1{,}92]$

B=1000 réplica $\rightarrow$ $\hat{\theta}^* = e^{\bar{x}^*}$

Percentis empíricos de $\hat{\theta}^*$ $\rightarrow$ IC percentil 95%

[0,75; 2,07]

$\rightarrow$ aprox. Normal não é muito boa nesse caso.

Tabela 4: percentiles of $\hat{\theta}^* = e^{\bar{x}^*}$ for a normal sample of size 10.

2,5%	5%	10%	16%	50%	84%	90%	95%	97,5%
0,75	0,82	0,90	0,98	1,25	1,61	1,75	1,93	2,07