19 лекция !!

Бутстрэп.

→ кусочек док.-в
→ примеры, где бутстрэп не работает.

Сценарий:

θ – неизв. пар-р $\hat{\theta}$ – оценка по выборке
$\hat{\theta}(X_1, \dots, X_n)$ X_i незав. один. расп.

$\mu = E(X_i)$
$\hat{\mu} = \bar{X} = \frac{X_1 + \dots + X_n}{n}$
Хрестом. пример.

Цель:

$[\hat{\theta}_L\,?$ ... $\hat{\theta}_R\,?]$

$\hat{\theta}_L(X_1 \dots X_n)$
$\hat{\theta}_R(X_1 \dots X_n)$

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R) = 1 - \alpha = 95\%$$

! θ - неизв. конст.
! $\hat{\theta}_L$ и $\hat{\theta}_R$ – С.В.

Шаг! упростим цель:
Вместо двустороннего ДИ хотим односторонний ДИ.

$\hat{\theta}_L$

$$P(\theta \geq \hat{\theta}_L) = 0.95$$

Шаг 2 присмотримся к классическому интервалу для μ.

$$\bar{X} - \mu \overset{\text{ЦПТ}}{\underset{\text{approx}}{\sim}} N\left(0; \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X} - \mu) \overset{\text{ЦПТ}}{\underset{\text{approx}}{\sim}} N(0; \sigma^2)$$

Блок предпосылок. А

$R = h(\hat{\theta}) - h(\theta)$ ← ошибка оценивания

- h – строго монот-ая ф-ция возраст.
- R распр-на симметр. отн-но 0.
  [$f_R(-t) = f_R(t)$ } $F_R(t) = 1 - F_R(-t)$]
- ф. ция распр $F_R(t)$ непрерывна.
- $F_R(t)$ строго возрастает.

[блок А выполнен примерно]

Шаг 3 супер идеальная ситуация

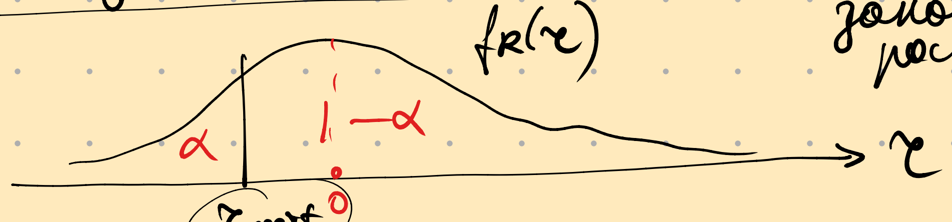
преобразование h известно

F – известная ф-ция распр. R

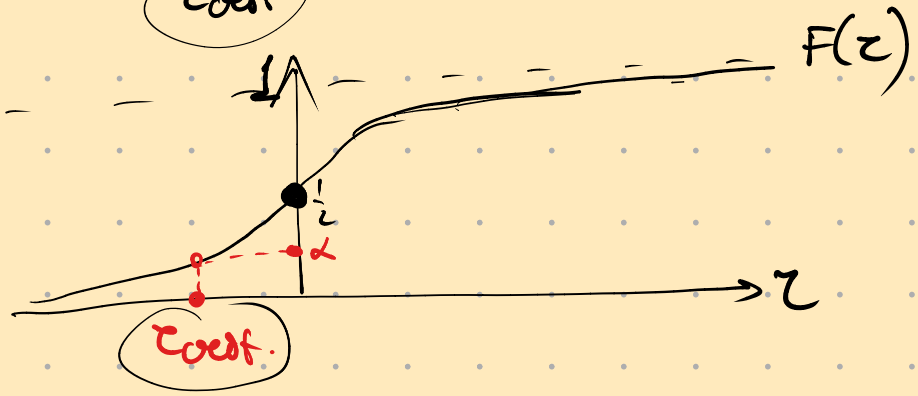
в примере:
$h(u) = u$
F_R – ф. распр. для $N(0; \frac{\sigma^2}{n})$
хрест. пример с μ и $\hat{\mu} = \bar{X}$

решим идеал. ситуацию в общем виде.

3.1



закон распр. R изв.



можно найти τ_{crit} т.к. F_R известно.

3.2 по симметрии

$$P(R \le \tau_{crit}) = \alpha \qquad P(R \ge -\tau_{crit}) = 1-\alpha$$

$$P(R \le -\tau_{crit}) = 1-\alpha$$

3.3 выворачиваем θ

$$P\left(h(\hat{\theta}) - h(\theta) \le -\tau_{crit}\right) = 1-\alpha$$

$$P\left(h(\theta) \ge h(\hat{\theta}) + \tau_{crit}\right) = 1-\alpha$$

h заявленная мон. возраст.

$$P\left(\theta \ge h^{-1}\left(h(\hat{\theta}) + \tau_{crit}\right)\right) = 1-\alpha$$

$$\hat{\theta}_L^{ideal} = h^{-1}\left(h(\hat{\theta}) + \tau_{crit}\right)$$

абсолютно точная граница CI /односторонняя/ кот-ую мы можем посчитать.

$$P\left(\theta \ge \hat{\theta}_L^{ideal}\right) = 1-\alpha$$

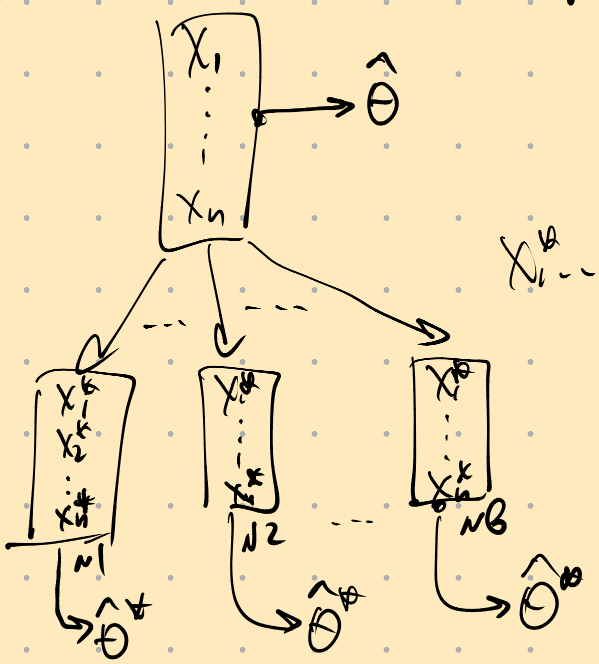
Шаг 4.

два источника случайности

① Природа генерирует исх-ую выб-ку: $X_1, \dots, X_n$

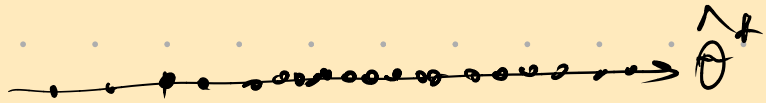
$P_*(\ )$ ② процесс построения бутстрэп выборки $X_1^*, \dots, X_n^*$

Вспомним наивный / перцентильный бутстрэп.



бутстрэп выборка $X_1^* \dots X_n^*$ генерируется из исходной $X_1, X_2 \dots X_n$ путем выбора элементов с возм-стью повт-ния (равн-но, того же размера)

n – число наблю-ий
B – число бут-выборок



B велико [B = 10000 на практике]

$$P_*(\hat\theta^* \le t) = K(t)$$ ← не знаем, но дешёво и точно умеем оценивать

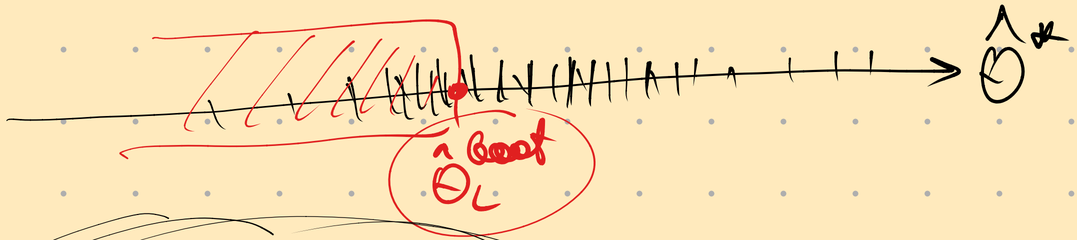
$$K(t) \approx \frac{\text{кол-во}\{\hat\theta^* \le t\}}{B}$$

по ЗБЧ выбрав $B \to \infty$ можем сделать сколь угодно точным.

Бутстрэп дал нам $\hat K(t)$, но можно считать, что знаем $K(t)$

Шаг 5

$\alpha = 5\%$ $\quad 1-\alpha = 95\%$ [на практике, конечно, двусторонний]



$$\hat{\theta}_L^{boot} = K^{-1}(\alpha), \quad B \to \infty$$

[это не сущ. ошибка]

[ненормально] Шаг 6.

F:
- непр.
- строго возраст
- $F(-t) = 1 - F(t)$

предп.

$$P\left(h(\hat{\theta}) - h(\theta) \le t\right) = F(t)$$

$$P_*\left(h(\hat{\theta}^*) - h(\hat{\theta}) \le t\right) = F(t)$$

Шаг 7

h неуб-ая строго монот.

h – неуб, F неуб

вычтем $h(\hat{\theta})$

$$P_*\left(\hat{\theta}^* \le \hat{\theta}_L^{boot}\right) = \alpha =$$

$$= P_*\left(h(\hat{\theta}^*) \le h(\hat{\theta}_L^{boot})\right) =$$

$$= P_*\left(\underbrace{h(\hat{\theta}^*) - h(\hat{\theta})}_{R^*} \le \underbrace{h(\hat{\theta}_L^{boot}) - h(\hat{\theta})}_{t}\right) = F_R(t)$$

$$h(\hat{\theta}_L^{boot}) - h(\hat{\theta}) = F^{-1}(\alpha)$$

$$h(\hat{\theta}_L^{boot}) = h(\hat{\theta}) + \underbrace{F^{-1}(\alpha)}_{\tau_{crit}}$$

$$\hat{\theta}_L^{boot} = h^{-1}\left(h(\hat{\theta}) + \tau_{crit}\right)$$

Шаг 8

<u>мораль истории</u>: при выполнении блока предпосылок A: $\hat{\theta}_L^{boot}$ совпадает $\hat{\theta}_L^{ideal}$, но! не требует ни знания h, ни знания F.

неф.

в <u>реальности</u>: предп A выполнена примерно, $\hat{\theta}_L^{boot} \approx \hat{\theta}_L^{ideal}$.

Упр <u>канонический пример провала бутстрэпа</u>

$X_1, X_2, \ldots, X_n$ ~ незав Unif$[0, a]$
a – неизв. пар-р.

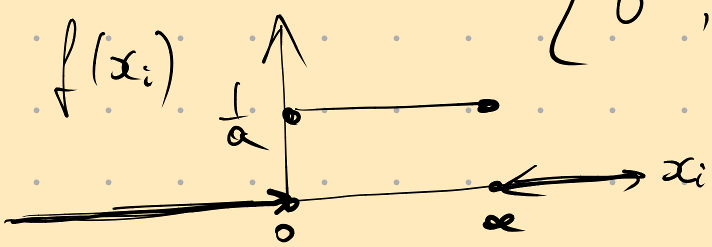
а) $\hat{a}_{ML}$? – оценка метода макс. правд.

б) наивный (перцентильный) бутстрэп для постр. 95% CI для a.
кол-но

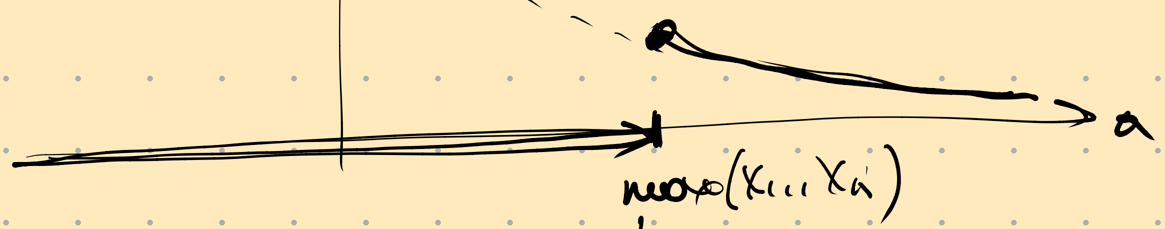
какая будет фактич-ая вер-сть накрытия.

в) убедитесь, что блок предп A сильно нар-н.
[рост n не приводит к вып-ию блока A]

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} \frac{1}{a} \cdot \frac{1}{a} \cdot \frac{1}{a} \cdot \ldots \cdot \frac{1}{a} & \text{все } x_i \in [0; a] \\ 0, & \text{хотя бы } 1 \; x_i \notin [0, a] \end{cases}$$

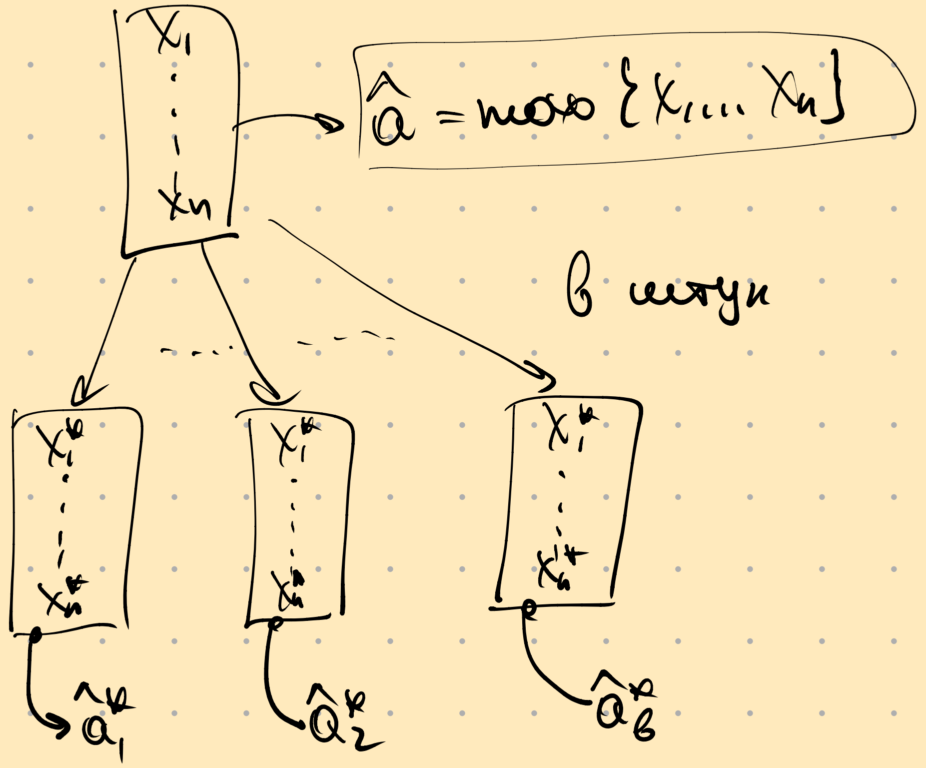


$$L(a) = f(x_1, \ldots, x_n) = \left(\frac{1}{a}\right)^n \cdot I(\max(x_1 \ldots x_n) \leq a)$$



$$\hat{a} = \max\{X_1 \dots X_n\}$$

б) наивный [перцентильный] бутстрэп ci.



$$\hat{a}^* = \max\{X_1^*, X_2^* \dots, X_n^*\}$$

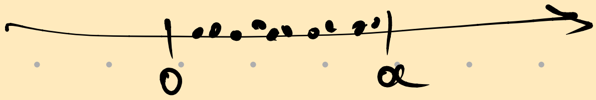
$$\hat{a}_j^* = \max\{X_{1j}^*, X_{2j}^*, X_{3j}^* \dots X_{nj}^*\}$$

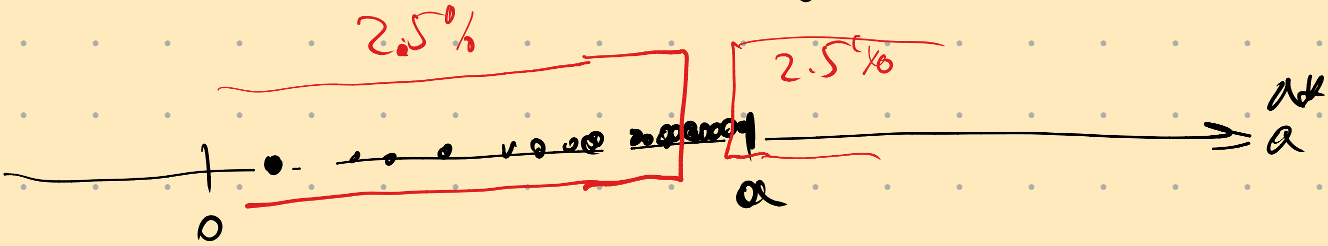
для $\forall$ бутстрэп выборки $\underline{\max\{X_1^* \dots X_n^*\} \le}$

$$\underline{\le \max\{X_1 \dots X_n\}}$$

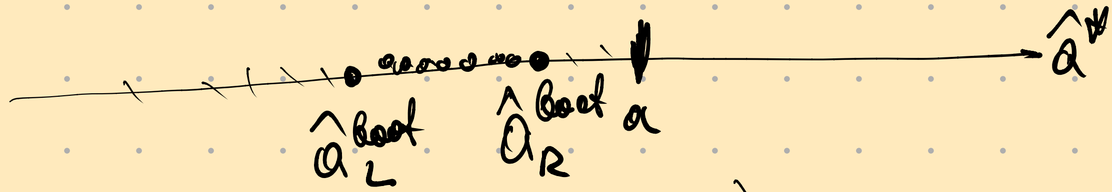
$$\hat{a}^* = \max\{X_1^* \dots X_n^*\} \le \max\{X_1, \dots, X_n\} \le a.$$

$$X_i \sim Unif[0;a]$$





95% двустор.



$$P(\hat a_L^{boot} \le a \le \hat a_R^{boot}) = 0$$

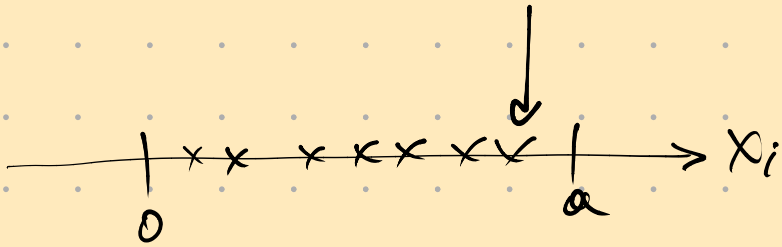
→ номин. вер-сть покрытия $= 95\%$

→ фактич. вер-сть покрытия $= 0\%$.

В) А нарушен ли блок предпосылок?

$\hat a - a \le 0$

$\hat a = \max\{X_1, \dots, X_n\}$



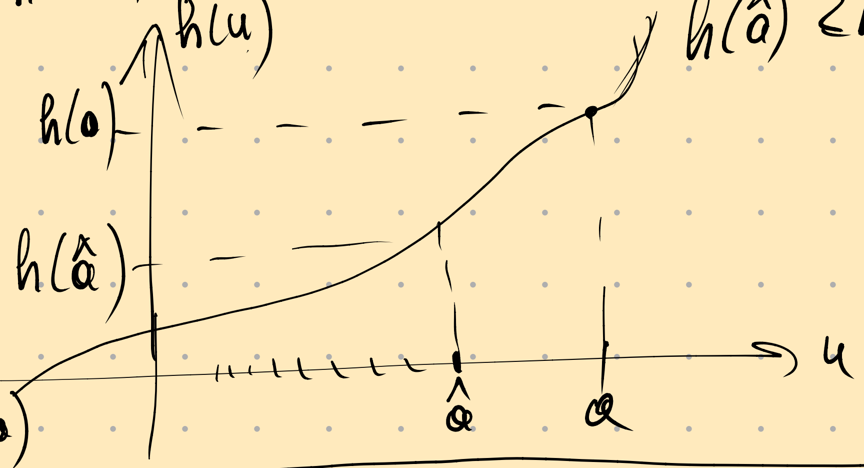
$\hat a - a$ не симм-но около 0.

h - строго монотонно возраст. ф-ция

$h(\hat a) - h(a) \le 0.$

$\hat a \le a$

$h(\hat a) \le h(a)$



$R = h(\hat a) - h(a)$

ошибка оц-ния не симм-ная отн-но 0.

[с ростом n симм-сть не появляется].