Лекция 18.

Бутстрэп / бутстреп / бутстрап
Bootstrap.

- техника для построения дов-х интервалов, [для неизв. пар-в]
  оценки дисперсии [оценок неизв-х пар-в].
- вычислительная (компьютер).
  - → техника реализ.
  - → примеры, когда бутстрэп не применим

$\boxed{\text{Пример 1}}$

идея бутстрэпа.

Задача. Есть нез. один-во расп-ые $X_1, X_2 \ldots X_n$

$E(X_i) = \mu$, $Var(X_i) = \sigma^2$ μ, σ^2 не известны.

Хотим построить 95% и асимпт.-ий интервал для μ.

$$\frac{\hat{a} - a}{se(\hat{a})} \xrightarrow{dist} N(0;1) \quad ?$$

Способ классический:

$\bar{X} \overset{asy}{\sim} Norm$

$$\frac{\bar{X} - \mu}{se(\bar{X})} \xrightarrow{dist} N(0;1)$$

$$se(\bar{X}) = \sqrt{\hat{\sigma}^2/n}$$

$$\hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

asy CI для μ: $[\bar{X} - 1{,}96 \cdot se(\bar{X}); \bar{X} + 1{,}96 \cdot se(\bar{X})]$

$$\lim_{n\to\infty} P(\mu \in CI) = 0{,}95 \quad ✓$$

• минус 1: нужна формула для $se(\bar{X})$
есть случаи когда $se(\hat{a})$ не имеет явного вида.

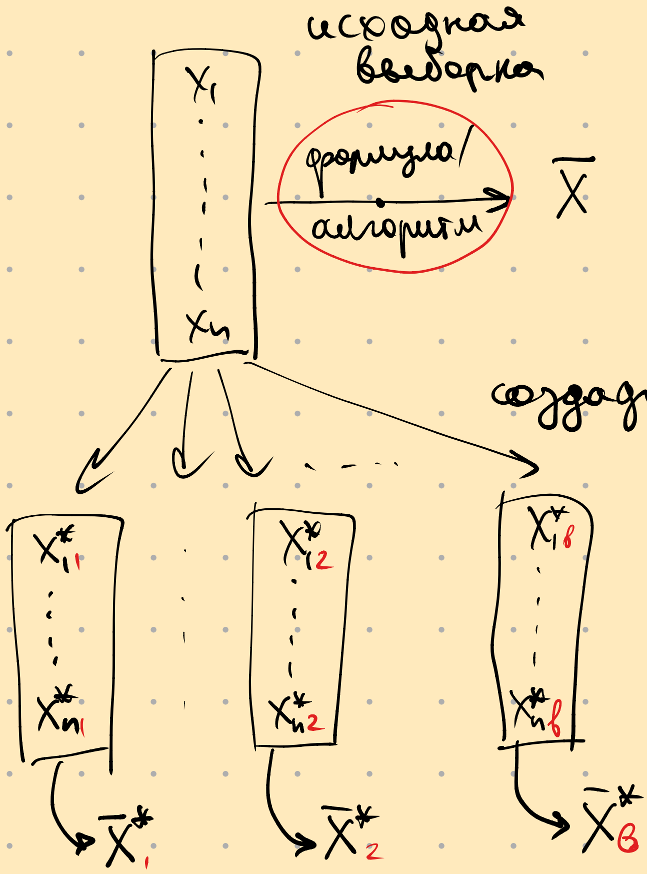
оц-ка пар-ра $\hat{a}$ имеет сложный вид
(м.б. даже нет явного вида)

• минус 2: а насколько быстро сх-ся ЦПТ?

• <u>наивный бутстрэп.</u>

бутстрэп – семейство алгоритмов.

идея: „вытащить" оценку закона распределения оценки $\hat{a}$ из самой выборки.



(оценка для неизв-го μ)

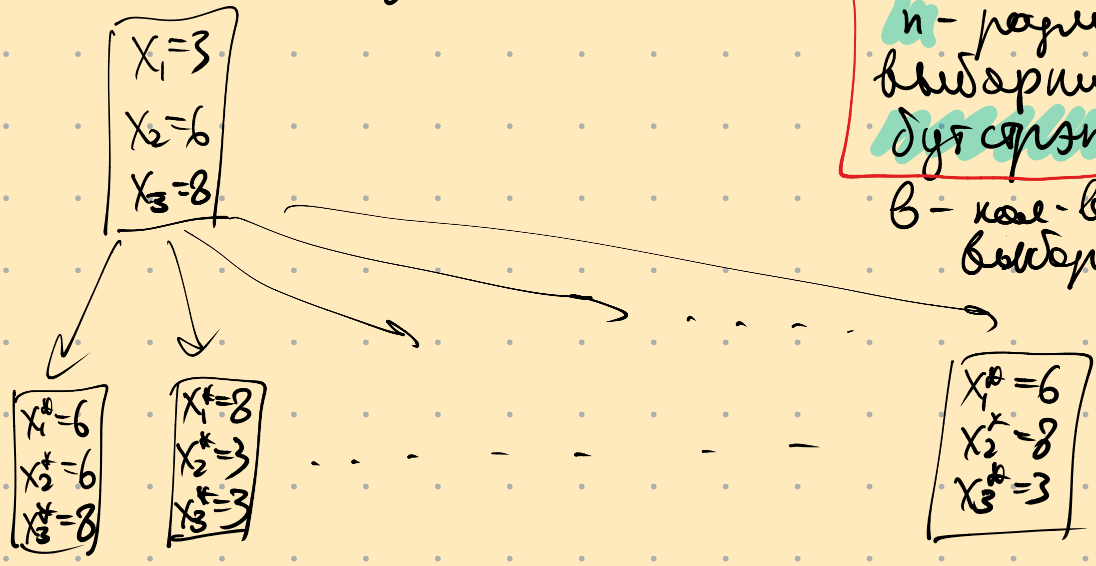
создадим искусственные бутстрэп выборки.

$B = 10000$ [кол-во бутстрэп-выборок]

[опустим индексы для номера бутстрэп выборки]

бутстрэп выборка строится на базе исходной выборки путём <u>случайной выборки</u> с повторениями и сохранением размера. (равновероятно)

игрушечный пример. ($n=3$)

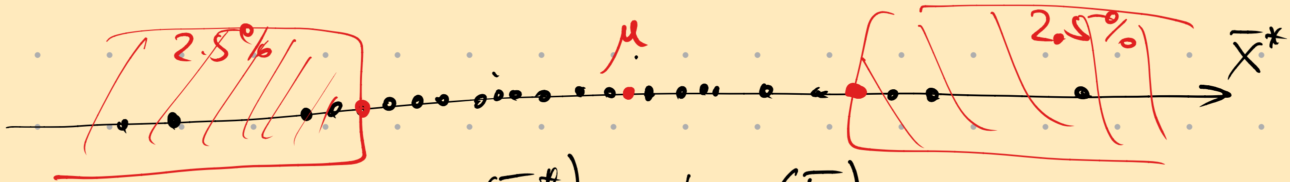


n – размер исх-ой выборки и каждой бутстрэп выборки

B – кол-во бутстрэп выборок

n = 100
B = 10000

наивный бутстрэп (naive Bootstrap)



идея 1: $Law(\bar{X}^*) \approx Law(\bar{X})$

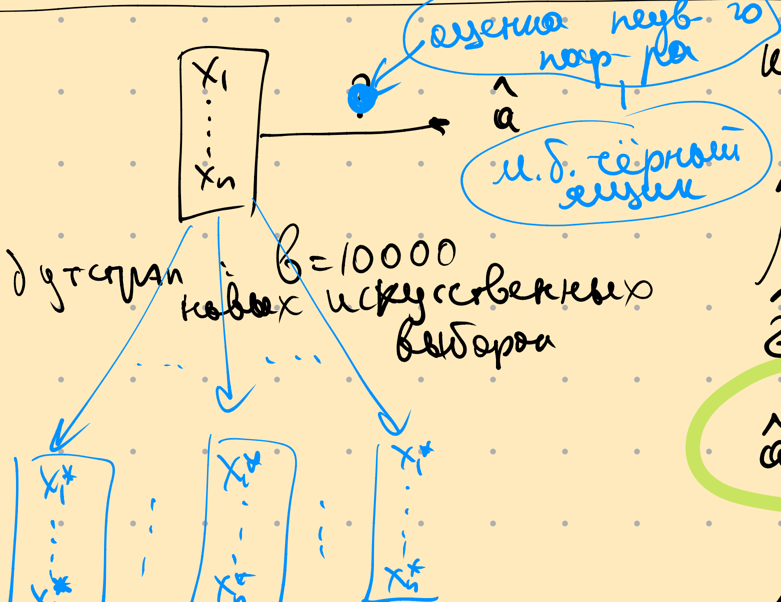
бутстрэп доверит-ный интервал для μ:

$$[q_L(\bar{X}^*);\ q_R(\bar{X}^*)]$$

Упр.: есть набл $X_1, \ldots, X_n$ незав, один. распр с неизв. ср. $E(X_i)=\mu$ и $Var(X_i)=\sigma^2$

неизв. пар-р $a = \frac{\sigma}{\mu}$

Как выглядит наивный бутстрэп для построения CI?



как построить единств-ую оценку?

$$\hat{\mu} = \bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

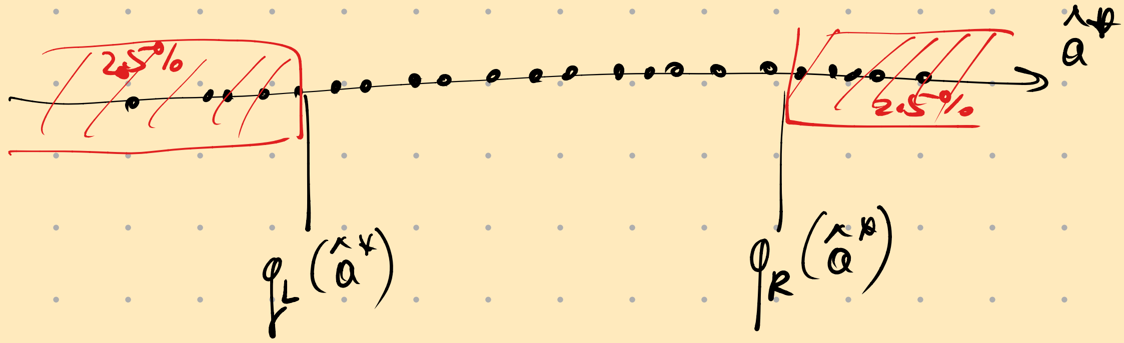
$$\hat{\sigma}^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$\hat{a} = \frac{\sqrt{\sum (X_i - \bar{X})^2} / \sqrt{n-1}}{\sum X_i / n}$$

есть одна оценка для неизв-го a.

$\hat{}\ \sqrt{\sum (X_i - \bar{X})^2}$

$\hat a_1^* \quad \ldots \quad \hat a_2^* \quad \ldots \quad \hat a_B^*$



boot - CI $\quad [q_L(\hat a^*) ; q_R(\hat a^*)]$

Q. $n = 50$ какова вер-сть того, что наблюдение X_3 не войдёт в бутстрэп выборку №2?

$$\underbrace{\frac{49}{50}\cdot\frac{49}{50}\cdot\ldots\cdot\frac{49}{50}}_{50\text{ раз}} = \left(\frac{49}{50}\right)^{50} = \left(1-\frac{1}{50}\right)^{50} \approx e^{-1} \approx 0{,}37$$

$$\left(1+\frac{1}{n}\right)^n \to e$$

$$\left(1+\frac{\alpha}{n}\right)^n \to e^{\alpha}$$

мораль:
в среднем каждая бутстрэп выборка использует 63% исходной выборки.

бутстрэп даёт возможность посчитать стандар-ую ошибку для $\hat a$.

$\hat a$ в единств-м экземпляре

бутстрэп репликации $\hat a_1^* \ldots \hat a_B^*$ много. $(B = 10000)$ $\quad$ B - кол-во бутстрэп выборок.

j - номер бутстрэп выборки

$$se_B(\hat a) = \sqrt{\frac{\sum_{j=1}^{B} \left(\hat a_j^* - \overline{\hat a^*}\right)^2}{B-1}}$$

$$\overline{\hat a^*} = \frac{\sum_{j=1}^{B} \hat a_j^*}{B}$$

python: arch.

Бутстрэп t-статистики (t-stat bootstrap)

Вернемся к оценке μ.

· классич. алгоритм

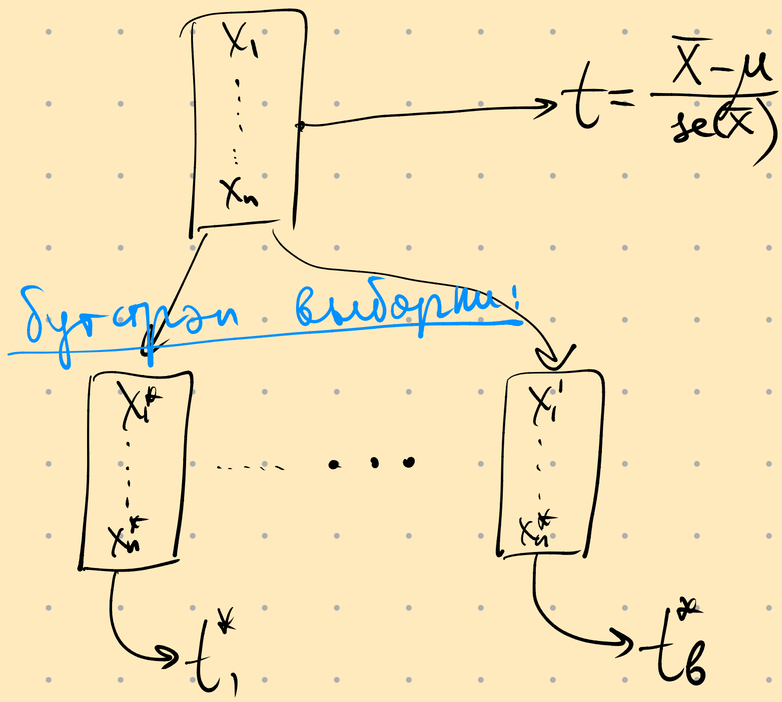
$$t = \frac{\overline{X} - \mu}{se(\overline{X})} \xrightarrow{dist} N(0,1)$$

$$se(\overline{X}) = \sqrt{\hat{\sigma}^2/n}$$

$$\hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

Бутстрэп t-статистики:

будем генерировать случ-ые выборки и решим t-статистики.

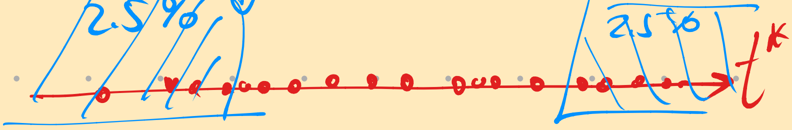


$$t_j^* = \frac{\overline{X}_j^* - \overline{X}}{se(\overline{X}_j^*)}$$

$q_L(t^*)$ $q_R(t^*)$

$B = 10000$ бутстрэп выборок.

исх-ая	бутстрэп №j
$X_1, \dots, X_n$	$X_1^*, \dots, X_n^*$
μ	$\overline{X}$
$\overline{X} = \frac{X_1 + \dots + X_n}{n}$	$\overline{X}^* = \frac{X_1^* + \dots + X_n^*}{n}$
$\hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$	$\hat{\sigma}^2_* = \frac{\sum (X_i^* - \overline{X})^2}{n-1}$
$se(\overline{X}) = \sqrt{\hat{\sigma}^2/n}$	$se(\overline{X}^*) = \sqrt{\frac{\hat{\sigma}^2_*}{n}}$
$t = \frac{\overline{X} - \mu}{se(\overline{X})}$	$t^* = \frac{\overline{X}^* - \overline{X}}{\dots}$



$se(\bar X)$ $se(\bar X^*)$

ЦПТ: при $n \to \infty$ мы должны увидеть норм $N(0;1)$ распр.

индекс j (номер бут-выборки) опущен.

бутстрэп: оценим распр-ие для t исходя из выборки $t_1^* \dots t_B^*$

$\boxed{\text{Ц.П.Т. } n \to \infty}$ $q_L \approx -1{,}96$ $q_R \approx +1{,}96$. $-1{,}96 \le \dfrac{\bar X - \mu}{se(\bar X)} \le 1{,}96$

$$\mu \in [\bar X - 1{,}96 \cdot se(\bar X); \bar X + 1{,}96 \cdot se(\bar X)]$$

$\boxed{\text{бутстрэп:}}$ и руки.

извлекаем квантили $q_L(t^*)$ $q_R(t^*)$ из выборки $t_1^* \dots t_B^*$

$$q_L(t^*) \le \frac{\bar X - \mu}{se(\bar X)} \le q_R(t^*)$$

$$q_L(t^*) \cdot se(\bar X) \le \bar X - \mu \le q_R(t^*) \cdot se(\bar X)$$

$$-q_L(t^*) \cdot se(\bar X) \ge \mu - \bar X \ge -q_R(t^*) \cdot se(\bar X)$$

$\boxed{\text{boot-CI (t-stat)}}$: $\mu \in [\bar X - q_R(t^*) \cdot se(\bar X); \bar X - q_L(t^*) \cdot se(\bar X)]$

Ц.П.Т $n \to \infty$ $q_L \to -1{,}96$ $q_R \to 1{,}96$

$$\mu \in [\bar X - 1{,}96 \cdot se(\bar X); \bar X + 1{,}96 \cdot se(\bar X)]$$

плюсы: быстрее скорость сходимости.

минус: нужно считать и $\hat\theta_j^*$ и $se(\hat\theta_j^*)$

$$t_j^* = \frac{\hat\theta_j^* - \hat\theta}{se(\hat\theta_j^*)}$$

(ещё более вычислительно ёмкое)

исп-ть бутстрэп для вычисления ст. ошибки бутстрэп реплики $\hat{a}_j^*$

$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \xrightarrow{\text{формула/чёрн ящик}} \hat{a}$ — оценка для неизв.-го a.

$b_1 = 10000$ [бут-выборки 1-го ур-ня]

$\begin{bmatrix} X_1^* \\ \vdots \\ X_n^* \end{bmatrix} \longrightarrow \hat{a}_1^*$ $\quad \dots \quad$ $\begin{bmatrix} X_1^* \\ \vdots \\ X_n^* \end{bmatrix} \longrightarrow \hat{a}_b^*$

$$t_1^* = \frac{\hat{a}_1^* - \hat{a}}{se(\hat{a}_1^*)} \qquad t_b^* = \frac{\hat{a}_b^* - \hat{a}}{se(\hat{a}_b^*)}$$

Q. откуда взять?

A запустить бутстрэп внутри бутстрэпа!

бутстрэп 2-го ур-ня

$b_2 = 10000$ $(b_2 = 1000)$

$\begin{bmatrix} X_1^{**} \\ \vdots \\ X_n^{**} \end{bmatrix} \longrightarrow \hat{a}_1^{**}$ $\quad \dots \quad$ $\begin{bmatrix} X_1^{**} \\ \vdots \\ X_n^{**} \end{bmatrix} \longrightarrow \hat{a}_{b_2}^{**}$

$$\hat{a}_1^{**} \;\dots\; \hat{a}_{b_2}^{**} \longrightarrow se(\hat{a}_1^*) = \sqrt{\frac{\sum_{j=1}^{b_2} \left(\hat{a}_j^{**} - \overline{\hat{a}^{**}}\right)^2}{b_2 - 1}}$$

выборок $= b_1 \times b_2$