universität wien

Statistics for Data Science, Winter 2023

Chapter 3:

Bootstrap and Jackknife

Overview

Bootstrap and Jackknife are simulation based methods for inference/uncertainty quantification.

CIs and testing

pros:

- Conceptually simple
- No need for mathematically complex probability calculations

cons:

- Computationally expensive
- Requires large sample size (as for the normal approximation)
- Can go wrong even in large samples!
- Use with care!

RECALL: SAMPLING DISTRIBUTION

$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0,1), \hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$

$\frac{\hat{\theta}_n - \theta}{se_\theta} \xrightarrow[n\to\infty]{d.} N(0,1)$, $se_\theta = \sqrt{\text{Var}_\theta[\hat{\theta}_n]} = \sqrt{\frac{\theta(1-\theta)}{n}}$

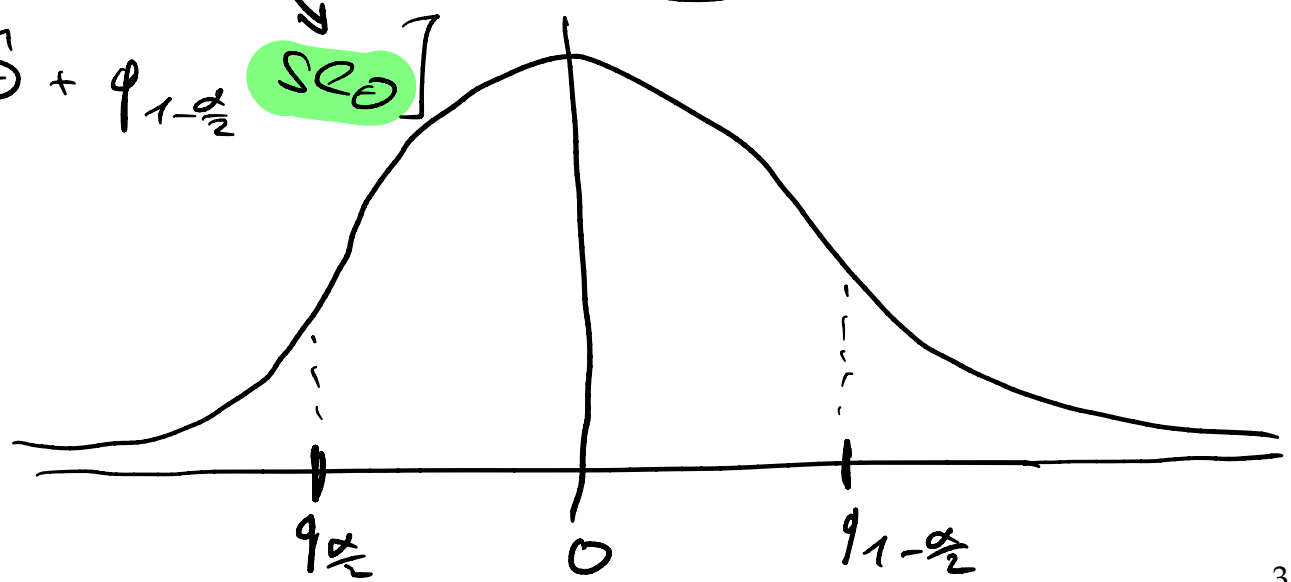
estimate se_θ by plug-in rule $se_{\hat{\theta}_n} = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$

$$P\left(q_{\frac{\alpha}{2}} \le \frac{\hat{\theta} - \theta}{se_\theta} \le q_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha \qquad \left[\frac{\hat{\theta}_n - \theta}{se_\theta} \approx N(0,1)\right]$$

unknown

$$\left[\hat{\theta} - q_{1-\frac{\alpha}{2}} se_\theta,\ \hat{\theta} + q_{1-\frac{\alpha}{2}} se_\theta\right]$$

$$\hat{se} = se_{\hat{\theta}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$



In general

$X_1, \ldots, X_n \overset{iid}{\sim} f_\theta, \theta \in \Theta,$

We often have a CLT:

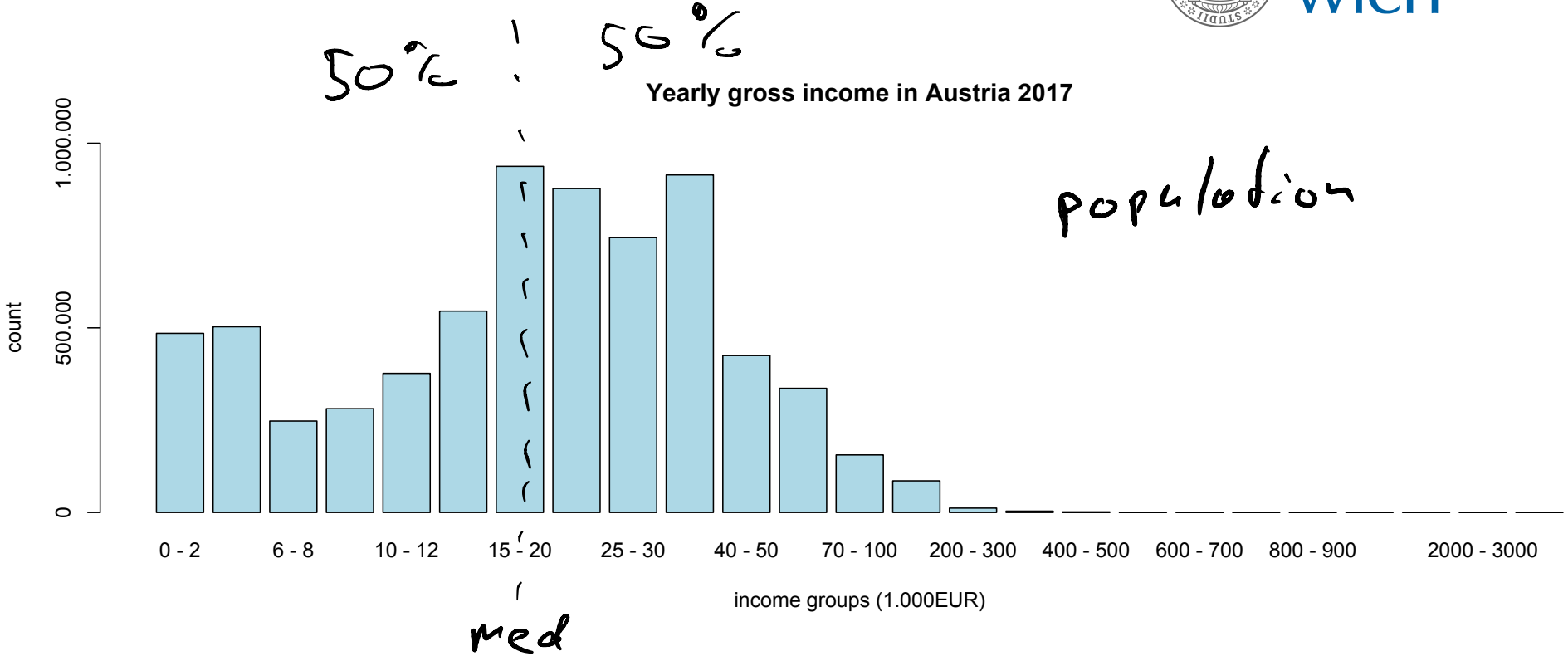
$$\frac{\hat{\theta}_n - \theta}{se_\theta} \xrightarrow[n\to\infty]{d.} N(0,1)$$

But the standard error $se_\theta := \sqrt{\mathrm{Var}_\theta[\hat{\theta}_n]}$ may be hard to compute or the plug-in idea may not be feasible.

First Bootstrap idea: se bootstrap
Estimate se_θ (without knowing its exact analytical form) and rely on the CLT.

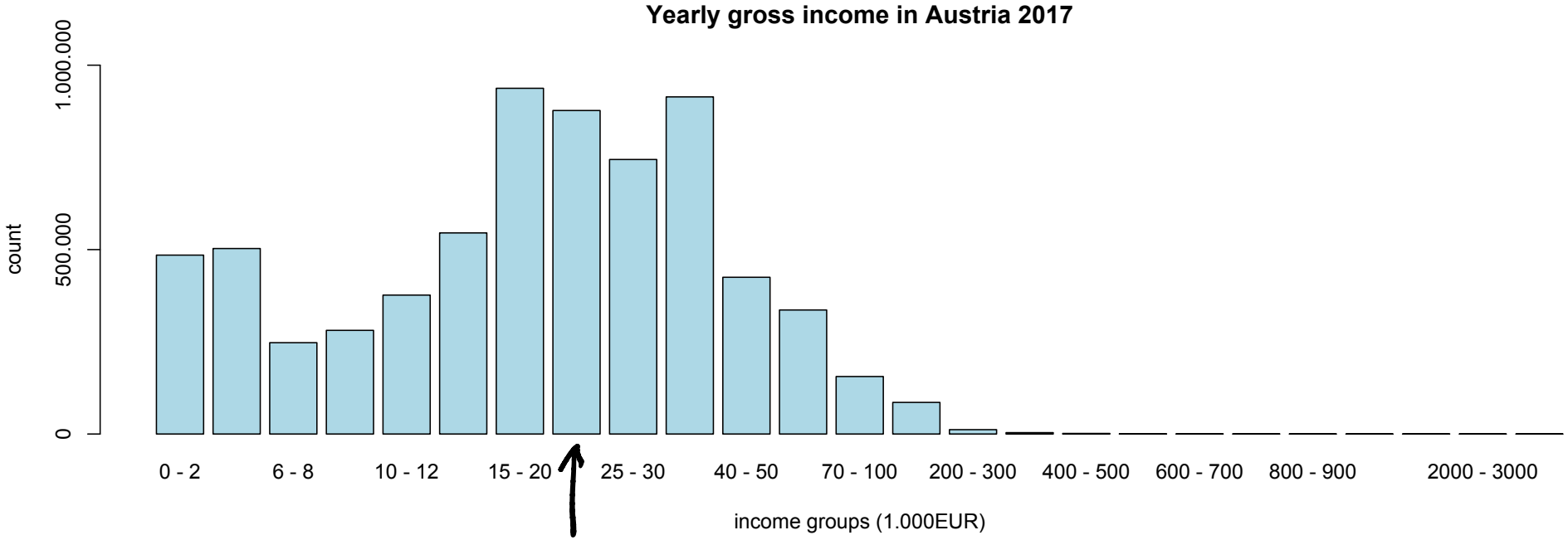
Second Bootstrap idea: pivotal bootstrap
Estimate quantiles of the sampling distribution of $\hat{\theta}_n - \theta$. No Gaussian approximation required.

EXAMPLE: MEDIAN INCOME



Recall: Median = 50% quantile = smallest number $m \in \mathbb{R}$ such that 50% of data are below or equal to m

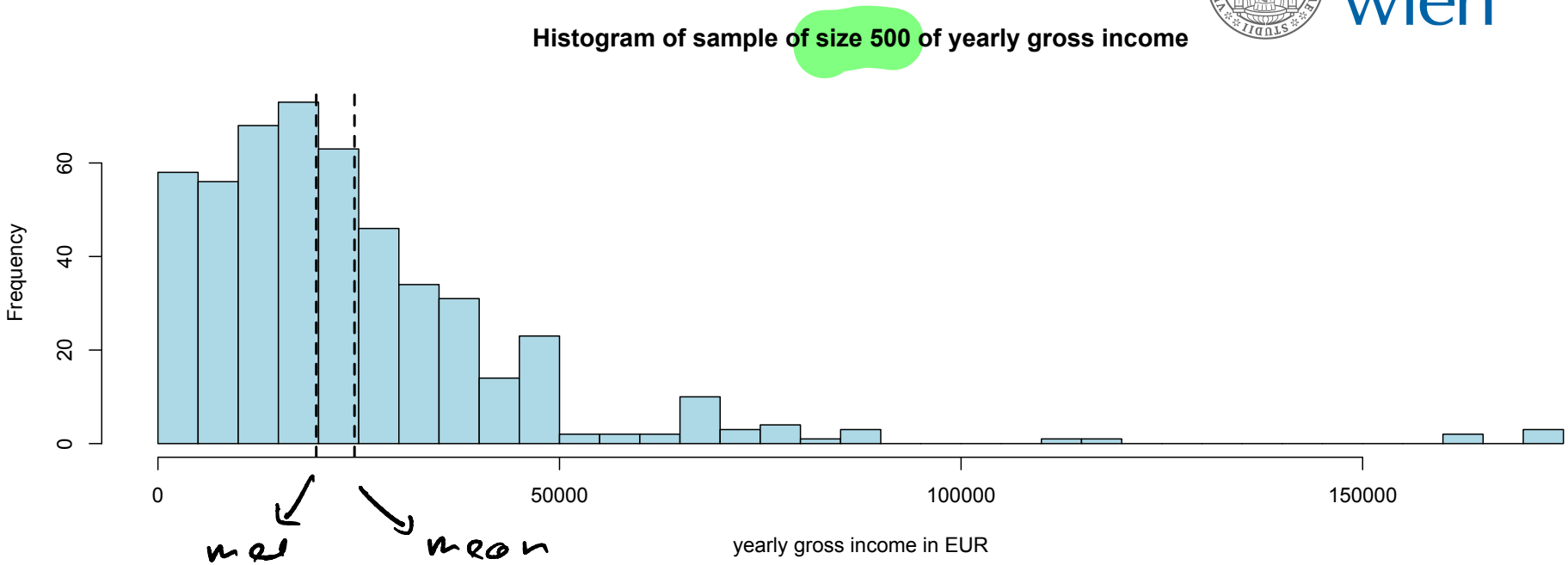
Example: median income



med 20k – 25k

```
> N/2
[1] 3466764
> cumsum(rev(dat$count))
 [1]  485050  987971 1235594 1516636 1893307 2438570
 [7] 3375906 4252996 4997282 5911340 6336321 6672594
[13] 6828489 6914181 6926012 6929730 6931288 6932039
[19] 6932500 6932774 6932960 6933085 6933416 6933475
[25] 6933529
```

EXAMPLE: MEDIAN INCOME



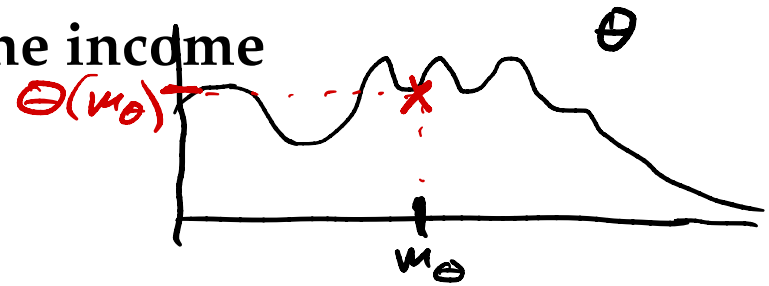
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8	10716	19712	24475	31148	174629

$\hat{\theta}$

Median should be in the $20k$ – $25k$ group!?

EXAMPLE: MEDIAN INCOME

A realistic (nonparametric) model for the income distribution:



$$\mathcal{X} = [0, \infty)^n,$$

$$\Theta = \{\text{all pdfs } \theta \in C^1([0, \infty)) : \theta(m_\theta) > 0\}, \quad m_\theta := \text{med}(\theta)$$

$$f_\theta(x) = \prod_{i=1}^{n} \theta(x_i), \quad x = (x_1, \ldots, x_n)' \in \mathcal{X}, \theta \in \Theta.$$

True population is discrete! Should use pmfs?

$\rightarrow$ Model is an approximation/idealization!

EXAMPLE: MEDIAN INCOME

A realistic (nonparametric) model for the income distribution:

$$\mathcal{X} = [0, \infty)^n,$$

$$\Theta = \{\text{all pdfs } \theta \in C^1([0,\infty)) : \theta(m_\theta) > 0\}, m_\theta := \text{med}(\theta)$$

$$f_\theta(x) = \prod_{i=1}^{n} \theta(x_i), \quad x = (x_1, \ldots, x_n)' \in \mathcal{X}, \theta \in \Theta.$$

Use sample median $\hat{m}_n := \hat{F}_n^\dagger(1/2)$ to estimate population median $m_\theta := \psi(\theta) := F_\theta^\dagger(1/2)$, $F_\theta(x) := \int_{-\infty}^{x} \theta(y)dy$.

It is well known that for every $\theta \in \Theta$,

$$\frac{\hat{m}_n - m_\theta}{se_\theta} \xrightarrow[n\to\infty]{d.} N(0,1)$$

plug-in for Θ? estimate Θ non-parametrically

where $se_\theta = \frac{1}{2\sqrt{n}\theta(m_\theta)}$.

How do we estimate that?!

BOOTSTRAP ESTIMATION OF SE

universität wien

Given data $X = (X_1, \ldots, X_n)'$, $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$, $\theta \in \Theta$, and an estimator $\hat{\theta}_n : \mathcal{X}_0^n \to \mathbb{R}$ of θ (or $\psi(\theta)$), we want to estimate $se_\theta(\hat{\theta}_n) = \sqrt{\mathrm{Var}_\theta[\hat{\theta}_n]}$.

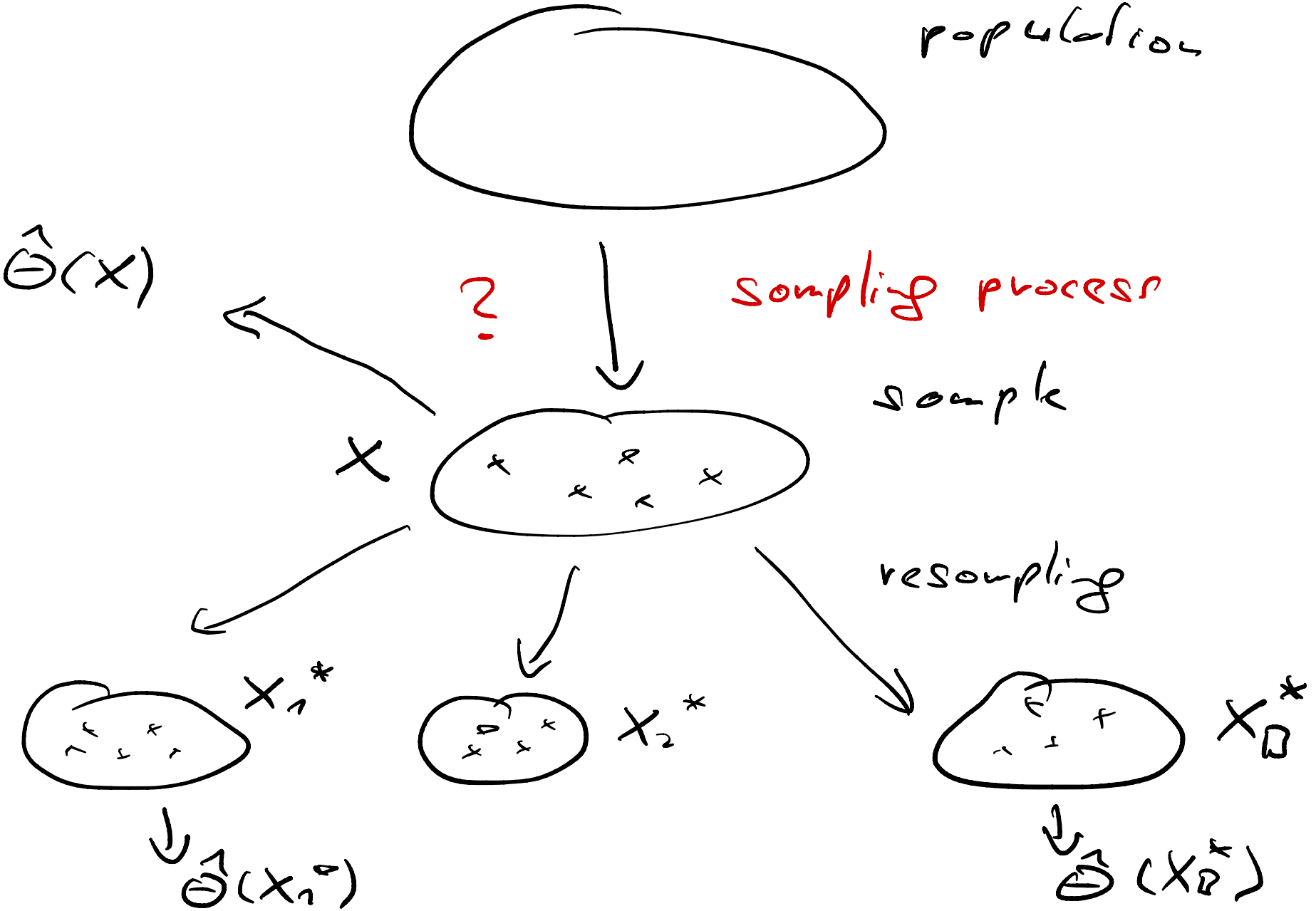
1. Draw a large number B of random samples of size n (with replacement) from the **sample!**

$$
\begin{aligned}
X_1^* &= (X_{1,1}^*, \ldots, X_{n,1}^*)' \\
X_2^* &= (X_{1,2}^*, \ldots, X_{n,2}^*)' \\
&\vdots \\
X_B^* &= (X_{1,B}^*, \ldots, X_{n,B}^*)'
\end{aligned}
$$

2. Compute $\hat{\theta}_n(X_1^*), \ldots, \hat{\theta}_n(X_B^*)$.
3. Approximate $\mathrm{Var}_\theta[\hat{\theta}_n]$ using the LLN (MC idea)

$$
\hat{se}_{boot}^2 = \frac{1}{B}\sum_{j=1}^{B} \hat{\theta}_n(X_j^*)^2 - \left(\frac{1}{B}\sum_{j=1}^{B} \hat{\theta}_n(X_j^*)\right)^2 .
$$

BOOTSTRAP ESTIMATION OF SE



General idea of the Bootstrap:

universität wien

Pretend that the sample is the population and simulate (imitate) the sampling process by 'resampling' from the sample.

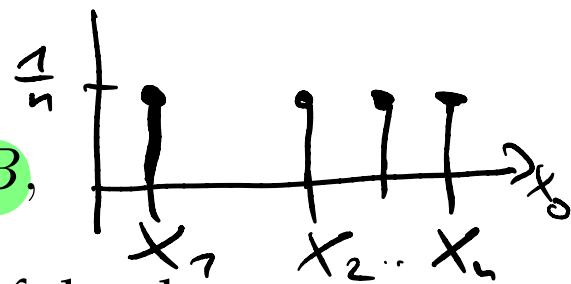
universität wien

Why should this work?

Assume that the original data sample $X = (X_1, \ldots, X_n)'$ from sample space $\mathcal{X} = \mathcal{X}_0^n$ is **given and non-random**.

- Then the Bootstrap draws

$$X_{i,j}^*, \; i = 1, \ldots, n, \; j = 1, \ldots, B,$$



are iid from the **empirical distribution** $\hat{F}_n$ of the data, which has pmf

$$\hat{p}_n(x_0) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i\}}(x_0), \quad x_0 \in \{X_1, \ldots, X_n\}.$$

- Thus, also the $X_1^*, \ldots, X_B^*$ are iid from the n-fold product

$$\hat{f}_n(x) := \prod_{i=1}^{n} \hat{p}_n(x_i), \quad x = (x_1, \ldots, x_n)' \in \{X_1, \ldots, X_n\}^n.$$

Bootstrap estimation of SE

- But if $X_1^*, \ldots, X_B^* \overset{iid}{\sim} \hat{f}_n$, then (by the LLN)

$$\hat{se}^2_{boot} = \frac{1}{B}\sum_{j=1}^{B} \hat{\theta}_n(X_j^*)^2 - \left(\frac{1}{B}\sum_{j=1}^{B} \hat{\theta}_n(X_j^*)\right)^2 \xrightarrow[B\to\infty]{i.p.} \mathrm{Var}_{\hat{f}_n}[\hat{\theta}_n].$$

- Because $X_1^* = (X_{1,1}^*, \ldots, X_{n,1}^*)' \sim \hat{f}_n$ follows an iid model with marginal pmf $\hat{p}_n$, we also write $\mathrm{Var}_{\hat{f}_n}[\hat{\theta}_n] = \mathrm{Var}_{\hat{p}_n}[\hat{\theta}_n]$.
- Since the sample $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$, we have $\hat{F}_n \approx F_\theta$ (recall HW3), $\hat{p}_n \approx p_\theta$ and one can often show that

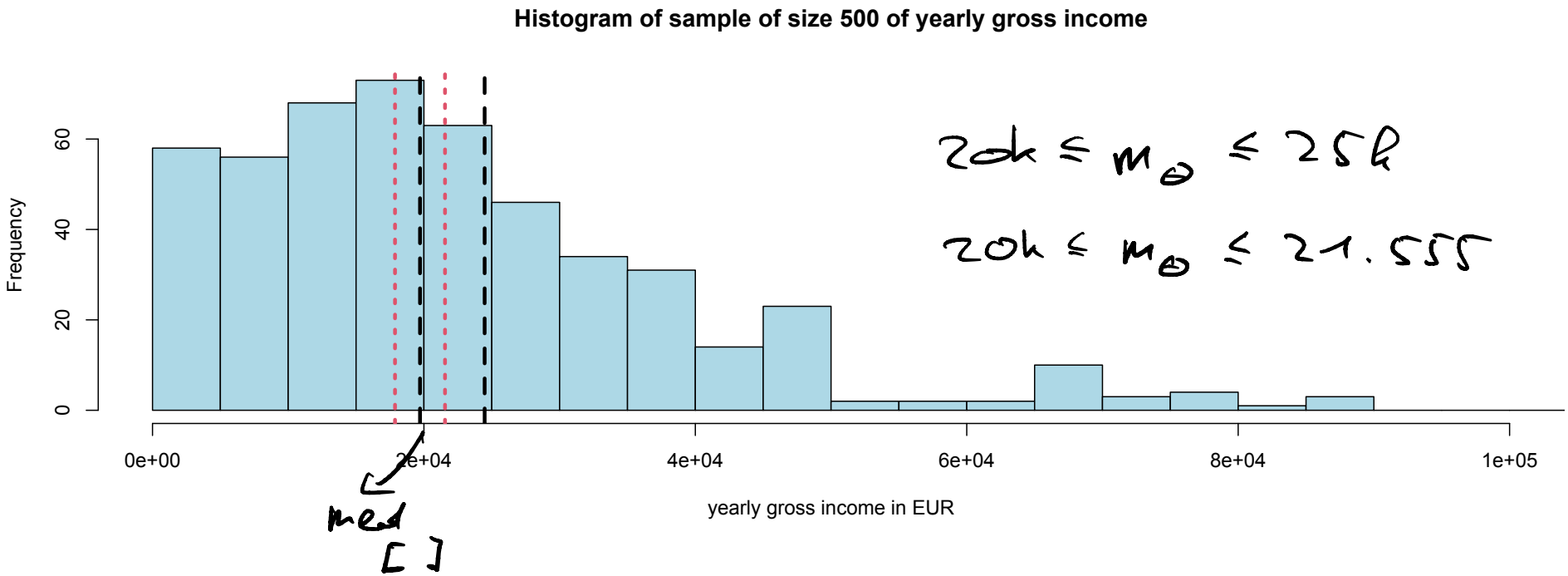
$$\mathrm{Var}_{\hat{p}_n}[\hat{\theta}_n] \approx \mathrm{Var}_{p_\theta}[\hat{\theta}_n] = \mathrm{Var}_\theta[\hat{\theta}_n], \quad \text{if } n \text{ is large.}$$

$$\underbrace{\hat{se}^2_{boot} \approx \mathrm{Var}_{\hat{p}_n}[\hat{\theta}_n]}_{\text{if B is large}}, \qquad \underbrace{\mathrm{Var}_{\hat{p}_n}[\hat{\theta}_n] \approx \mathrm{Var}_\theta[\hat{\theta}_n]}_{\text{if } n \text{ is large}} = se^2_\theta(\hat{\theta}_n)$$

EXAMPLE: MEDIAN INCOME

Compute an approximate 95% ($\alpha = 0.05$) bootstrap confidence interval (by bootstrap estimation of the standard error)

$$CI_\alpha = [\hat{m}_n - q^{(N)}_{1-\frac{\alpha}{2}} \hat{se}_{boot}, \hat{m}_n + q^{(N)}_{1-\frac{\alpha}{2}} \hat{se}_{boot}]$$



$20k \leq m_\Theta \leq 25k$

$20k \leq m_\Theta \leq 21.555$

```
> c(CI_l, CI_u)
[1] 17870 21555
```

```
> CI_u - CI_l
[1] 3685
```

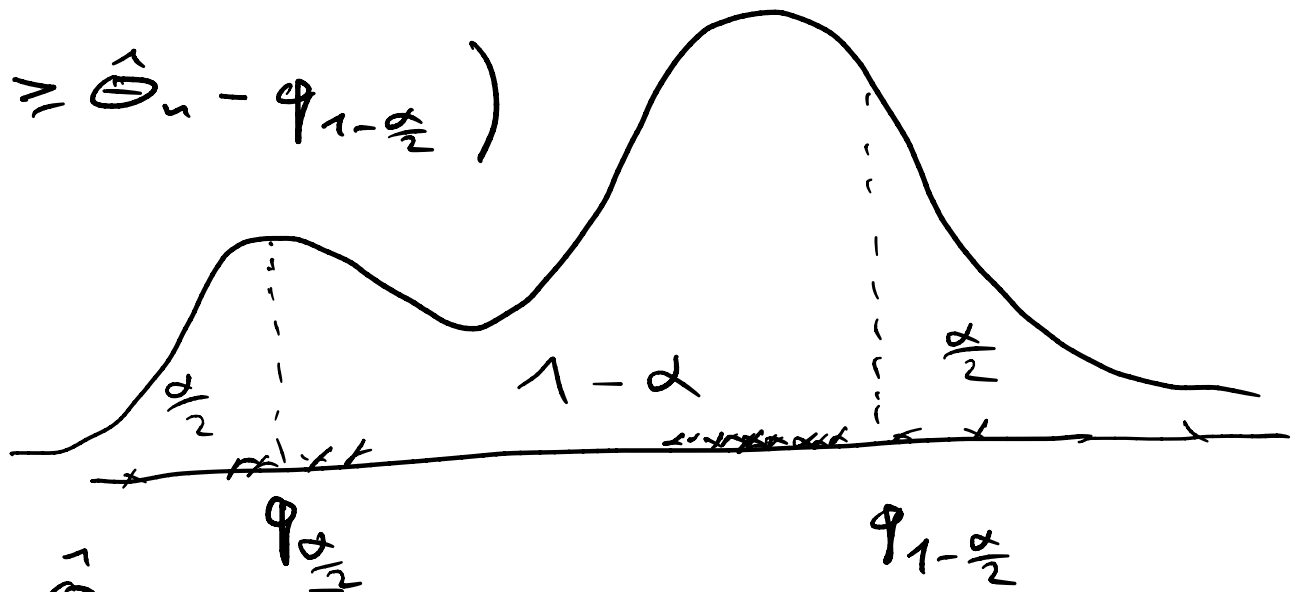
PIVOTAL BOOTSTRAP CI

Given data $X = (X_1, \ldots, X_n)'$, $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$, $\theta \in \Theta$, and an estimator $\hat{\theta}_n : \mathcal{X}_0^n \to \mathbb{R}$ of θ (or $\psi(\theta)$), we want to estimate the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of $\hat{\theta}_n(X) - \theta$.

Why? Say $\hat{\theta}_n - \theta \sim g_n$ is the unknown sampling distribution.

$$P_\theta\left(q_{\frac{\alpha}{2}} \le \hat{\theta}_n - \theta \le q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \qquad \hat{\theta}_n - \theta \sim g_n$$

$$= P_\theta\left(\hat{\theta}_n - q_{\frac{\alpha}{2}} \ge \theta \ge \hat{\theta}_n - q_{1-\frac{\alpha}{2}}\right)$$



$$CI_\alpha = \left[\hat{\theta}_n - q_{1-\frac{\alpha}{2}}, \hat{\theta}_n - q_{\frac{\alpha}{2}}\right]$$

PIVOTAL BOOTSTRAP CI

Given data $X = (X_1, \ldots, X_n)'$, $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$, $\theta \in \Theta$, and an estimator $\hat{\theta}_n : \mathcal{X}_0^n \to \mathbb{R}$ of θ (or $\psi(\theta)$), we want to estimate the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of $\hat{\theta}_n(X) - \theta$.

1. Draw a large number B of random samples of size n (with replacement) from the **sample!**

$$X_1^*, X_2^*, \ldots, X_B^*$$

2. Compute empirical (bootstrap) quantiles $\hat{q}^*_{\alpha/2}$ and $\hat{q}^*_{1-\frac{\alpha}{2}}$ of

$$\hat{\theta}_n(X_1^*) - \hat{\theta}_n(X), \ldots, \hat{\theta}_n(X_B^*) - \hat{\theta}_n(X).$$

3. $CI_\alpha = [\hat{\theta}_n(X) - \hat{q}^*_{1-\frac{\alpha}{2}}, \hat{\theta}_n(X) - \hat{q}^*_{\alpha/2}]$

EXAMPLE: MEDIAN INCOME

Comparison of 95% bootstrap CIs:

$B = 100$

method	lower	median	upper	length
bootstrap se	17870	19712	21555	3685
pivotal bootstrap	18074	19712	21452	3378

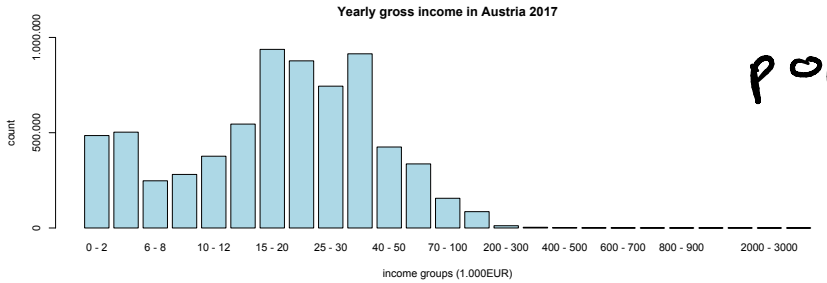
$B = 1000$

method	lower	median	upper	length
bootstrap se	18100	19712	21324	3224
pivotal bootstrap	18161	19712	21388	3227

EXAMPLE: MEDIAN INCOME

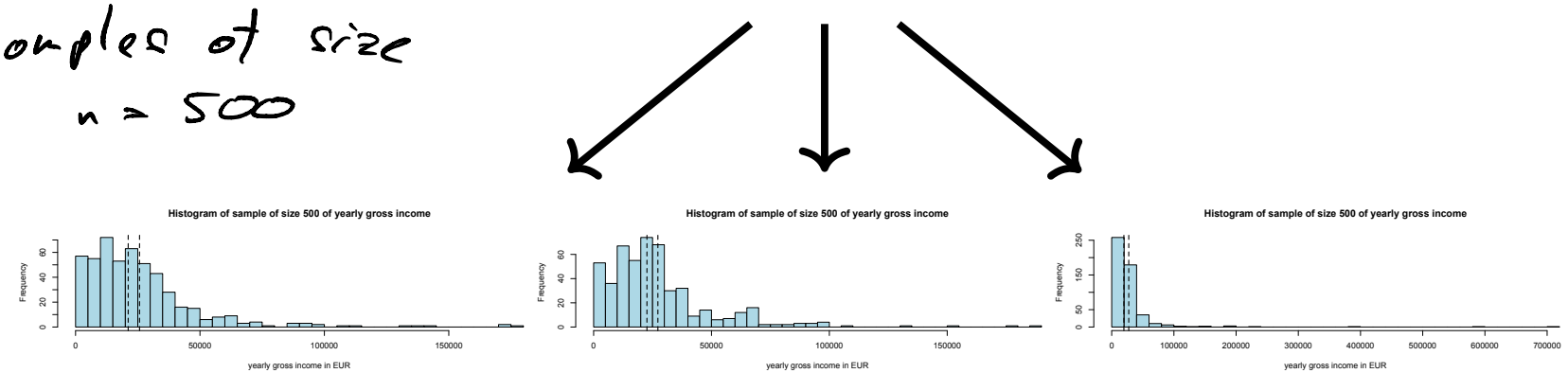
Can we actually trust this CI? Check by simulation!

$$m_\theta = F_\theta^\dagger(1/2)$$

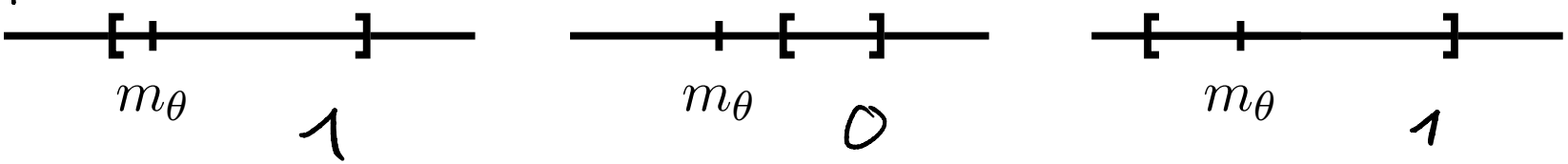


population

samples of size $n = 500$



Bootstrap samples



EXAMPLE: MEDIAN INCOME

Simulation results:

$1 - \alpha = 0.95$

number of (Montecarlo) samples drawn $= 1000$

$B = 1000$

sample size $n = 100$

method	coverage prob.	average length
se bootstrap	0.932	7601
pivotal bootstrap	0.871	7404

$\geq 1-\alpha = 0,95$

sample size $n = 1000$

method	coverage prob.	average length
se bootstrap	0.935	2412
pivotal bootstrap	0.917	2377

Failure of the Bootstrap

Recall: Bootstrap idea

- Regard your data sample as the population.
- Draw B iid random (bootstrap) samples from the sample (re-sample) like in a MC simulation.
- Compute your estimator on each of the bootstrap samples.
- Use this resulting 'bootstrap distribution' of your estimator as an approximation of its true unknown sampling distribution.

Recall

- In the median income example we simulated the actual coverage probability of the bootstrap CIs and found

$$P_\theta(m_\theta \in CI_\alpha) \approx 1 - \alpha.$$

- We did that for the true data generating parameter $\theta \in \Theta$. (we cheated!)
- We concluded that the bootstrap works relatively well for n large.
- In practice we would try many different choices for θ.

Failure of the Bootstrap

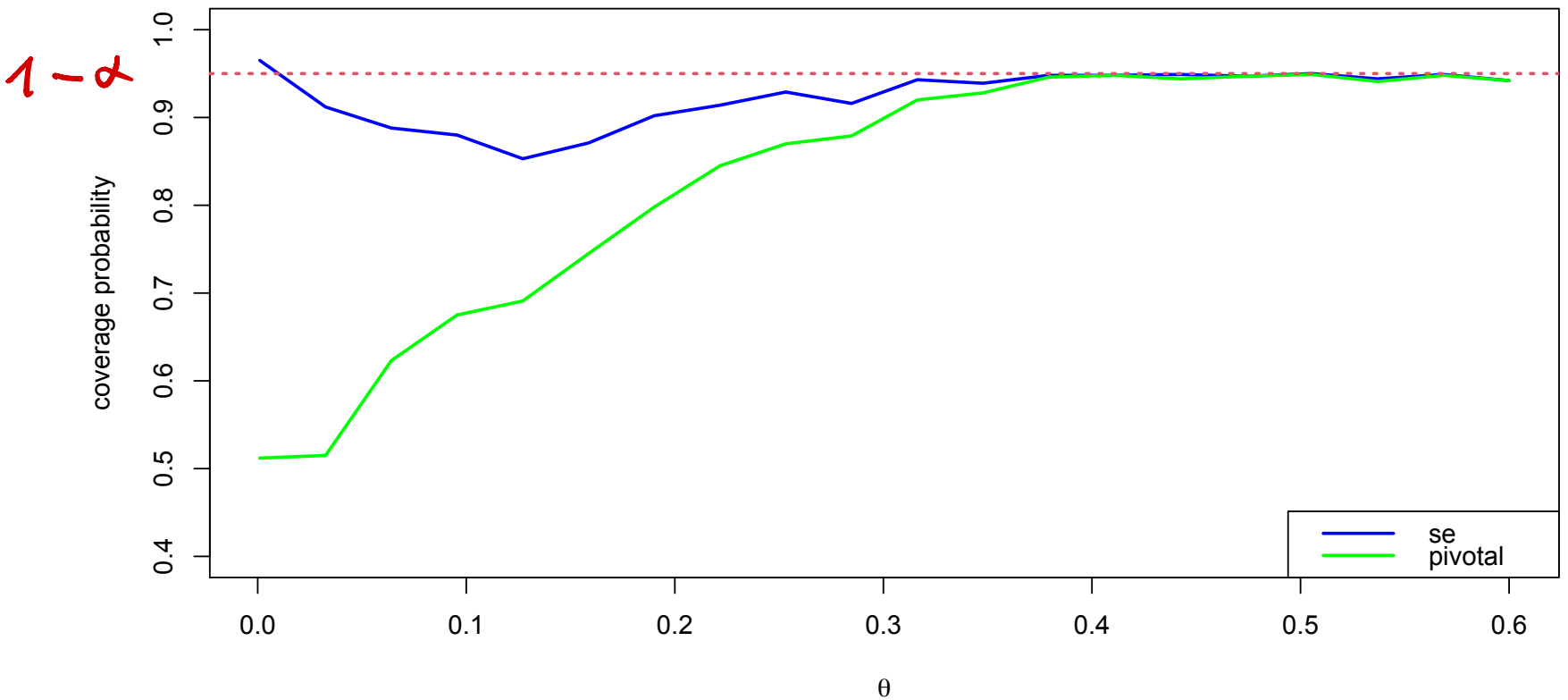
universität wien

- Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \Theta = [0, \infty)$.
- Goal: Bootstrap inference on θ.
- The classical estimator (MLE) is

$$\hat{\theta}_n = \max\{0, \bar{X}_n\}, \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Failure of the Bootstrap

- $n = 100, \alpha = 0.05, B = 1000$
- 1000 MC iterations

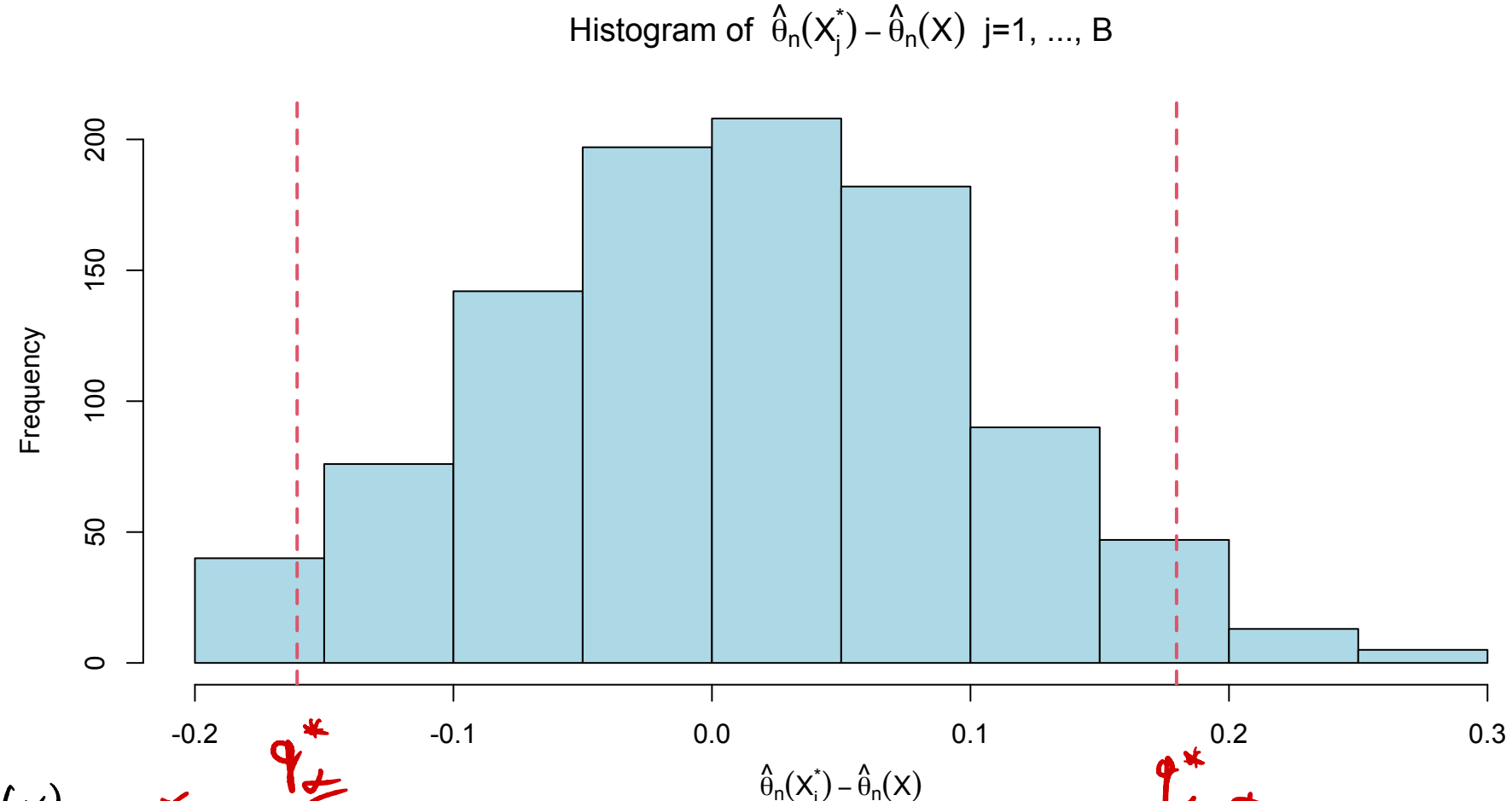


Failure of the Bootstrap

What went wrong?

Look at the 'bootstrap distribution' for a sample $X = (X_1, \ldots, X_n)$ with $\bar{X}_n = 0.185$.

$\hat{\theta}_n(X_j^*) = \max\{0, \bar{X}_j^*\}$

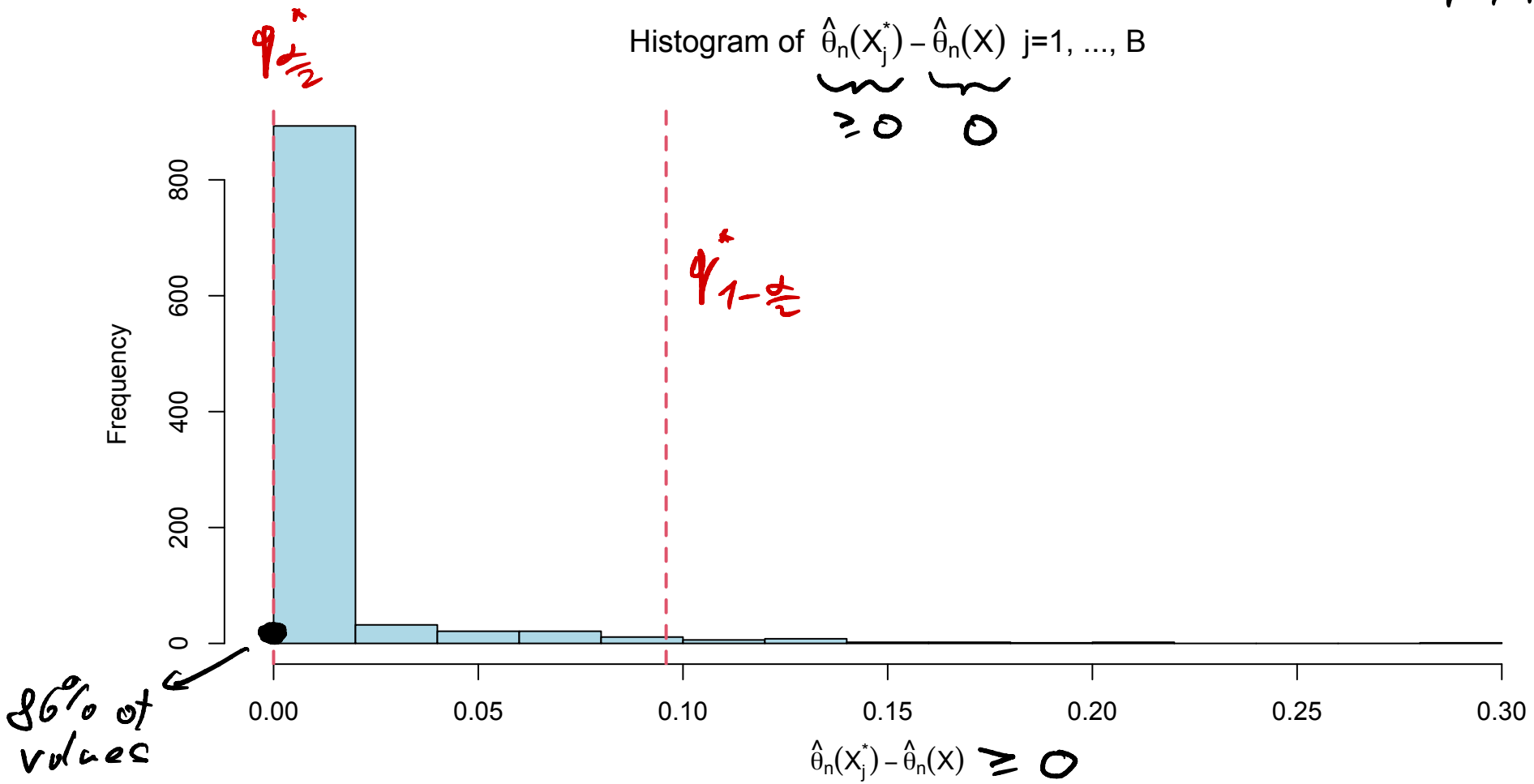


$q^*_{\alpha/2}$ $q^*_{1-\alpha/2}$

$\left[\hat{\theta}_n(X) - q^*_{1-\alpha/2},\ \hat{\theta}_n(X) - q^*_{\alpha/2}\right]$

Failure of the Bootstrap

Look at the 'bootstrap distribution' for a sample with $\bar{X}_n = -0.122$.



Here, actually 86% of bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$.

Failure of the Bootstrap

- Look at the 'bootstrap distribution' for a sample with $\bar{X}_n = -0.122$.
- Notice: for every $j \in \{1, \ldots, B\}$, we have

$$\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = \max\{0, \bar{X}_j^*\} - \max\{0, \bar{X}_n\} \geq 0,$$

$$\Rightarrow 0 \leq \hat{q}^*_{\alpha/2} \leq \hat{q}^*_{1-\alpha/2}.$$

- Therefore,

$$CI_\alpha = [\hat{\theta}_n(X) - \hat{q}^*_{1-\alpha/2}, \hat{\theta}_n(X) - \hat{q}^*_{\alpha/2}] \subseteq (-\infty, 0] \not\ni \theta > 0.$$

$= 0$ $= 0$

Failure of the Bootstrap

Recall: $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$

$\alpha = 0.05$

Thus, intuitively, for very small $\theta > 0$,

$$P_\theta(\theta \in CI_{0.05}) = P_\theta(\theta \in CI_{0.05} | \bar{X}_n < 0) P_\theta(\bar{X}_n < 0)$$

$= 0$ $\qquad \approx \frac{1}{2}$

$\bar{X}_n \sim N(\theta, \frac{1}{n})$

$$+ P_\theta(\theta \in CI_{0.05} | \bar{X}_n \geq 0) P_\theta(\bar{X}_n \geq 0)$$

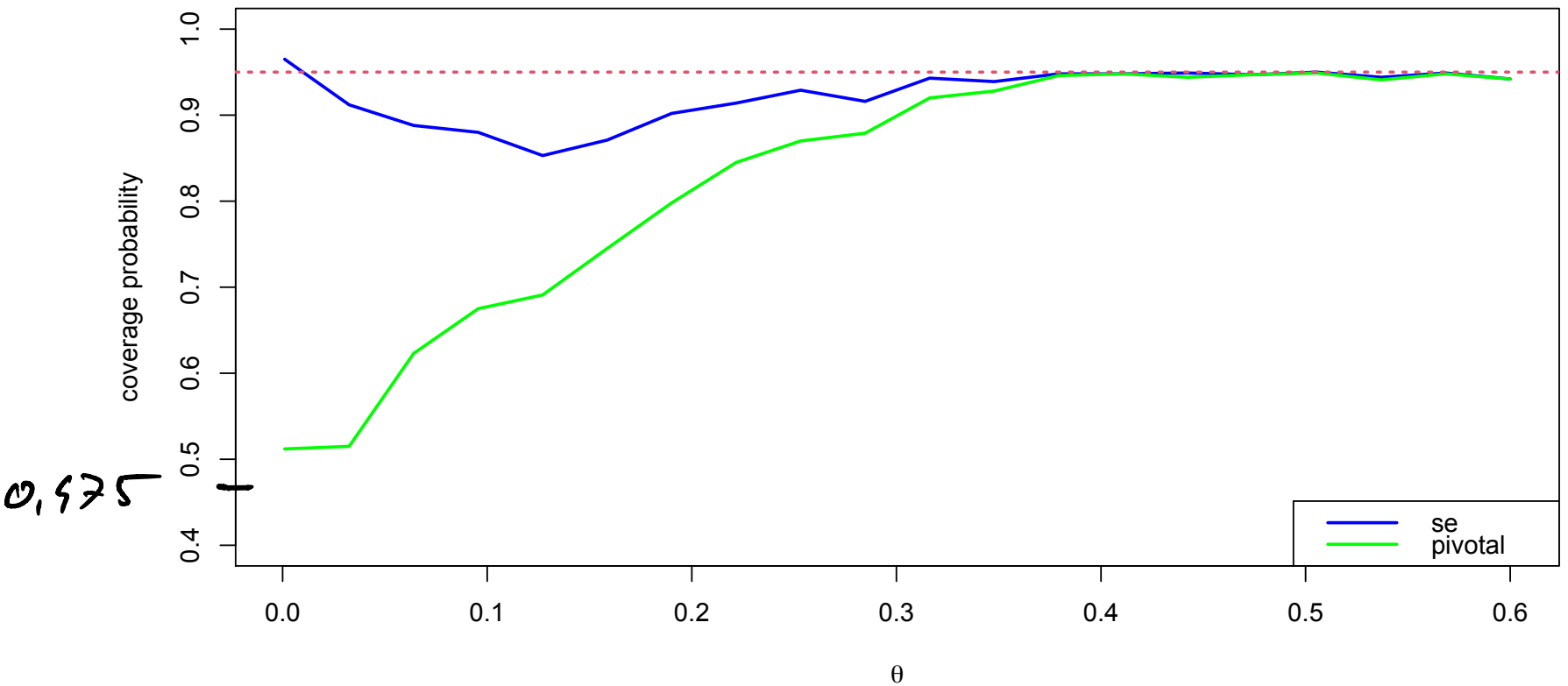
≈ 0.95 $\qquad \approx \frac{1}{2}$

≤ 1

$\approx \quad 0.475$

Failure of the Bootstrap

- $n = 100, \alpha = 0.05, B = 1000$
- 1000 MC iterations

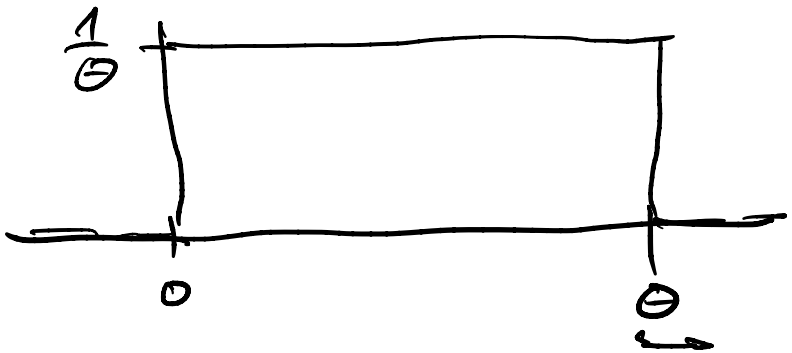


Failure of the Bootstrap

Is the se-bootstrap always superior to the pivotal method?

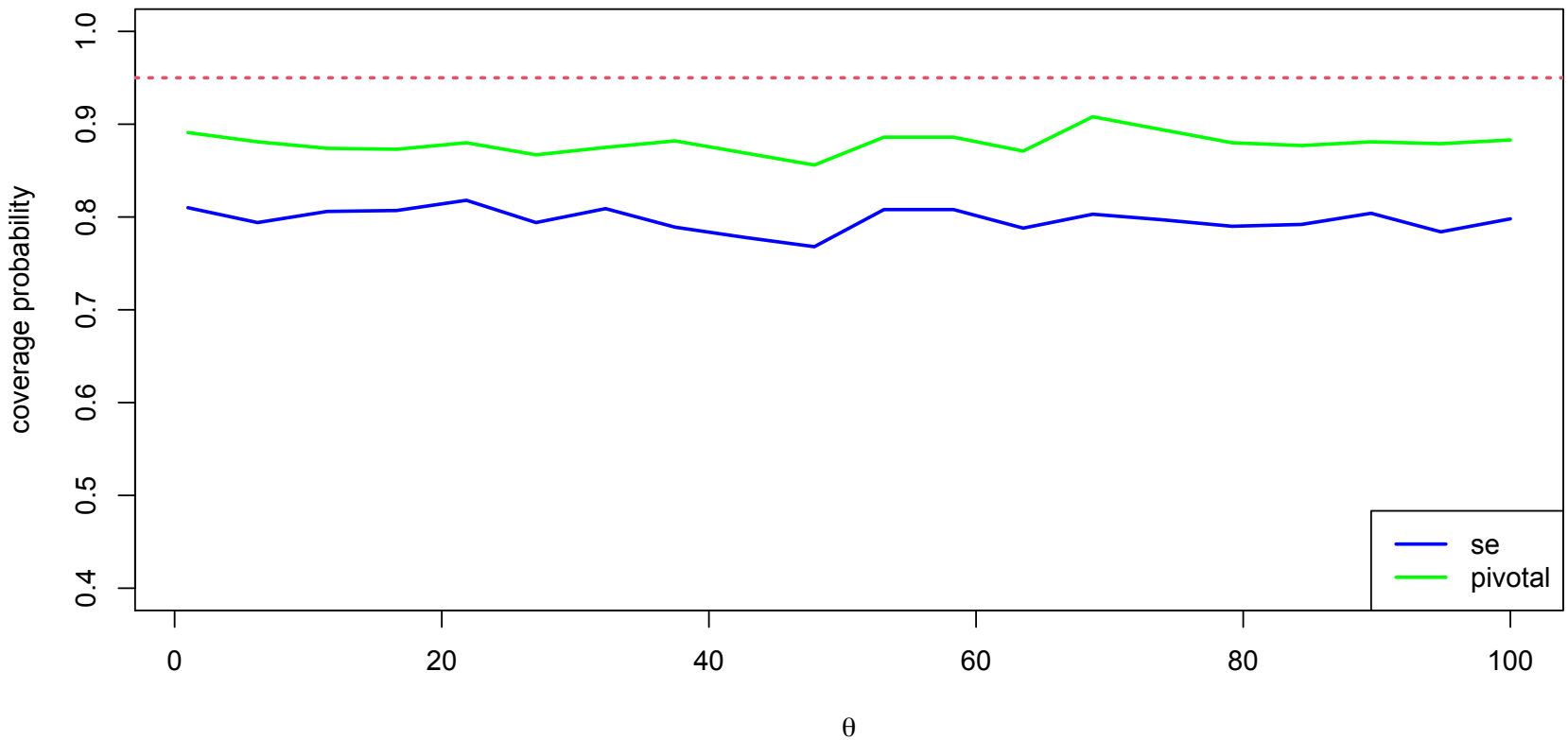
- Consider $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[0, \theta]$, $\theta \in \Theta = (0, \infty)$.
- Goal: Bootstrap inference on θ.
- The classical estimator (MLE) is

$$\hat{\theta}_n = \max\{X_1, \ldots, X_n\}.$$



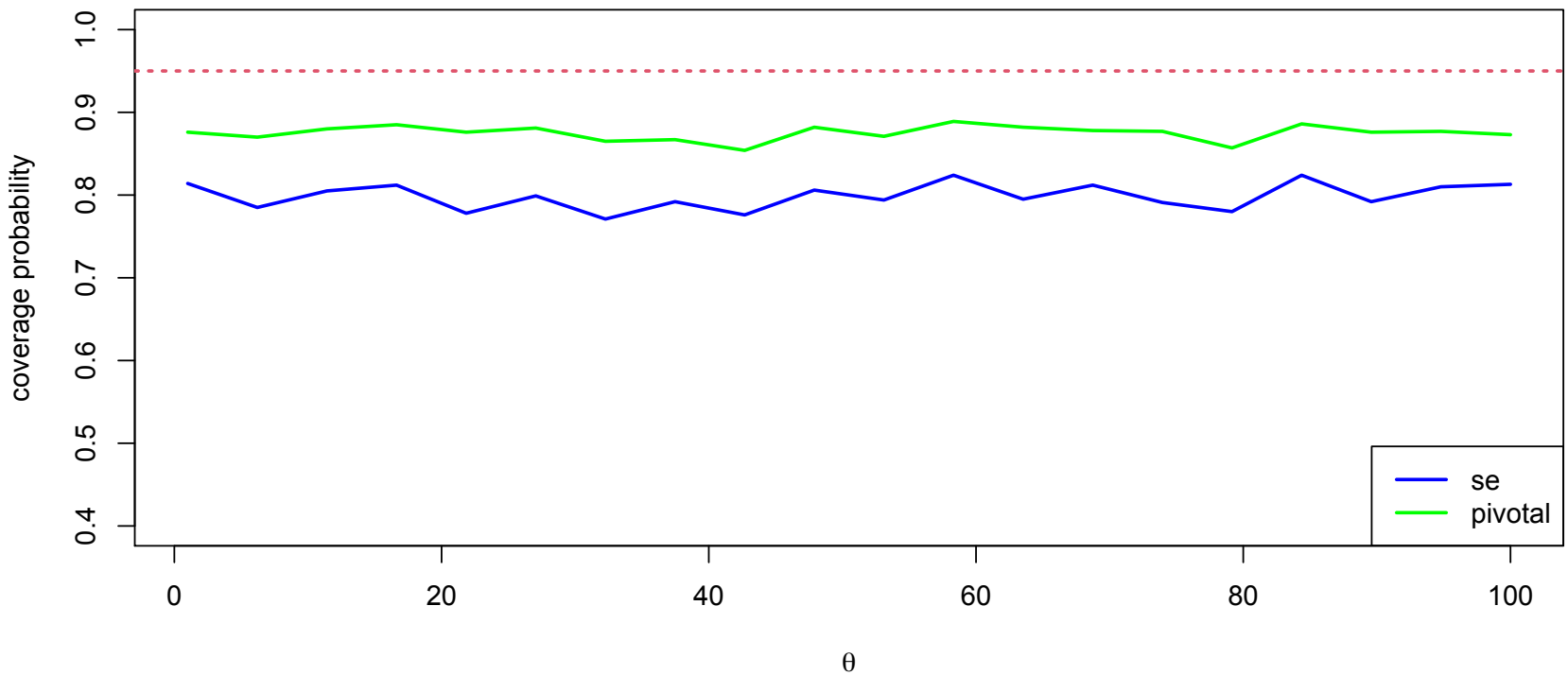
Failure of the Bootstrap

- $n = 100, \alpha = 0.05, B = 1000$
- 1000 MC iterations



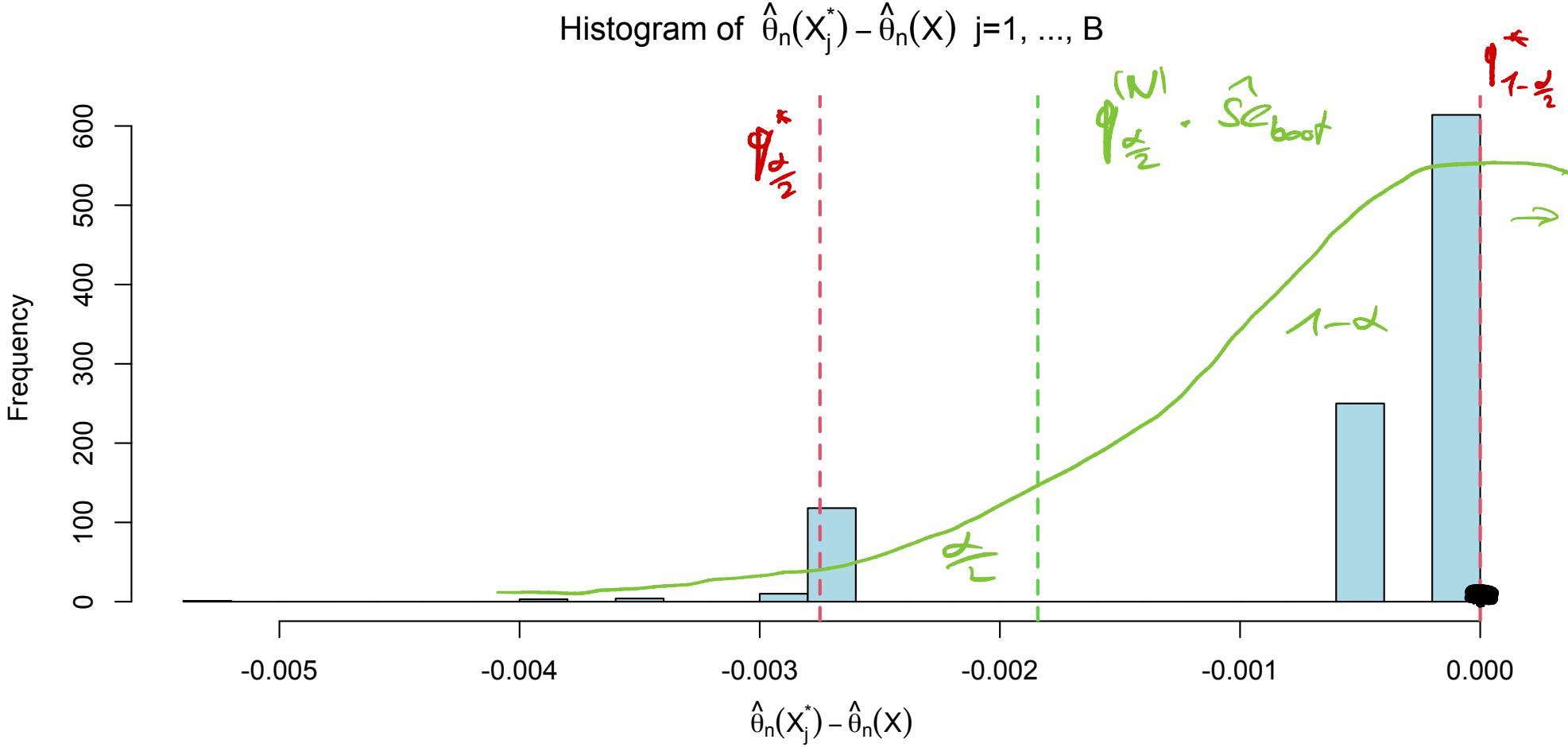
Failure of the Bootstrap

- $n = 1000$, $\alpha = 0.05$, $B = 1000$
- 1000 MC iterations



Failure of the Bootstrap

Look at the bootstrap distribution of one given sample of size $n = 1000$ with $\theta = 1$.



Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$.

Failure of the Bootstrap

Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$. Why?

assume no ties

$$\hat{\theta}_n(X_1^*) - \hat{\theta}_n(X) = \max\{X_{1,1}^*, \ldots, X_{n,1}^*\} - \max\{X_1, \ldots, X_n\}$$

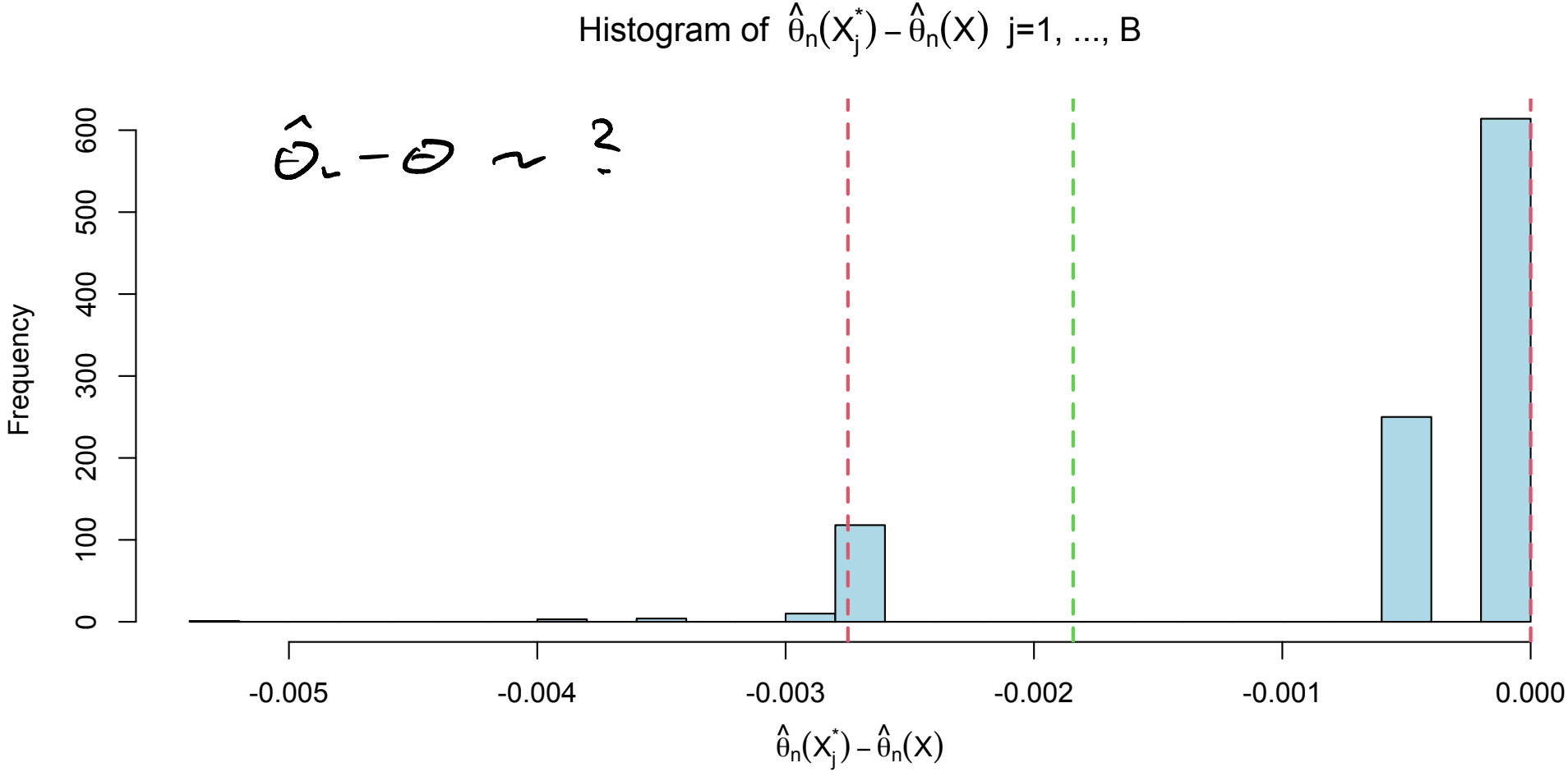
- This is equal to 0 if, and only if, in our bootstrap sample $X_{1,1}^*, \ldots, X_{n,1}^*$ we happen to draw the largest sample point from $\{X_1, \ldots, X_n\}$.
- What is the probability of that happening?

$$1 - \left(\frac{n-1}{n}\right)^n = 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow[n\to\infty]{} 1 - e^{-1}$$

$$\approx 0{,}63$$

Failure of the Bootstrap

Look at the bootstrap distribution of one given sample of size $n = 1000$ with $\theta = 1$.



Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$.

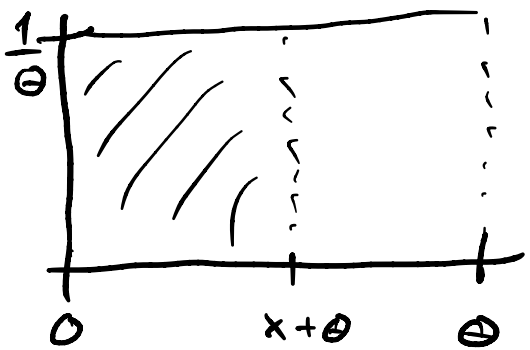
What is the actual sampling distribution of $\hat{\theta}_n - \theta$?

Recall:
$X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[0, \theta], \theta \in \Theta = (0, \infty), \hat{\theta}_n = \max\{X_1, \ldots, X_n\}.$

$$P_\theta(\hat{\theta}_n - \theta \leq x) = P_\theta\left(\max\{X_1 \ldots X_n\} \leq x + \theta\right)$$

$$= P_\theta\left(X_1 \leq x+\theta \text{ and } X_2 \leq x+\theta \ldots \text{ and } X_n \leq x+\theta\right)$$

$$\overset{\text{indep.}}{=} \prod_{i=1}^{n} P_\theta(X_i \leq x+\theta)$$



$$= \begin{cases} 0 & \text{if } x+\theta < 0, \text{ if } x < -\theta \\ \frac{x+\theta}{\theta} & \text{if } 0 \leq x+\theta \leq \theta, -\theta \leq x \leq 0 \\ 1 & \text{if } x+\theta > \theta, x > 0 \end{cases}$$

$$P_\theta(\hat{\theta}_n - \theta \le x) = \begin{cases} 0 & x < -\theta \\ \left(1 + \frac{x}{\theta}\right)^n & -\theta \le x \le 0 \\ 1 & x > 0 \end{cases}$$

$$g_n(x) = \frac{d}{dx} P_\theta(\hat{\theta}_n - \theta \le x)$$

$$= \begin{cases} 0, & x < -\theta \\ n\left(1 + \frac{x}{\theta}\right)^{n-1} \frac{1}{\theta}, & -\theta \le x \le 0 \\ 0, & x > 0 \end{cases}$$

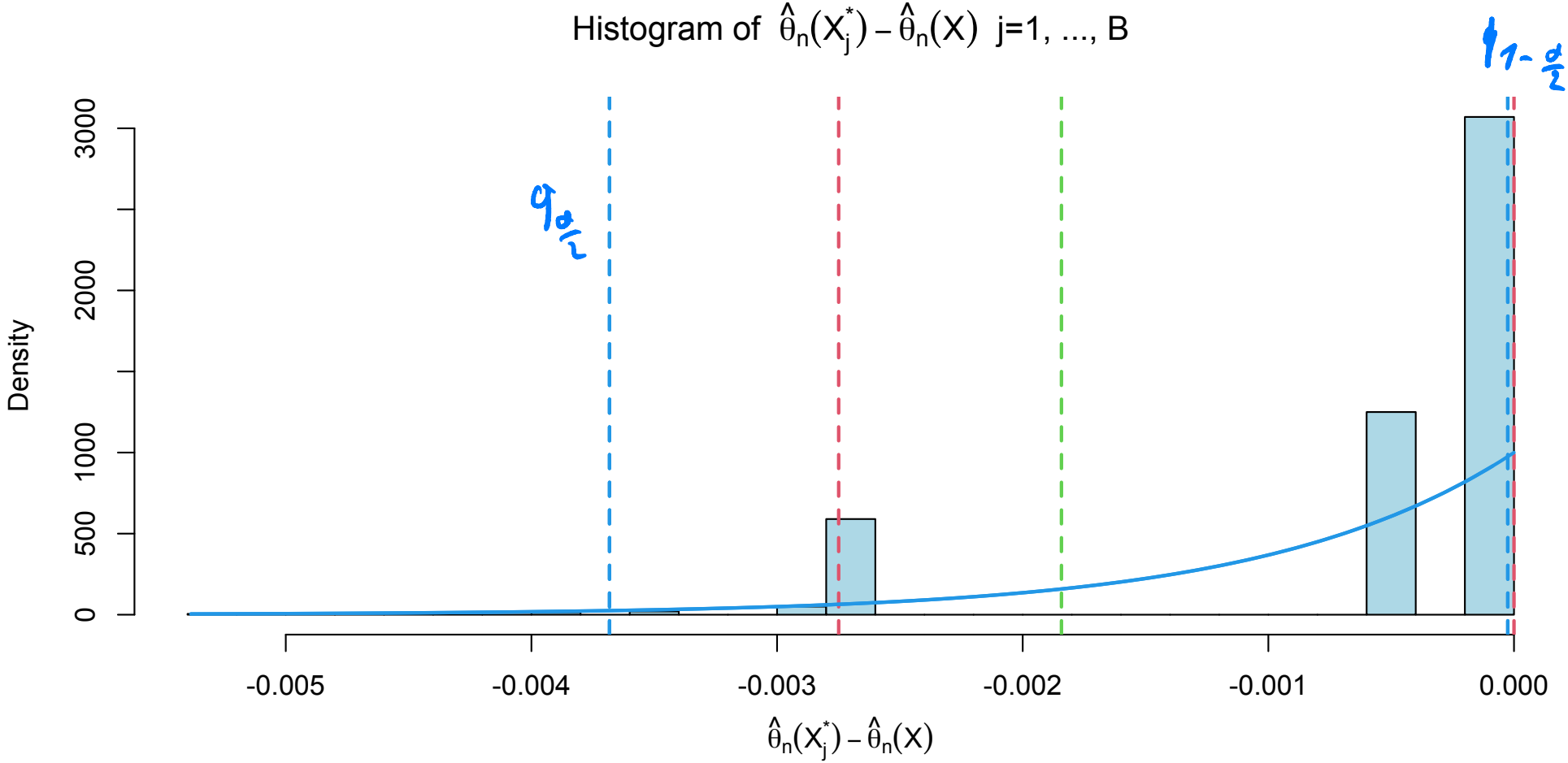
pdf of sampling distribution

$$q_\alpha^{(\hat{\theta}_n)} = \left[\alpha^{\frac{1}{n}} - 1\right]\theta$$

HW

Failure of the Bootstrap

Bootstrap vs. true sampling distribution of $\hat{\theta}_n - \theta$. $\theta = 1, n = 1000$.



Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$.

universität wien

Always take a look at the bootstrap distribution!!!

Summing up:

- Computationally expensive simulation method.
- Very flexible, generic, no need for problem specific formulas.
- Produces only approximate inference; requires large samples.
- Can go wrong!
- In general: Works, if the normal approximation works.