Бутстреп

а) Моделирование Монте-Карло, бутстреп

__Бутстреп__ – метод оценивания доверительных интервалов и подсчета ошибок

Пусть задана i.i.d. выборка $X_1, \ldots, X_n \sim F$**,** $T_n = g(X_1, \ldots, X_n)$ **-**
1 **статистика от нее**

Необходимо оценить $\mathbb{V}_F(T_n)$

__Пример__. $T_n = \overline{X}_n$ **- оценка среднего значения, тогда** $\mathbb{V}_F(T_n) = \sigma^2/n$**, где** $\sigma^2 = \int (x-\mu)^2 dF(x)$ **и** $\mu = \int x dF(x)$

<u>Основная идея</u>:

Шаг 1. Оценить $\mathbb{V}_F(T_n)$ **с помощью** $\mathbb{V}_{\widehat{F}_n}(T_n)$

Шаг 2. Аппроксимировать значение $\mathbb{V}_{\widehat{F}_n}(T_n)$ **используя моделирование**

Монте-Карло

<u>Пример</u>. В случае $T_n = \overline{X}_n$ **получаем, что** $\mathbb{V}_{\widehat{F}_n}(T_n) = \widehat{\sigma}^2/n$**, где** $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)$**, при этом шаг 2 не требуется. Шаг 2 требуется в более сложных случаях, в которых не удается выписать явной формулы для** $\mathbb{V}_{\widehat{F}_n}(T_n)$**.**

Основная идея метода Монте-Карло состоит в следующем.

Пусть задана i.i.d. выборка $Y_1, \dots, Y_B$**, распределение элементов которой равно** G**. По закону больших чисел при** $B \to \infty$

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{\mathrm{P}} \int y\, dG(y) = \mathbb{E}(Y)$$

В общем случае для функции h**, у которой среднее значение относительно распределения** G **конечно, получаем, что**

$$\frac{1}{B}\sum_{j=1}^{B} h(Y_j) \xrightarrow{\mathrm{P}} \int h(y) dG(y) = \mathbb{E}(h(Y))$$

В частности,

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j-\overline{Y})^2 = \frac{1}{B}\sum_{j=1}^{B}Y_j^2-\left(\frac{1}{B}\sum_{j=1}^{B}Y_j\right)^2$$

$$\xrightarrow{\mathrm{P}} \int y^2 dF(y)-\left(\int y\,dF(y)\right)^2=\mathbb{V}(Y)$$

=> можно использовать выборочную дисперсию сгенерированной выборки для оценки дисперсии $\mathbb{V}(Y)$

b) Оценка дисперсии на основе бутстрепа

Итак, значение $\mathbb{V}_{\widehat{F}_n}(T_n)$ можно оценить с помощью моделирования

$\mathbb{V}_{\widehat{F}_n}(T_n)$ **= дисперсия T_n в случае, когда распределение элементов выборки равно $\widehat{F}_n$ => надо моделировать выборки, в которых каждый из элементов имеет распределение $\widehat{F}_n$, после чего оценить $\mathbb{V}_{\widehat{F}_n}(T_n)$**

$$
\begin{array}{ccccc}
F & \Longrightarrow & X_1, \dots, X_n & \Longrightarrow & T_n = g(X_1, \dots, X_n) \\
\widehat{F}_n & \Longrightarrow & X_1^*, \dots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \dots, X_n^*)
\end{array}
$$

$\widehat{F}_n$ **<=> наблюдения могут принимать одно из значений** $X_1, \ldots, X_n$**, при этом вероятность получить любое из этих значений равна** $1/\mathrm{n}$

6

Таким образом, чтобы получить псевдовыборку $\widehat{F}_n \quad \Longrightarrow \quad X_1^*, \ldots, X_n^*$**, надо случайным образом выкинуть номера** $\{i_1, \ldots, i_n\}$**, где** $P(i_j = k) = 1/n,\ k = 1, \ldots, n$**, тогда** $\{X_1^*, \ldots, X_n^*\} = \{X_{i_1}, \ldots, X_{i_n}\}$ **(выборка с возвращением)**

<u>Оценка дисперсии с помощью бутстрепа</u>:

1. **Смоделировать** $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
2. **Подсчитать** $T_n^* = g(X_1^*, \ldots, X_n^*)$
3. **Повторить шаги 1 и 2, получить** $T_{n,1}^*, \ldots, T_{n,B}^*$
4. **Подсчитать оценку дисперсии по формуле**

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*\right)^2$$

<u>Псевдокод для оценки дисперсии оценки медианы</u>:

Входные данные `X = (X(1), ..., X(n)):`

```
T <- median(X)
```

8

```
Tboot <- вектор длины B

for(i in 1:B){
    Xstar <- выборка объема n с возвращением из X
    Tboot[i] <- median(Xstar)
    }
se <- sqrt(variance(Tboot))
```

$$\mathbb{V}_F(T_n) \approx \mathbb{V}_{\widehat{F}_n}(T_n) \approx v_{boot}$$

<u>Пример</u> (данные о моментах времени между последовательными импульсами вдоль нервного волокна):

9 $\theta = T(F) = \int (x-\mu)^3 dF(x)/\sigma^3$ **- коэффициент ассиметрии**

$$\widehat{\theta} = T(\widehat{F}_n) = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\widehat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^3}{\widehat{\sigma}^3} = 1.76$$

Применяя метод оценки дисперсии на основе бутстрепа при $B = 1,000$ **получаем, что стандартная ошибка оценки коэффициента ассиметрии равна 0.16**

с) Оценка доверительных интервалов на основе бутстрепа

<u>Метод 1</u>: нормальный интервал.

$$T_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}_{\mathrm{boot}},$$

где $\widehat{\mathsf{se}}_{\mathrm{boot}} = \sqrt{v_{\mathrm{boot}}}$ **- оценка стандартной ошибки на основе бутстрепа. Метод хорошо работает, если распределение данных близко к нормальному распределению**

<u>Метод 2</u>: центральный интервал. Пусть $\theta = T(F)$ **и** $\widehat{\theta}_n = T(\widehat{F}_n)$**. Положим** $R_n = \widehat{\theta}_n - \theta$**. Обозначим через** $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ **- повторную выборку значений** $\widehat{\theta}_n = T(\widehat{F}_n)$ **на основе бутстрепа, а через** $H(r)$ **- распределение величины** R_n**, то есть**

$$H(r) = \mathbb{P}_F(R_n \leq r)$$

Определим доверительный интервал согласно формуле $C_n^\star = (a, b)$**, где**

$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right)$ **и** $b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$**.**

Очевидно, что

$$
\begin{aligned}
\mathbb{P}(a \le \theta \le b) &= \mathbb{P}(a - \widehat{\theta}_n \le \theta - \widehat{\theta}_n \le b - \widehat{\theta}_n) \\
&= \mathbb{P}(\widehat{\theta}_n - b \le \widehat{\theta}_n - \theta \le \widehat{\theta}_n - a) \\
&= \mathbb{P}(\widehat{\theta}_n - b \le R_n \le \widehat{\theta}_n - a) \\
&= H(\widehat{\theta}_n - a) - H(\widehat{\theta}_n - b) \\
&= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha
\end{aligned}
$$

Однако, распределение $H(r)$ неизвестно. Будем использовать оценку распределения $H(r)$ на основе бутстрепа, то есть положим

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R^*_{n,b} \le r)$$

,

где $R^*_{n,b} = \widehat{\theta}^*_{n,b} - \widehat{\theta}_n$. Пусть r^*_β обозначает β выборочную квантиль, подсчитанную по выборке значений $(R^*_{n,1}, \ldots, R^*_{n,B})$, а θ^*_β обозначает β выборочную квантиль, подсчитанную по выборке значений $(\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B})$. Несложно показать, что $r^*_\beta = \theta^*_\beta - \widehat{\theta}_n$.

Тогда приблизительный доверительный интервал с доверительной вероятностью $1-\alpha$ **имеет вид** $C_n = (\widehat{a}, \widehat{b})$ **, где**

$$\widehat{a} \;=\; \widehat{\theta}_n - \widehat{H}^{-1}\left(1-\frac{\alpha}{2}\right) = \widehat{\theta}_n - r^*_{1-\alpha/2} = 2\widehat{\theta}_n - \theta^*_{1-\alpha/2}$$

$$\widehat{b} \;=\; \widehat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) \qquad = \widehat{\theta}_n - r^*_{\alpha/2} \;= 2\widehat{\theta}_n - \theta^*_{\alpha/2}$$

13 **Таким образом получаем, что центральный доверительный интервал, построенный на основе бутстрепа, имеет вид**

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2},\ 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}\right)$$

Отметим, что $\mathbb{P}_F(T(F) \in C_n) \to 1-\alpha$ **при** $n \to \infty$

Метод 3: интервал на основе процентилей. Доверительный интервал, построенный на основе процентилей, имеет вид

$$C_n = \left(\theta^*_{\alpha/2},\ \theta^*_{1-\alpha/2}\right)$$

Действительно, допустим, что существует монотонное преобразование $U = m(T)$, для которого $U \sim N(\phi, c^2)$, где $\phi = m(\theta)$
14 **(предполагается только, что такое преобразование существует; конкретный вид преобразования может быть неизвестен). Пусть $U_b^* = m(\theta^*_{n,b})$. Обозначим через u^*_β - β выборочную квантиль последовательности чисел U_b^*, $b = 1, \ldots, B$. Так как монотонное преобразование сохраняет квантили, то $u^*_{\alpha/2} = m(\theta^*_{\alpha/2})$. Также $U \sim N(\phi, c^2)$, поскольку $\alpha/2$**

квантиль величины U равна $\phi - z_{\alpha/2}c$. Таким образом, $u^*_{\alpha/2} = \phi - z_{\alpha/2}c$ и, аналогично, $u^*_{1-\alpha/2} = \phi + z_{\alpha/2}c$.

Итак,

15
$$\begin{aligned}
\mathbb{P}(\theta^*_{\alpha/2} \le \theta \le \theta^*_{1-\alpha/2}) &= \mathbb{P}(m(\theta^*_{\alpha/2}) \le m(\theta) \le m(\theta^*_{1-\alpha/2})) \\
&= \mathbb{P}(u^*_{\alpha/2} \le \phi \le u^*_{1-\alpha/2}) \\
&= \mathbb{P}(U - cz_{\alpha/2} \le \phi \le U + cz_{\alpha/2}) \\
&= \mathbb{P}(-z_{\alpha/2} \le \frac{U-\phi}{c} \le z_{\alpha/2}) \\
&= 1 - \alpha
\end{aligned}$$

Точное «нормализующее» преобразование существует редко, однако существует много преобразований, которые позволяют делать распределения близкими к нормальному. Например, одно- и двухпараметрические семейства преобразований Бокса-Кокса:

$$y = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \log(x), \lambda = 0 \end{cases}$$

$$y = \begin{cases} \dfrac{(x + \lambda_2)^{\lambda_1} - 1}{\lambda_1}, \lambda_1 \neq 0 \\ \log(x + \lambda_2), \lambda_1 = 0 \end{cases}$$

<u>Пример</u> (данные о моментах времени между последовательными импульсами вдоль нервного волокна):

95% интервал для коэффициента ассиметрии

Нормальный интервал: $(1.44,\ 2.09)$

Центральный интервал: $(1.48,\ 2.11)$

Интервал на основе процентилей: $(1.42,\ 2.03)$

Все три типа интервалов имеют сравнимую точность

<u>Пример</u> (данные о значениях холестерина в плазме): построим доверительный интервал на основе бутстрепа для разности медиан

```
x1 <- первая выборка; x2 <- вторая выборка; n1 <- length(x1)
n2 <- length(x2); th.hat <- median(x2) - median(x1)
B <- 1000; Tboot <- вектор длины B
for(i in 1:B){
```

18

```
     xx1 <- выборка длины n1 с возвращением из x1
     xx2 <- выборка длины n2 с возвращением из x2
     Tboot[i] <- median(xx2) - median(xx1)}
se <- sqrt(variance(Tboot))
Normal     <- (th.hat - 2*se, th.hat + 2*se)
percentile <- (quantile(Tboot,.025), quantile(Tboot,.975))
pivotal    <- ( 2*th.hat-quantile(Tboot,.975),
                 2*th.hat-quantile(Tboot,.025) )
```

95% интервал для разности медиан имеет вид

Нормальный интервал: $(3.7,\ 33.3)$

Центральный интервал: $(5.0,\ 34.0)$

Интервал на основе процентилей: $(5.0,\ 33.3)$

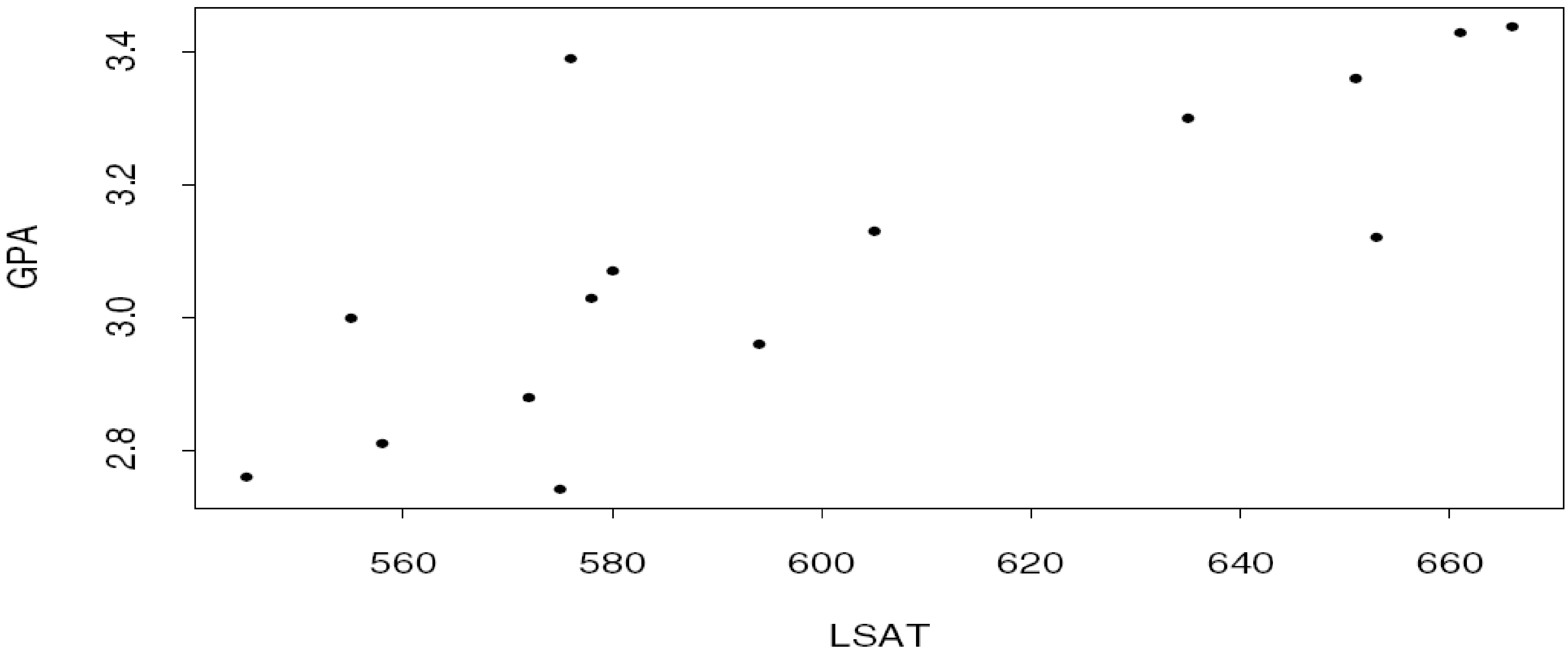
<u>Пример</u>: данные о LSAT **- Law School Admissible Test и** GPA **- Grade Point Average.**

19

LSAT	576	635	558	578	666	580	555	661
	651	605	653	575	545	572	594	
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
	3.36	3.13	3.12	2.74	2.76	2.88	3.96	

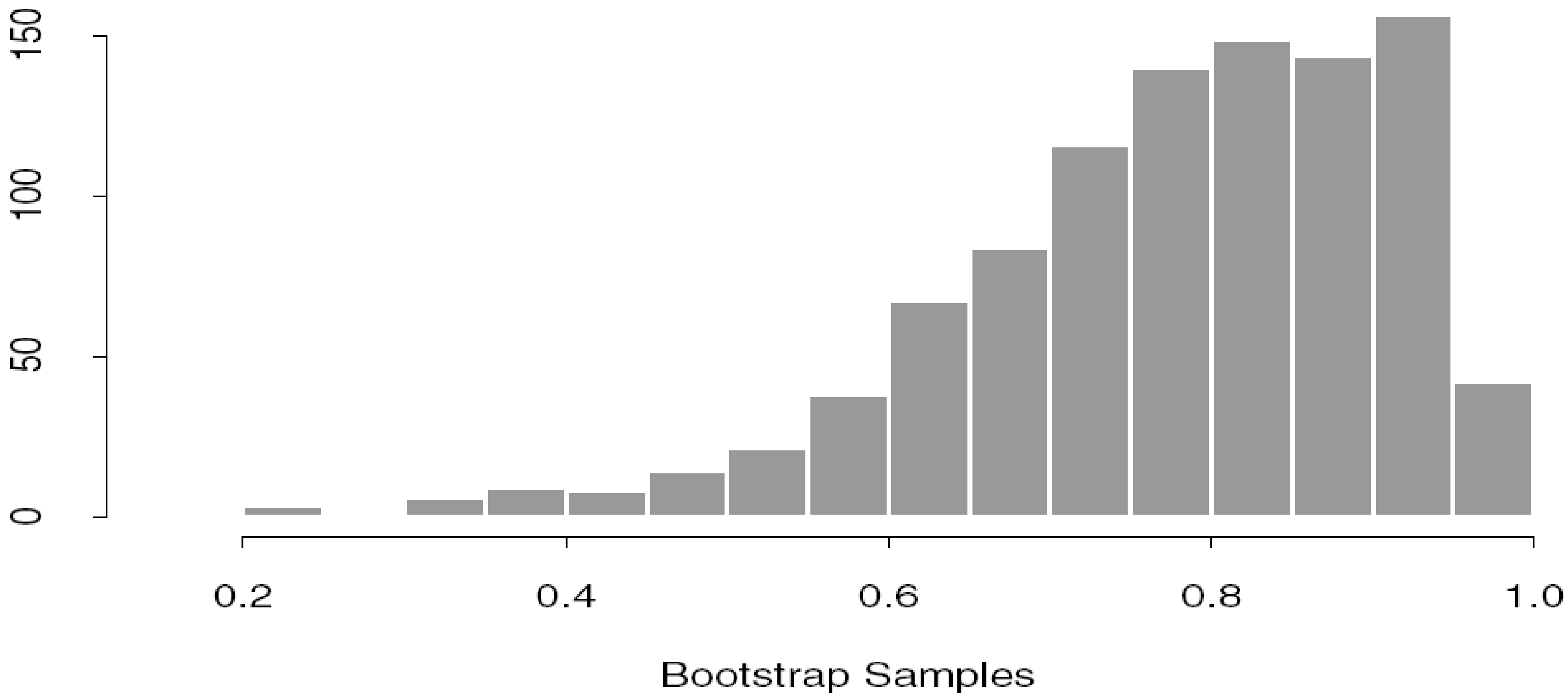
$X_i = (Y_i, Z_i)$ **, где** $Y_i = \mathrm{LSAT}_i$ **и** $Z_i = \mathrm{GPA}_i$

20



$$\theta = \frac{\int\int (y-\mu_Y)(z-\mu_Z)dF(y,z)}{\sqrt{\int (y-\mu_Y)^2 dF(y) \int (z-\mu_Z)^2 dF(z)}}$$

$$\widehat{\theta} = \frac{\sum_i (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sqrt{\sum_i (Y_i - \overline{Y})^2 \sum_i (Z_i - \overline{Z})^2}}, \ \widehat{\theta} = .776$$



Гистограмма значений $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ корреляций, подсчитанных по бустреп-выборкам

Гистограмма = аппроксимация распределения выборочных значений $\widehat{\theta}$

На основе бутстрепа с $B = 1000$ была получена оценка стандартной ошибки оценки $\widehat{\mathsf{se}} = .137$

95% интервал для коэффициента корреляции имеет вид

Нормальный интервал: $.78 \pm 2\widehat{\mathsf{se}} = (.51, 1.00)$

Интервал на основе процентилей: $(.46, .96)$

Причина сильного отличия – малый объем выборки

<u>Пример</u>: доказательство эквивалентности эффекта от медицинских препаратов

номер	плацебо	старое	новое	старое-плацебо	новое-старое
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

Обозначим $Z =$ **старое-плацебо и** $Y =$ **старое-новое. Считается, что эффект от препаратов совпадает, если** $|\theta| \leq .20$**, где**

$$\theta = \frac{\mathbb{E}_F(Y)}{\mathbb{E}_F(Z)}$$

24

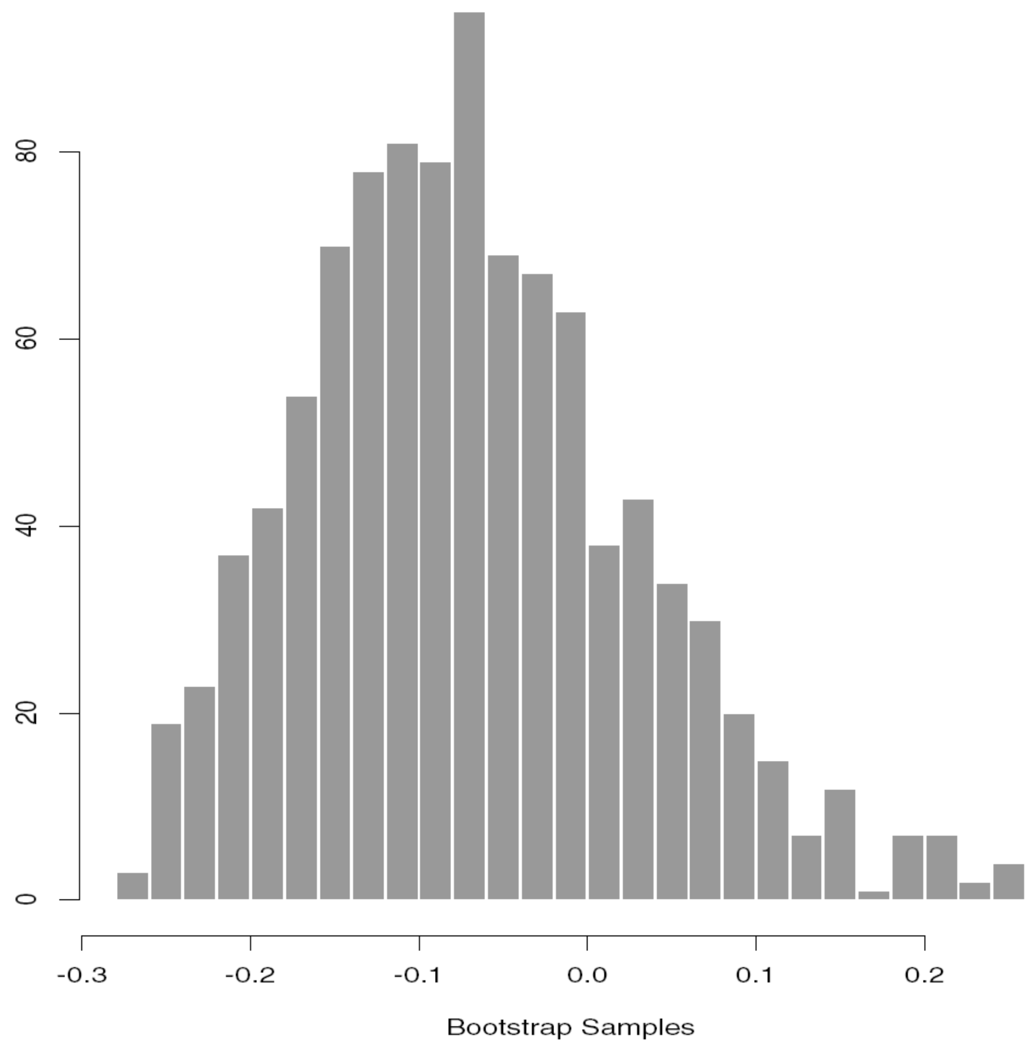
$$\widehat{\theta} = \frac{\overline{Y}}{\overline{Z}} = \frac{-452.3}{6342} = -0.0713$$

Оценка стандартной ошибки оценки параметра с помощью бутстрепа равна $\widehat{\mathsf{se}} = 0.105$

При $B = 1000$ **95% центральный интервал равен** $(\text{-}0.24, 0.15) \notin (\text{-}0.20, 0.20)$ **=> эквивалентность показана не была**

На рис. изображена гистограмма значений θ, полученных на основе бутстреп-выборки

25



d) Метод складного ножа

Пусть $T_n = T(X_1, \ldots, X_n)$ **- значение статистики, подсчитанное на основе простой выборки** $\boldsymbol{X_1, \ldots, X_n}$

26 **Обозначим через** $T_{(-i)}$ **значение статистики, подсчитанное на основе выборки** $\boldsymbol{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n}$

Пусть $\overline{T}_n = n^{-1} \sum_{i=1}^{n} T_{(-i)}$**. Оценка дисперсии** $\mathrm{var}(T_n)$ **по методу складного ножа равна**

$$v_{\text{jack}} = \frac{n-1}{n} \sum_{i=1}^{n} (T_{(-i)} - \overline{T}_n)^2$$

Оценка стандартной ошибки по методу складного ножа равна $\widehat{\mathsf{se}}_{jack} = \sqrt{v_{jack}}$**.**

Можно показать, что $v_{\mathrm{jack}}/\mathrm{var}(\mathrm{T_n}) \xrightarrow{\mathrm{P}} 1$ **при** $n \to \infty$**. Однако в отличие от бутстрепа оценка стандартной ошибки оценки квантили на основе метода складного ножа является несостоятельной**