Лекция 11.Бутстраппинг

Буре В.М., Грауэр Л.В.

ШАД

Санкт-Петербург, 2013

Идея бутстраппинга

Рассмотрим случайную величину ξ с неизвестной функцией распределения $F(x)$.
$(x_1, \ldots, x_n)$ — выборка объема n из генеральной совокупности ξ.

Истинное распределение данных можно хорошо приблизить эмпирическим.
Эмпирическая функция распределения

$$F^*(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[x_i \leq x]}$$

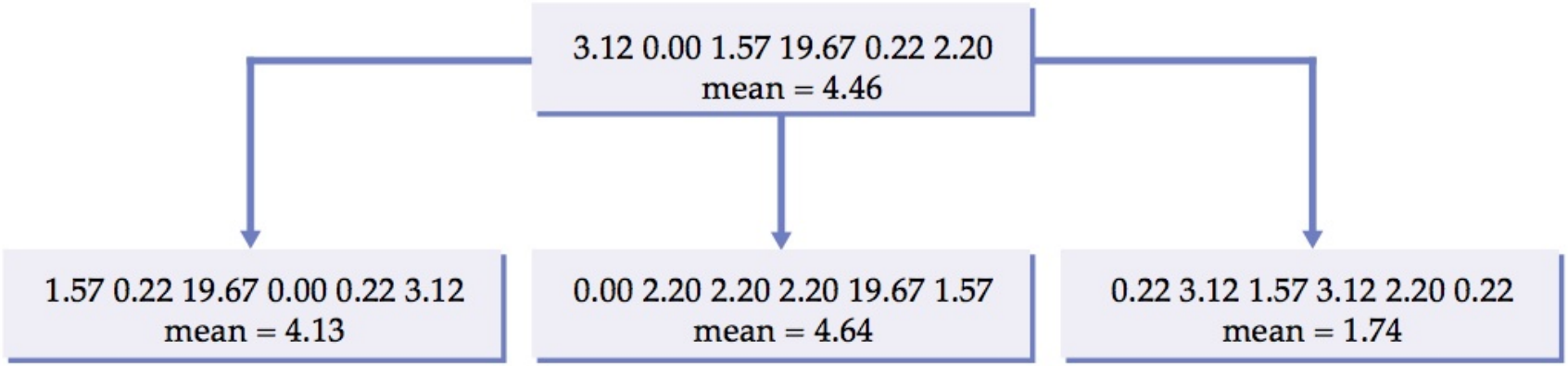
равномерно почти наверное стремится к $F(x)$ при $n \to \infty$ по лемме Гливенко-Кантелли, где $I_{[\cdot]}$ — индикатор-функция.

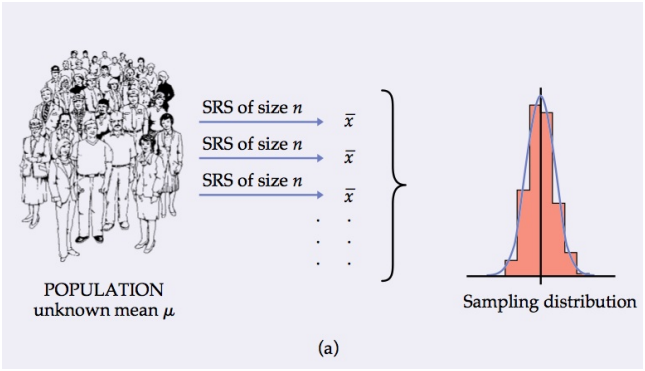
Идея бутстрепа по Б. Эфрону (1989):
трансформировать приближение для распределения данных в приближенное распределение статистик, используя метод статистических испытаний Монте-Карло, многократно извлекая повторные выборки из эмпирического распределения.

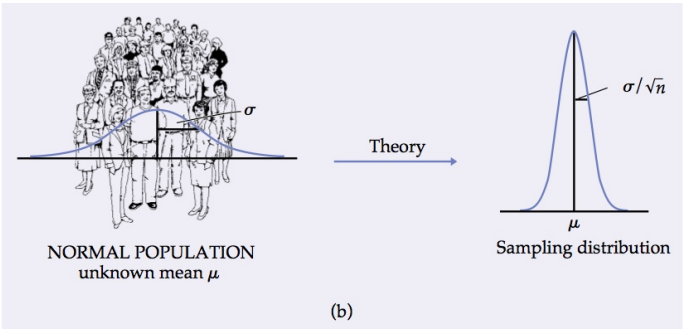
Пусть B — число бутстрапповских выборок. Для $b = 1, 2, \ldots, B$ построим бутстраповскую выборку

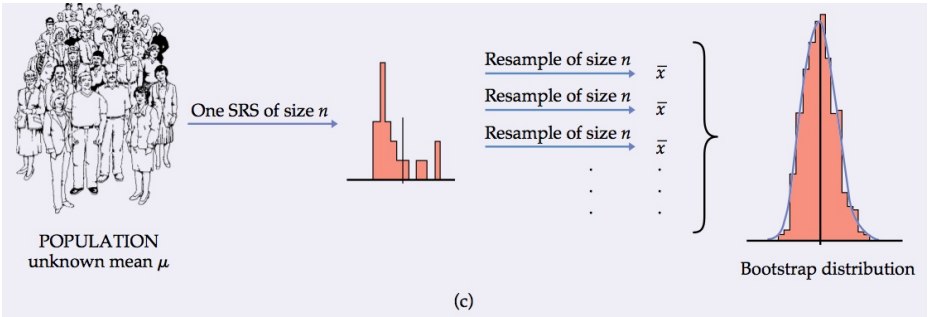
$$\{x_1^*, x_2^*, \ldots, x_n^*\}_b \tag{1}$$

А именно: берется конечная совокупность из n членов исходнои выборки $x_1, x_2, \ldots, x_n$, откуда на каждом шаге из n последовательных итерации с помощью датчика случаиных чисел, равномерно распределенных на интервале $[1, n]$, «вытягивается» произвольныи элемент x_k, которыи снова «возвращается» в исходную выборку.


3.12 0.00 1.57 19.67 0.22 2.20
mean = 4.46
1.57 0.22 19.67 0.00 0.22 3.12
mean = 4.13
0.00 2.20 2.20 2.20 19.67 1.57
mean = 4.64
0.22 3.12 1.57 3.12 2.20 0.22
mean = 1.74


SRS of size n
SRS of size n
SRS of size n
$\bar{x}$
$\bar{x}$
$\bar{x}$
POPULATION
unknown mean μ
Sampling distribution
(a)


σ
Theory
$\sigma/\sqrt{n}$
μ
NORMAL POPULATION
unknown mean μ
Sampling distribution
(b)


One SRS of size n
Resample of size n
Resample of size n
Resample of size n
$\bar{x}$
$\bar{x}$
$\bar{x}$
POPULATION
unknown mean μ
Bootstrap distribution
(c)

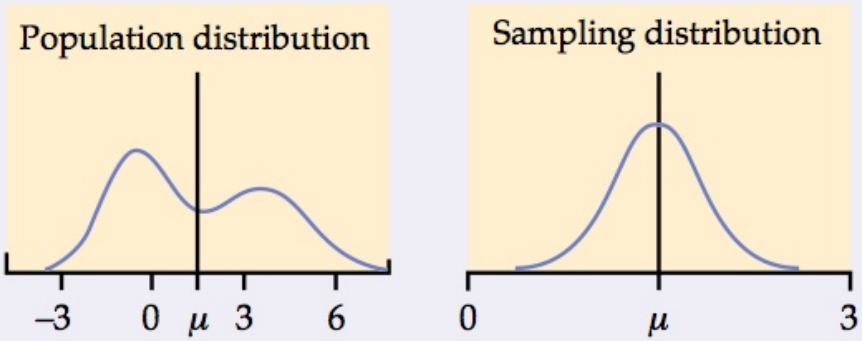
Пробутстрапим статистику $\hat{\theta} = \hat{\theta}(\{x_1, x_2, \ldots, x_n\})$.
Для каждой выборки (1) вычислим бутстраповскую статистику

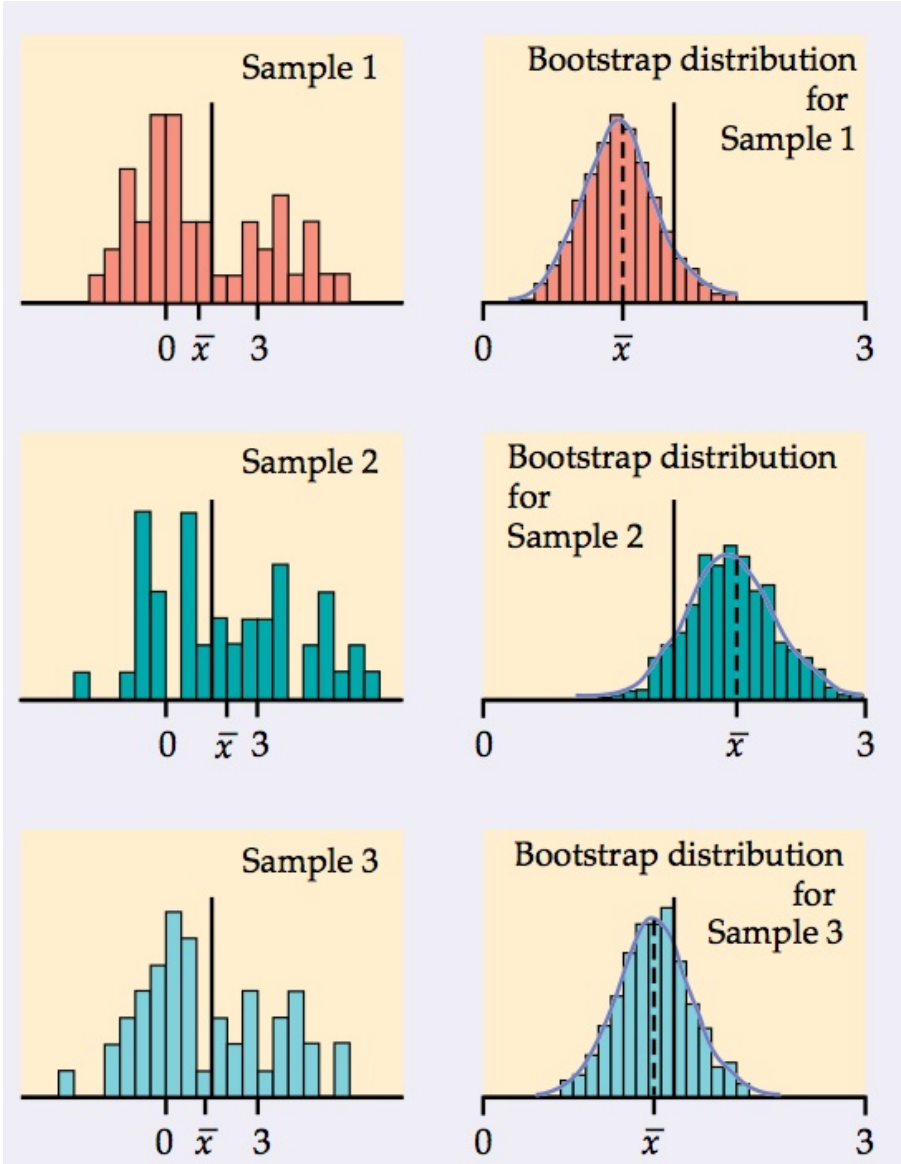
$$\hat{\theta}^*_b = \hat{\theta}(\{x^*_1, x^*_2, \ldots, x^*_n\}_b), \quad b = 1, \ldots, B. \tag{2}$$

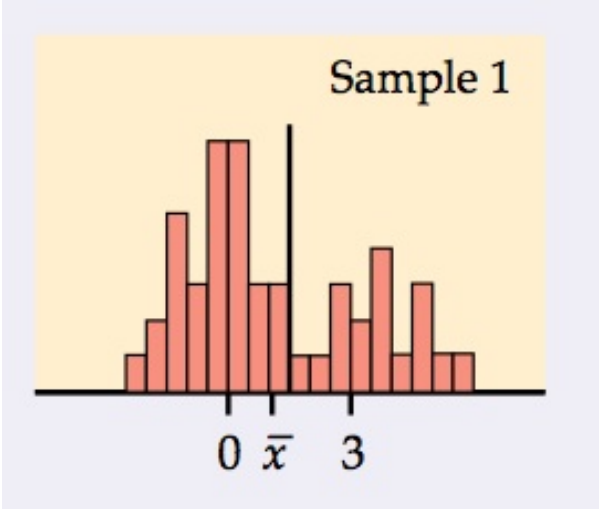
Полученный набор статистик $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$ с приписанием каждой веса $1/B$ составляет приближенное бутстраповское распределение статистики $\hat{\theta}$ с характеристиками

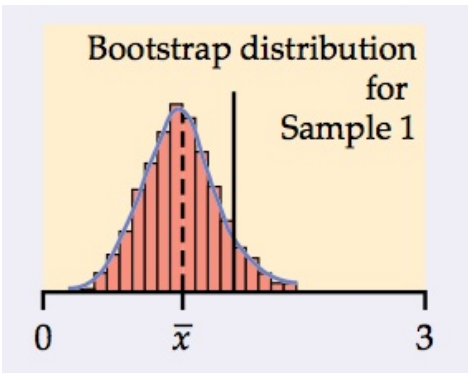
$$E(\hat{\theta}^*) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b$$

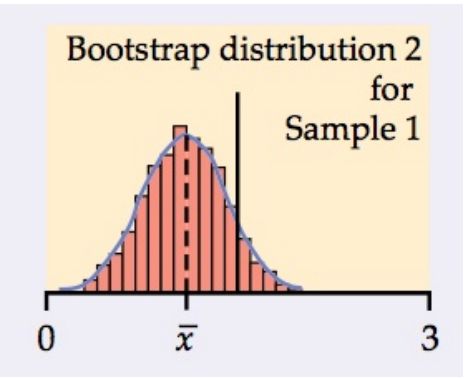
$$SE_{boot}(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - E(\hat{\theta}^*))^2}$$

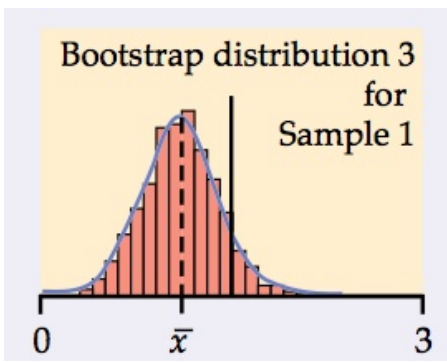

Population distribution
–3
0
μ
3
6
Sampling distribution
0
μ
3

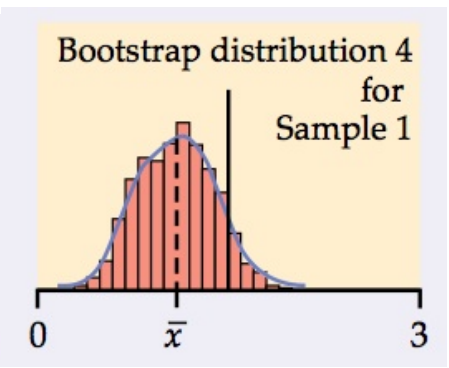

Sample 1
0
x̄
3
Bootstrap distribution for Sample 1
0
x̄
3
Sample 2
0
x̄
3
Bootstrap distribution for Sample 2
0
x̄
3
Sample 3
0
x̄
3
Bootstrap distribution for Sample 3
0
x̄
3


Sample 1
0
$\bar{x}$
3


Bootstrap distribution for Sample 1
0
$\bar{x}$
3


Bootstrap distribution 2 for Sample 1
0
$\bar{x}$
3


Bootstrap distribution 3 for Sample 1
0
$\bar{x}$
3


Bootstrap distribution 4 for Sample 1
0
$\bar{x}$
3

Рецентрирование

Когда речь идет о разнозностях или расстояних между выборочными и популяционными объектами, необходимо рецентрирование
Правильным бутстраповским аналогом разности $\hat{\theta} - \theta$ является $\hat{\theta}^* - \hat{\theta}$.

Пример 1

Бутстрапируя t-статистику при нулевой гипотезе $H_0 : \theta = 0$

$$t = \frac{\hat{\theta}}{se(\hat{\theta})},$$

следует использовать бутстрап аналог

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{se_{boot}(\hat{\theta}^*)}$$

Пример 2

Бутстраповским аналогом статистики отношения правдобподобия

$$LR = 2\left(\max_{\theta} \ln L(\theta) - \max_{\theta:g(\theta)=0} \ln L(\theta)\right)$$

будет

$$LR^* = 2\left(\max_{\theta} \ln L(\theta) - \max_{\theta:g(\theta)=\hat{\theta}} \ln L(\theta)\right)$$

Рецентрировать $se_{boot}(\hat{\theta}^*)$ нет необходимости так, как это нормирующий множитель.

Бутстраповская корректировка смещения

Бутстрап позволяет скорректировать смещение, связанное с конечностью выборки.
Пусть $\hat{\theta}$ — состоятельная смещенная оценка неизвестного параметра θ ($E(\hat{\theta}) \neq \theta$).
Смещение оценки $\hat{\theta}$ равно $C(\hat{\theta}) = E(\hat{\theta}) - \theta$.
Вычислив бутстраповский аналог этого смещения

$$C^*(\hat{\theta}) = E(\hat{\theta}^*) - \hat{\theta},$$

можно скорректировать исходную статистику

$$\hat{\theta}_{BC} = \hat{\theta} - C^*(\hat{\theta}) = 2\hat{\theta} - E(\hat{\theta}^*) = 2\hat{\theta} - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

Бутстрап не способен справиться с асимптотическим смещением.

Пример 3

Пусть μ — неизвестное математическое ожидание случайной величины ξ. Рассмотрим статистику μ^2. В качестве оценки μ возьмем $\bar{x}_n$, в качестве оценки μ^2 — $\bar{x}_n^2$.

Бутстраповский аналог смещения $C^*(\bar{x}_n) = E^*(\bar{x}_n^*) - \bar{x}_n = 0$, так как $E^*(\bar{x}_n^*) = \bar{x}_n$.

Смещение $\bar{x}_n^2$ равно

$$C(\bar{x}_n^2) = E(\bar{x}_n^2) - \mu^2 = D(\bar{x}_n) = \frac{1}{n} D\xi$$

Бутстраповский аналог смещения в данном случае есть

$$C^*(\bar{x}_n^2) = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}_n^2\right).$$

Тогда оценкой μ^2, скорректированной на смещение, будет

$$\bar{x}_n^2 - C^*(\bar{x}_n^2) = \frac{n+1}{n}\bar{x}_n^2 - \frac{1}{n^2}\sum_{i=1}^{n} x_i.$$

Доверительные интервалы

Пусть $\hat{\theta}$ — состоятельная оценка параметра θ. Построим доверительные интервалы параметра θ.

Эфронов доверительный интервал

Используем бутстраповское распределение статистики $\hat{\theta}$ (2) для построения доверительного интервала.
В качестве квантилей $q^*_{\alpha_1}$, $q^*_{1-\alpha_2}$ возьмем порядковые статистики $\hat{\theta}^*_{([B\alpha_1])}$ и $\hat{\theta}^*_{([B(1-\alpha_2)]+1)}$.
Эфронов доверительный интервал с покрытием $1-\alpha_1-\alpha_2$ равен

$$CI_{efr} = \left(\hat{\theta}^*_{([B\alpha_1])}, \hat{\theta}^*_{([B(1-\alpha_2)]+1)}\right).$$

Так как не используется рецентрирование, в данном случае смещение исходной выборки лишь усиливается. Применим для симметричных распределений или в случае малых смещений.

Квантильный доверительный интервал (рецентрированный)

Пробутстрапим статистику $\hat{\theta} - \theta$, бутстраповский аналог которой $\hat{\theta}^* - \hat{\theta}$.

В качестве квантилей $q^{*\%}_{\alpha_1}$, $q^{*\%}_{1-\alpha_2}$ возьмем соответствующие порядковые статистики порядков $[B\alpha_1]$ и $[B(1-\alpha_2)]+1$ и построим квантильный доверительный интервал с покрытием $1-\alpha_1-\alpha_2$

$$CI_{\%} = \left(\hat{\theta} - q^{*\%}_{1-\alpha_2}, \hat{\theta} - q^{*\%}_{\alpha_1}\right).$$

T-квантильный доверительный интервал

Пробутстрапим статистику $\dfrac{\hat{\theta} - \theta}{se(\hat{\theta})}$, бутстраповский аналог которой $\dfrac{\hat{\theta}^* - \hat{\theta}}{se^*(\hat{\theta})}$. Получим бутстраповские квантили $q^{*\%t}_{\alpha_1}$, $q^{*\%t}_{1-\alpha_2}$ и построим t-квантильный доверительный интервал с покрытием $1-\alpha_1-\alpha_2$

$$CI_{\%t} = \left(\hat{\theta} - se(\hat{\theta})q^{*\%t}_{1-\alpha_2}, \hat{\theta} - se(\hat{\theta})q^{*\%t}_{\alpha_1}\right).$$

Если $\alpha_1 = \alpha_2 = \alpha/2$, то можно бутстрапировать не ценитрированную оценку или t-статитистику, а их модули $|\hat{\theta} - \theta|$ и $|\hat{\theta} - \theta|/se(\hat{\theta})$, соответственно.

Симметричный квантильный доверительный интервал

$$CI_{|\%|} = \left(\hat{\theta} - q^{*|\%|}_{1-\alpha/2}, \hat{\theta} + q^{*|\%|}_{1-\alpha/2}\right).$$

Симметричный t-квантильный доверительный интервал

$$CI_{|\%|t} = \left(\hat{\theta} - se(\hat{\theta})q^{*|\%|t}_{1-\alpha/2}, \hat{\theta} + se(\hat{\theta})q^{*|\%|t}_{1-\alpha/2}\right).$$

Асимптотическое рафинирование

Определение 1

Пусть $\hat{\varphi}$ — некоторая статисктика, инстинное распределение которой $F_{\hat{\varphi}}(x)$, $F^*_{\hat{\varphi}}(x)$ — бутстраповское распределение этой статистики. Говорят, что с помощью бутстрапа достигается асимптотическое рафинирование, если ошибка аппроксимации истинного распределения $F_{\hat{\varphi}}(x)$ бутстраповским $F^*_{\hat{\varphi}}(x)$ большего порядка малости, чем ошибка аппроксимации асимптотическим распределением при стремлении объема выборки к бесконечности.

Рассмотрим асимптотически нормальную статистику

$$\hat{\varphi} = \frac{\hat{\theta} - \theta}{se(\hat{\theta})} \xrightarrow{d} N(0,1).$$

Разложения Эджворта истинного и бутстраповского распределений вокруг асимптотического выглядят следующим образом

$$F_{\hat{\varphi}}(x) = \Phi(x) + \frac{h_1(x,F)}{\sqrt{n}} + \frac{h_2(x,F)}{n} + o\left(\frac{1}{n\sqrt{n}}\right),$$

$$F^*_{\hat{\varphi}}(x) = \Phi(x) + \frac{h_1(x,F*)}{\sqrt{n}} + \frac{h_2(x,F*)}{n} + o\left(\frac{1}{n\sqrt{n}}\right),$$

где $h_1(x,F)$ — четная по x, непрерывная по F функция, $h_2(x,F)$ — нечетная по x, непрерывная по F функция

Ошибки аппроксимации точного распределения асимптотическим и бутстраповским, соответственно, равны

$$\Phi(x) - F_{\hat{\varphi}}(x) = \frac{h_1(x, F)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{\sqrt{n}}\right),$$

$$F^*_{\hat{\varphi}}(x) - F_{\hat{\varphi}}(x) = \frac{h_1(x, F^*) - h_1(x, F)}{\sqrt{n}} + o\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right),$$

Сравнивая порядки ошибок аппроксимации, можно сделать вывод, что в данном случае использование бутстрапа приводит к асимптотическому рафинированию.

На практике это означает, что в достаточно больших выборках ошибка бутстраповского приближения, намного меньше, чем ошибка асимптотического приближения.

Аналогично можно показать, что
если статистика

$$\hat{\varphi} = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, D_\theta),$$

то бутстрап не приводит к асимптотическому рафинированию

$$\Phi(x, D_\theta) - F_{\hat{\varphi}}(x) = O\left(\frac{1}{\sqrt{n}}\right), \quad F^*_{\hat{\varphi}}(x) - F_{\hat{\varphi}}(x) = O\left(\frac{1}{\sqrt{n}}\right);$$

если статистика

$$\hat{\varphi} = \frac{|\hat{\theta} - \theta|}{se(\hat{\theta})} \xrightarrow{d} |N(0, 1)|,$$

то бутстрап приводит к асимптотическому рафинированию, причем ошибка большего порядка малости

$$\Phi(x) - F_{\hat{\varphi}}(x) = O\left(\frac{1}{n}\right), \quad F^*_{\hat{\varphi}}(x) - F_{\hat{\varphi}}(x) = O\left(\frac{1}{n\sqrt{n}}\right).$$

Проверка статистических гипотез

$H_0 : \theta = \theta_0$

Для проверки гипотезы $H_0 : \theta = \theta_0$ с помощью бутстрапирования в случае

- двусторонней альтернативной гипотезы $H_1 : \theta \neq \theta_0$, пробустрапировав модуль t-статистики $|\hat{\theta} - \theta|/se(\hat{\theta})$, получим бутстраповское распределение $|\hat{\theta}^* - \hat{\theta}|/se^*(\hat{\theta})$ и ее квантиль $q^{*|\%|t}_{1-\alpha/2}$.
  Гипотеза H_0 отвергается на уровне значимости α, если $|\hat{\theta} - \theta_0|/se(\hat{\theta}) > q^{*|\%|t}_{1-\alpha/2}$.
- односторонней альтернативной гипотезы $H_1 : \theta > \theta_0$, пробустрапировав t-статистику $(\hat{\theta} - \theta)/se(\hat{\theta})$, получим бутстраповское распределение $(\hat{\theta}^* - \hat{\theta})/se^*(\hat{\theta})$ и ее квантиль $q^{*\%t}_{1-\alpha}$.
  Гипотеза H_0 отвергается на уровне значимости α, если $(\hat{\theta} - \theta_0)/se(\hat{\theta}) > q^{*\%t}_{1-\alpha}$.

Проверка статистических гипотез

$H_0 : \theta_1 = \theta_2$

Рассмотрим выборки $(x_1, \ldots, x_n)$ и $(y_1, \ldots, y_m)$ из генеральных совокупностей ξ и η. Проверим гипотезу о равенстве параметров двух распределений $H_0 : \theta_1 = \theta_2$ при двусторонней альтернативной гипотезе $H_1 : \theta_1 \neq \theta_2$.

Бутстрапирование

Пробутстрапируем статистики $\hat{\theta}_1(\{x_1, \ldots, x_n\})$ и $\hat{\theta}_2(\{y_1, \ldots, y_m\})$
Для каждого $b = 1, \ldots, B$ получим значение бустрап статистики

$$\hat{\theta}_1^*(\{x_1^*, \ldots, x_n^*\}_b) - \hat{\theta}_2^*(\{y_1^*, \ldots, y_m^*\}_b).$$

Построим распределение бутстрап статистики $\hat{\theta}_1^* - \hat{\theta}_2^*$ и ее квантили $q^*_{\alpha/2}$ и $q^*_{1-\alpha/2}$.
Гипотеза H_0 отвергается на уровне значимости α, если $0 > q^*_{1-\alpha/2}$ либо $0 < q^*_{\alpha/2}$.

Пермутейшн

Сформируем единую выборку

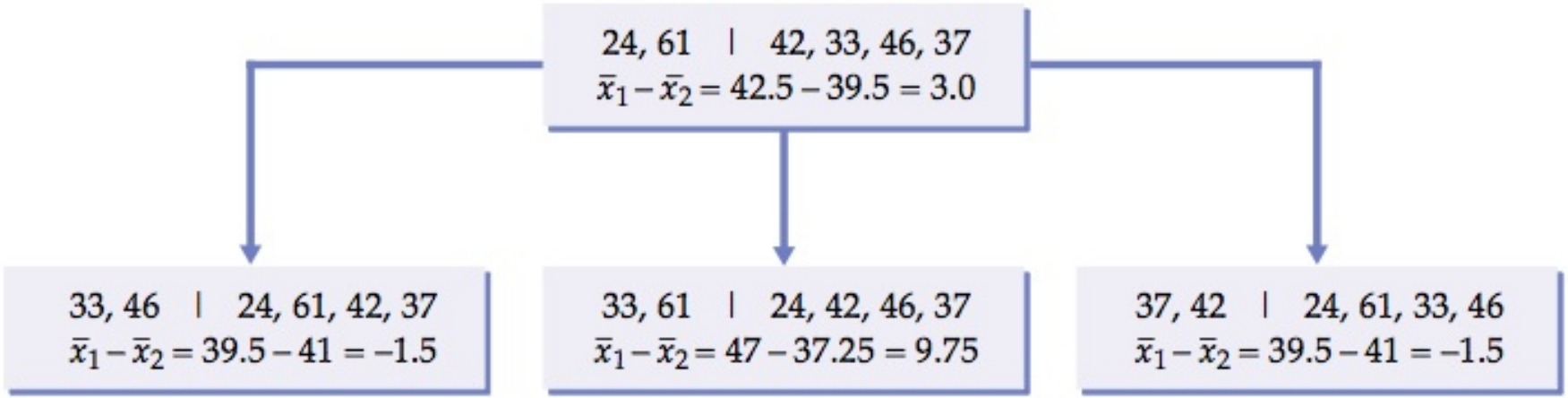
$$(x_1, \ldots, x_n, y_1, \ldots, y_m). \tag{3}$$

Случайным образом без возвращения извлечем n наблюдений из выборки (3) $x_1^*, \ldots, x_n^*$, остальные элементы обобщенной выборки обозначим $y_1^*, \ldots, y_m^*$. Получим значение статистики

$$\hat{\theta}_1^*(\{x_1^*, \ldots, x_n^*\}_b) - \hat{\theta}_2^*(\{y_1^*, \ldots, y_m^*\}_b).$$

Повторим данный ресэмплинг B раз.
Построим распределение пермутейшн статистики $\hat{\theta}_1^* - \hat{\theta}_2^*$ и ее квантили $q_{\alpha/2}^*$ и $q_{1-\alpha/2}^*$.

Гипотеза H_0 отвергается на уровне значимости α, если $\hat{\theta}_1 - \hat{\theta}_2 > q_{1-\alpha/2}^*$ либо $\hat{\theta}_1 - \hat{\theta}_2 < q_{\alpha/2}^*$.


24, 61 | 42, 33, 46, 37
$\bar{x}_1 - \bar{x}_2 = 42.5 - 39.5 = 3.0$
33, 46 | 24, 61, 42, 37
$\bar{x}_1 - \bar{x}_2 = 39.5 - 41 = -1.5$
33, 61 | 24, 42, 46, 37
$\bar{x}_1 - \bar{x}_2 = 47 - 37.25 = 9.75$
37, 42 | 24, 61, 33, 46
$\bar{x}_1 - \bar{x}_2 = 39.5 - 41 = -1.5$

Проверка статистических гипотез

$H_0 : m_1 = m_2$, связные выборки

Рассмотрим выборку $(x_1, y_1), \ldots, (x_n, y_n)$ из двумерной генеральной совокупности (ξ, η). Проверим гипотезу о равенстве математических ожиданий двух распределений $H_0 : m_1 = m_2$ при двусторонней альтернативной гипотезе $H_1 : m_1 \neq m_2$.

Бутстрапирование

Перейдем к выборке $(d_1, \ldots, d_n)$, где $d_i = x_i - y_i$ и нулевой гипотезе $H_0 : [m_1 - m_2 =]\theta = 0$.

Построим на основе выборки $(d_1, \ldots, d_n)$ t-статистику и пробустрапируем модуль t-статистики $|\hat{\theta}|/se(\hat{\theta})$, получим бутстраповское распределение $|\hat{\theta}^* - \hat{\theta}|/se^*(\hat{\theta})$ и ее квантиль $q^{*|\%|t}_{1-\alpha/2}$.

Гипотеза H_0 отвергается на уровне значимости α, если $|\hat{\theta}|/se(\hat{\theta}) > q^{*|\%|t}_{1-\alpha/2}$.

Пермутейшн

Для каждого парного наблюдения (x_i, y_i) случайным образом переставим местами x_i и y_i внутри пары.

Получим ресэмплированную выборку $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$.

Повторим действия B раз.

Построим распределение пермутейшн статистики $\hat{\theta}_1^* - \hat{\theta}_2^*$ и ее квантили $q^*_{\alpha/2}$ и $q^*_{1-\alpha/2}$.

Гипотеза H_0 отвергается на уровне значимости α, если $\hat{\theta}_1 - \hat{\theta}_2 > q^*_{1-\alpha/2}$ либо $\hat{\theta}_1 - \hat{\theta}_2 < q^*_{\alpha/2}$.

Ресэмплинг в регрессиях

Рассмотрим линейную регрессионную модель

$$y_i = X_i\beta + \varepsilon_i.$$

Существует два основных вида ресэмплинга в регрессионных моделях

- ресэмплинг парных наблюдений $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, при котором бутсрапированные данные $(y_1, x_1)^*, (y_2, x_2)^*, \ldots, (y_n, x_n)^*$ извлекаются независимо с равными вероятностями $1/n$, и
- ресэмплинг остатков:
  Построив регрессионную модель $X\hat{\beta}$, извлекаем случайным образом ϵ_i^* из набора центрированных стандартизированных остатков $e_1, \ldots, e_n$ и устанавливаем y_i^* равными

$$y_i^* = \hat{\beta}x_i + \epsilon_i^*, \quad i = 1, \ldots, n.$$

Литература

- Эфрон Б.Нетрадиционные методы многомерного статистического анализа. М.:Финансы и статистика, 1988
- Efron B., Tibshirani R.J.An introduction to the bootstrap. N.Y.: Chapman & Hall, 1993
- Moore D.Bootstrap methods and permutation tests The practice of business statistics.
- Эконометрический ликбез: бутстрап. Квантиль, 2007, №3, стр. 1-66
- Chibara L., Hesterberg T.Mathematical statistics with resampling and R. Wiley