Теория и практика применения бутстрэпа

2024

Содержание

1. Введение

1.1. История метода бутстрэпа

Словом «бутстрэп» обозначают некоторое семейство или множество статистических методов или алгоритмов, предназначенных для нахождения оценок параметров распределения и доверительных интервалов. Эти методы основаны на генерации большого числа выборок из исходной, имеющейся в распоряжении исследователя выборки с возвращениями с использованием датчика (псевдо-) случайных чисел. Таким образом, из-за объемности вычислений реализация бутстрэпа предполагается на компьютере.

Одной из первых работ, в которой была изложена суть метода бутстрэпа, стала статья Брэдли Эфрона «Bootstrap methods: another look at the jackknife» (1979 г.), вдохновленная более ранней работой о методе Jackknife. Позже разрабатывалась теория для нахождения более точных оценок дисперсий параметров, повышения устойчивости метода на выборках небольших размеров, с включением байесовского подхода и т.д.

1.2. Классическая теория построения доверительных интервалов

Пусть имеется выборка $\mathbf{X} = (X_1, ..., X_n)$ и известно, что она была сгенерирована из нормального распределения $\mathcal{N}(\mu, \sigma^2)$ с неизвестным математическим ожиданием μ и известной дисперсией σ^2. В таком случае известным результатом является построение следующего доверительного интервала:

$$\mathbb{P}\left(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

где z_α — квантиль стандартной нормальной случайной величины порядка α.

Кроме того, похожий результат может быть получен в случае неизвестного параметра σ^2. В таком случае используется несмещенная оценка дисперсии

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

и получается, что стандартизированная статистика имеет распределение Стьюдента

$$n \cdot \frac{\bar{X} - \mu}{\hat{\sigma}} \sim t_{n-1}$$

Тогда точный доверительный интервал приобретает вид

$$\mathbb{P}\left(\bar{X} - t_{1-\frac{\alpha}{2};n-1}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2};n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

При большом числе наблюдений стандартизированная случайная величина $n \cdot \frac{\bar{X}-\mu}{\hat{\sigma}}$, согласно центральной предельной теореме, по распределению сходится к стандартному нормальному распределению вне зависимости от вида закона распределения X (при условии существования конечного математического ожидания μ и конечной дисперсии σ^2), т. е.

$$n \cdot \frac{\bar{X} - \mu}{\hat{\sigma}} \xrightarrow[n \to \infty]{dist} \mathcal{N}(0, 1)$$

Поэтому может быть получен следующий асимптотический доверительный интервал:

$$\mathbb{P}\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Итак, было показано, что в случае построения доверительного интервала для параметра среднего в случае известного закона распределения имеется хорошо разработанная теория и могут быть получены точные или асимптотические формулы. Кроме того, такие результаты широко известны в случае построения доверительных интервалов для разницы математических ожиданий, дисперсий и отношения дисперсий.

Однако возникает вопрос, какую схему действий предпринять в случае, если закон распределения, порождающий данные, неизвестен, а также если стоит задача построить доверительный интервал для такой не менее важной характеристики, как, например, медиана распределения, асимптотический закон распределения которой не является широко известным.

По-прежнему можно построить $\hat{\theta}$ как точечную оценку неизвестного параметра θ. Но теоретических знаний о распределении $\hat{\theta}$ у нас может не быть, либо же эта теория недоступна исследователю. В таком случае можно использовать бутстрэп для построения доверительного интервала. Однако прежде стоит разобраться с тем, каким образом можно создавать новые выборки на основе исходной.

1.3. Построение псевдовыборок

Существует два распространенных метода построения случайной выборки.

- Выборка без возвращения (without replacement, simple random sampling):

Предположим, мы поочереди берем 10 карт наугад из колоды из 52 карт, не возвращая ни одну из карт обратно в колоду между взятиями. Это называется выборкой без возвращению или простой случайной выборкой. При таком методе в нашей выборке из 10 карт не будет дубликатов карт. То есть, в данном подходе, если мы захотим создать выборку, размер которой совпадает с исходной, будет получена сама исходная выборка, что ограничивает возможности в исследовании распределения данных.

- Выборка с возвращением (with replacement):

Теперь предположим, что мы берем 10 карт наугад из колоды, но после каждого взятия кладем карту обратно в колоду и перемешиваем карты. Это называется выборкой с возвращением. При использовании этого метода выборка из 10 карт может иметь дубликаты. Возможно и такое, что была вытянута шестерка червей все 10 раз. Такая процедура позволяет создавать отлиные от исходной новые выборки, размер которых при этом совпадает с размером исходной выборки.

1.4. Суть метода бустрэпа

На основе имеющейся выборки может быть рассчитана выборочная или эмпирическая функция распределения $\hat{F}(x)$, которая для каждого x показывает долю наблюдений в выборке, не превосходящих x:

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}\{X_i \leqslant x\}}{n}$$

Известно, что эта функция обладает рядом «хороших» свойств: состоятельность, эффективность, асимптотическая нормальность, поэтому является «хорошей» оценкой для истинной функции распределения $F(x)$, порождающей данные.

Итак, основной идеей бутстрэпа в данном случае будет генерация B бутстрэпированных выборок $\mathbf{X}_1^*, ..., \mathbf{X}_B^*$ с повторениями из исходной выборки $\mathbf{X}$ и подсчет оценки неизвестного параметра распределения $\{\hat{\theta}_j^*\}_{b=1}^{B}$ для получения распределения $\hat{\theta}$.

2. Теория основных версий бутстрэпа

2.1. Перцентильный бутстрэп

Одной из основных и наиболее простых техник бустрэпа является *перцентильный бутстрэп (percentile bootstrap)*. Его алгоритм следующий:

1. Из исходной выборки $\mathbf{X}$ генерируется B бутстрэпированных выборок $\{\mathbf{X}_b^*\}_{b=1}^{B}$;
2. Для каждой бутстрэпированный выборки оценивается параметр $\{\hat{\theta}_b^*\}_{b=1}^{B}$, затем оцененные параметры сортируются по возрастанию: $\hat{\theta}^*_{(1)} \leqslant \hat{\theta}^*_{(2)} \leqslant \ldots \leqslant \hat{\theta}^*_{(B)}$, и из упорядоченных оценок составляется вектор $\hat{\theta}^* = \left(\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}\right)$;
3. Оцененным $(1-\alpha)$-процентным доверительным интервалом истинного параметра θ будет отрезок от $\frac{\alpha}{2}$ квантиля до $1-\frac{\alpha}{2}$ квантиля вектора $\hat{\theta}^*$: $CI_E = \left[\hat{\theta}^*_{\frac{\alpha}{2}}; \hat{\theta}^*_{1-\frac{\alpha}{2}}\right]$.

Такой доверительный интервал называется *эфроновским* (E — Efron). Насколько же точен данный метод построения доверительных интервалов? Введем обозначения: $CL_\alpha = 1-\alpha$ — теоретический уровень значимости доверительного интервала, $\widehat{CL}_{\alpha,n}$ — фактический (выборочный) уровень значимости доверительного интервала. Разность между теоретическим и фактическим уровнем значимости может быть оценена следующим образом: $|CL_\alpha - \widehat{CL}_{\alpha,n}| = O\left(\frac{1}{\sqrt{n}}\right)$.

Оцененный таким образом доверительный интервал в среднем будет занижать вероятность накрытия.

Кроме того, если выборочное распределение было асимметричным, это также приведет к асимметрии доверительного интервала, полученного с помощью перцентильного бутстрэпа.

2.2. Обратный перцентильнй бутстрэп

Модификацией первого метода можно считать обратный перцентильный бутстрэп (Reverse Percentile Bootstrap). Его аглоритм:

1. Из исходной выборки $\mathbf{X}$ генерируется B бутстрэпированных выборок $\{\mathbf{X}_b^*\}_{b=1}^B$;
2. Для каждой бутстрэпированный выборки оценивается параметр $\{\hat{\theta}_b^*\}_{b=1}^B$, далее рассчитывается отклонение полученной оценки от выборочной оценки $d_b^* = \hat{\theta}_b^* - \hat{\theta}$, полученные разницы сортируются по возрастанию: $d_{(1)}^* \leqslant d_{(2)}^* \leqslant \ldots \leqslant d_{(B)}^*$, и собирается вектор $d^* = \left(d_{(1)}^*, d_{(2)}^*, \ldots, d_{(B)}^*\right)$;
3. Оцененным $(1-\alpha)$-процентным доверительным интервалом разницы $d = \hat{\theta} - \theta$ будет отрезок от квантиля $\frac{\alpha}{2}$ до квантиля $1 - \frac{\alpha}{2}$ вектора d^*: $\widehat{CI}_d = \left[d_{\frac{\alpha}{2}}^*; d_{1-\frac{\alpha}{2}}^*\right]$. Тогда для получения оценки доверительного интервала для параметра θ необходимо выполнить ряд преобразований:

$$d_{\frac{\alpha}{2}}^* \leqslant d \leqslant d_{1-\frac{\alpha}{2}}^* \Leftrightarrow d_{\frac{\alpha}{2}}^* \leqslant \hat{\theta} - \theta \leqslant d_{1-\frac{\alpha}{2}}^* \Leftrightarrow \hat{\theta} - d_{1-\frac{\alpha}{2}}^* \leqslant \theta \leqslant \hat{\theta} - d_{\frac{\alpha}{2}}^*$$

   Таким образом, $CI_H = \left[\hat{\theta} - d_{1-\frac{\alpha}{2}}^*; \hat{\theta} - d_{\frac{\alpha}{2}}^*\right]$.

Данный вид доверительного интервала называется *холловским* (H — Hall). Скорость сходимости обратного перцентильного бутстрэпа совпадает со скоростью сходимости обычного перцентильного бутстрэпа: $|CL_\alpha - \widehat{CL}_{\alpha,n}| = O\left(\frac{1}{\sqrt{n}}\right)$, однако холловский доверительный интервал дает более корректный результат для асимметричных распределений.

2.3. Стьюдентизированный бутстрэп (бутстрэп t-статистики)

Далее рассмотрим метод, который позволит рассчитывать стандартную ошибку оценки, и учитывать ее влияние на распределение. Идея этого подхода заключается в том, чтобы приблизить величину $\frac{\hat{\theta}-\theta}{se\left(\hat{\theta}\right)}$ величиной $\frac{\hat{\theta}^*-\hat{\theta}}{se\left(\hat{\theta}^*\right)}$. Алгоритм стьюдентизированного бутстрэпа:

1. Из исходной выборки $\mathbf{X}$ генерируется B бутстрэпированных выборок $\{\mathbf{X}_b^*\}_{b=1}^B$;
2. Для каждой бутстрэпированный выборки оценивается параметр $\{\hat{\theta}_b^*\}_{b=1}^B$, далее считается стандартная ошибка полученной оценки. В случае построения доверительного интервала для математического ожидания можно воспользоваться формулой

$$\widehat{se}\left(\hat{\theta}_b^*\right) = \sqrt{\frac{\sum_{i=1}^n (X_{i,b}^* - \bar{X}_b^*)^2}{n \cdot (n-1)}} = \frac{1}{\sqrt{n}} \sigma_{X_b^*}$$

   где $\bar{X}_b^*$ — среднее по бутстрэпированной выборке b, $\sigma_{X_b^*}$ — среднеквадратичное отклонение бутстрэпированной выборки b. Далее вычисляется B бутстрэпированных t-статистик:

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\widehat{se}\left(\hat{\theta}_b^*\right)}$$

полученные t-статистики сортируются по возрастанию: $t_{(1)}^* \leqslant t_{(2)}^* \leqslant \ldots \leqslant t_{(B)}^*$, и собирается вектор $t^* = \left(t_{(1)}^*, t_{(2)}^*, \ldots, t_{(B)}^*\right)$;

3. Оцененным $(1-\alpha)$-процентным доверительным интервалом стандартизированного параметра $\theta^{st} = \frac{\hat{\theta}-\theta}{se(\hat{\theta})}$ будет отрезок от квантиля $\frac{\alpha}{2}$ до квантиля $1-\frac{\alpha}{2}$ вектора t^*: $\widehat{CI}_{\theta^{st}} = \left[t_{\frac{\alpha}{2}}^*; t_{1-\frac{\alpha}{2}}^*\right]$. Тогда для получения оценки доверительного интервала для параметра θ необходимо выполнить ряд преобразований:

$$t_{\frac{\alpha}{2}}^* \leqslant \theta^{st} \leqslant t_{1-\frac{\alpha}{2}}^* \Leftrightarrow t_{\frac{\alpha}{2}}^* \leqslant \frac{\hat{\theta}-\theta}{se\left(\hat{\theta}\right)} \leqslant t_{1-\frac{\alpha}{2}}^* \Leftrightarrow \hat{\theta} - t_{1-\frac{\alpha}{2}}^* \cdot se\left(\hat{\theta}\right) \leqslant \theta \leqslant \hat{\theta} - t_{\frac{\alpha}{2}}^* \cdot se\left(\hat{\theta}\right)$$

Таким образом, $CI_t = \left[\hat{\theta} - t_{1-\frac{\alpha}{2}}^* \cdot se\left(\hat{\theta}\right); \hat{\theta} - t_{\frac{\alpha}{2}}^* \cdot se\left(\hat{\theta}\right)\right]$.

Эта версия бутстрэпа имеет ряд преимуществ над предыдущими версиями. Прежде всего, построенный с помощью стьдентизированного бутстрэпа доверительный интервал быстрее сходится к теоретическому: $|CL_\alpha - \widehat{CL}_{\alpha,n}| = O\left(\frac{1}{n}\right)$. Также стоит добавить, что такой доверительный интервал будет более корректно учитывать асимметрию выборочного распределения.

2.4. Бутстрэп в бутстрэпе

Однако точная формула стандартной ошибки, например, медианы либо других выборочных характеристик может быть неизвестна. В такой ситуации для подсчета стандартной ошибки может быть применен бутстрэп в бутстрэпе:

1. Из каждой бутстрэпированной выборки $\{\mathbf{X}_b^*\}_{b=1}^B$, генерируется B^* псевдовыборок $\mathbf{X}_{b,1}^{**}, \mathbf{X}_{b,2}^{**}, ..., \mathbf{X}_{b,B^*}^{**}$;
2. Для каждой бутстрэпированной выборки второго уровня оценивается параметр $\hat{\theta}_{b,b^*}^{**}$, $1 \leqslant b \leqslant B^*$. Далее вычисляется стандартная ошибка по формуле

$$\widehat{se}\left(\hat{\theta}_b\right) = \sqrt{\frac{\sum_{k=1}^{B^*}\left(\hat{\theta}_{b,k}^{**} - \bar{\hat{\theta}}_b^{**}\right)^2}{B^*-1}}$$

где $\bar{\hat{\theta}}_b^{**} = \frac{1}{B^*}\sum_{k=1}^{B^*}\hat{\theta}_{b,k}^{**}$ — среднее значение параметра по бутстрэпированным выборкам второго уровня.

2.5. Графическая иллюстрация точности методов

В данном разделе приводится графическая иллюстрация точности построения доверительных интервалов для математического ожидания (среднего) с помощью ЦПТ, перцентильного бутстрэпа и стьюдентизированного бутстрэпа. При использовании обеих версий бутстрэпа количество генерируемых выборок равняется 10^5. В качестве распределений использовались нормальное с параметрами $\mu = 10$, $\sigma^2 = 25$ и равномерное на отрезке $[0; 10]$.

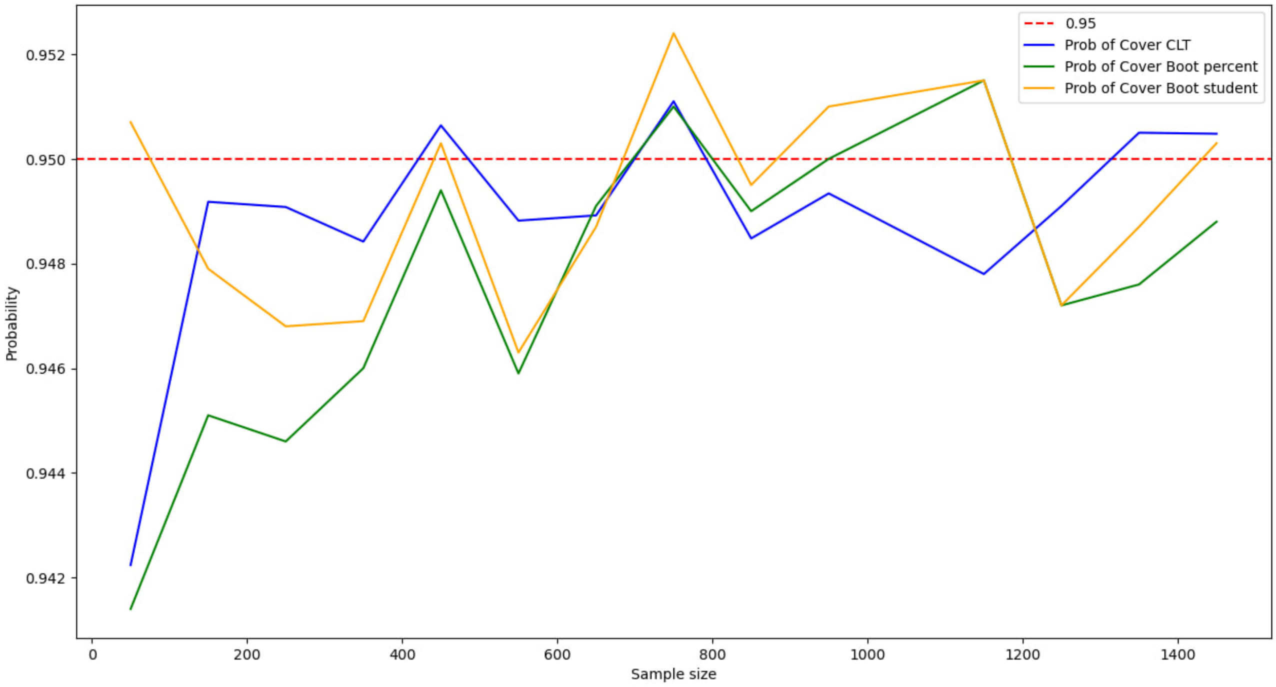

Рисунок 2.1. Зависимость оцениваемой вероятности накрытия от размера выборки, нормальное распределение. Число экспериментов для ЦПТ $= 5 \cdot 10^5$, для бутстрэпа $= 10^5$

Рис. 2.1 демонстрирует, что доверительный интервал, построенный при помощи бутстрэпа t-статистики, при небольших размерах выборки оказывается точнее доверительного интервала на основе ЦПТ и перцентильного бутстрэпа. Однако при объемах выборки в несколько сотен наблюдений в среднем более точным оказывается результат применения ЦПТ.

Для равномерного распределения (Рис. 2.2) вероятности накрытия на основе бутстрэпа также оказываются ближе к 0.95 только при размерах выборки до 200 — 300 наблюдений. Для бо́льших размеров выборки доверительные интервалы, оцененные с помощью ЦПТ, оказываются стабильно ближе к теоретическим.

Далее более внимательно сравним точность работы бутстрэпа по сравнению с ЦПТ для выборок маленького разбера (30, 50 и 100).

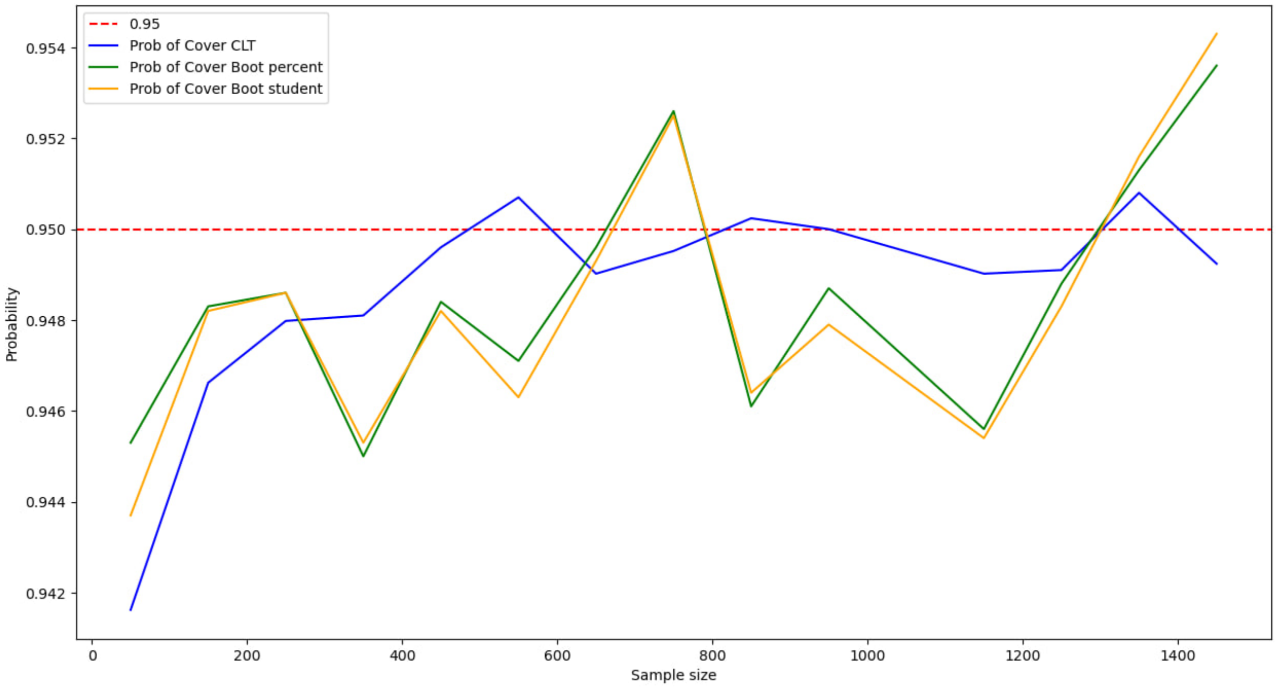

Рисунок 2.2. Зависимость оцениваемой вероятности накрытия от размера выборки, равномерное распределение. Число экспериментов для ЦПТ $= 5 \cdot 10^5$, для бутстрэпа $= 10^5$

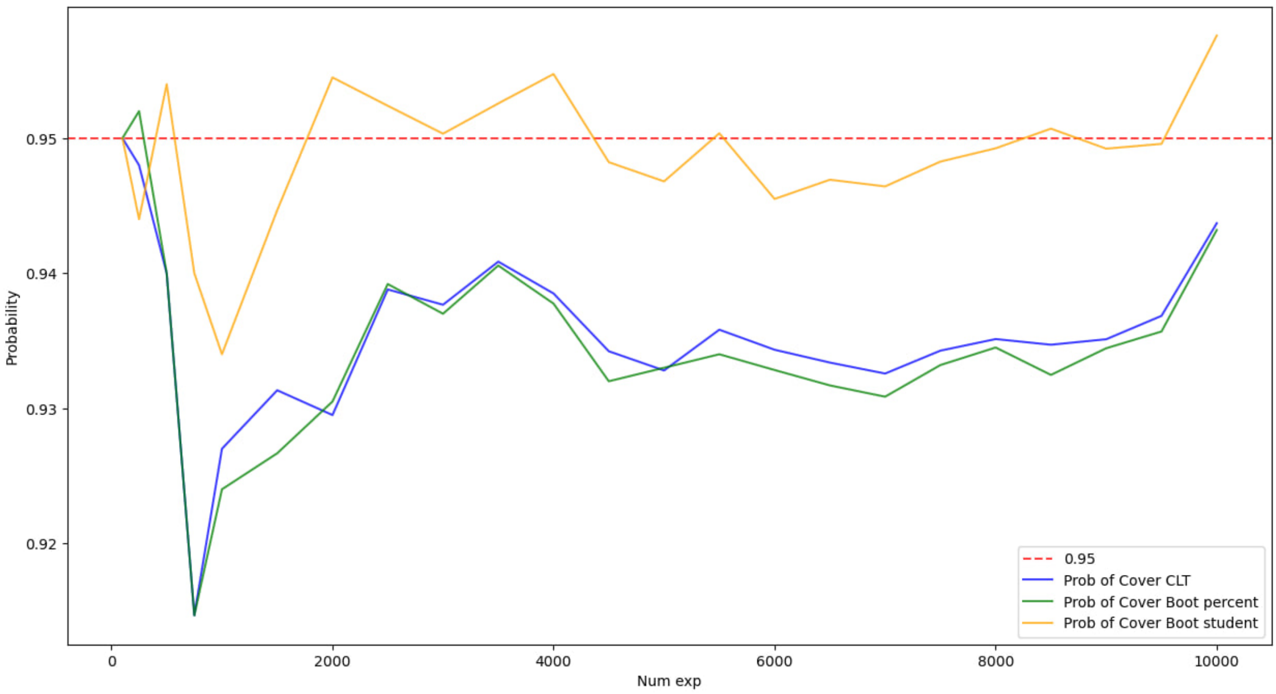

Рисунок 2.3. Зависимость оцениваемой вероятности накрытия от числа экспериментов, нормальное распределение, 30 наблюдений

Все три графика (Рис. 2.3, 2.4, 2.5) демонстрируют, что поведение вероятности накрытия истинного параметра (математического ожидания) очень схоже для ЦПТ и перцентильного бутстрэпа. При этом они оба систематически уступают стьдентизированному бутстрэпу в точности оценивания, особенно в ситуации работы с выборкой размером 30 наблюдений.

В случае оценки доверительного интервала для математического ожидания равномерного распределения результаты менее однозначны. Точность построения д. и. схожа для всех трех методов, причем все они склонны занижать размер оцененного доверительного интервала по сравнению с теоретическим (Рис. 2.6, 2.7, 2.8). Это происходит из-за низкого эксцесса равномерного распределения (случайные величины не концентрируются вблизи математического ожидания).

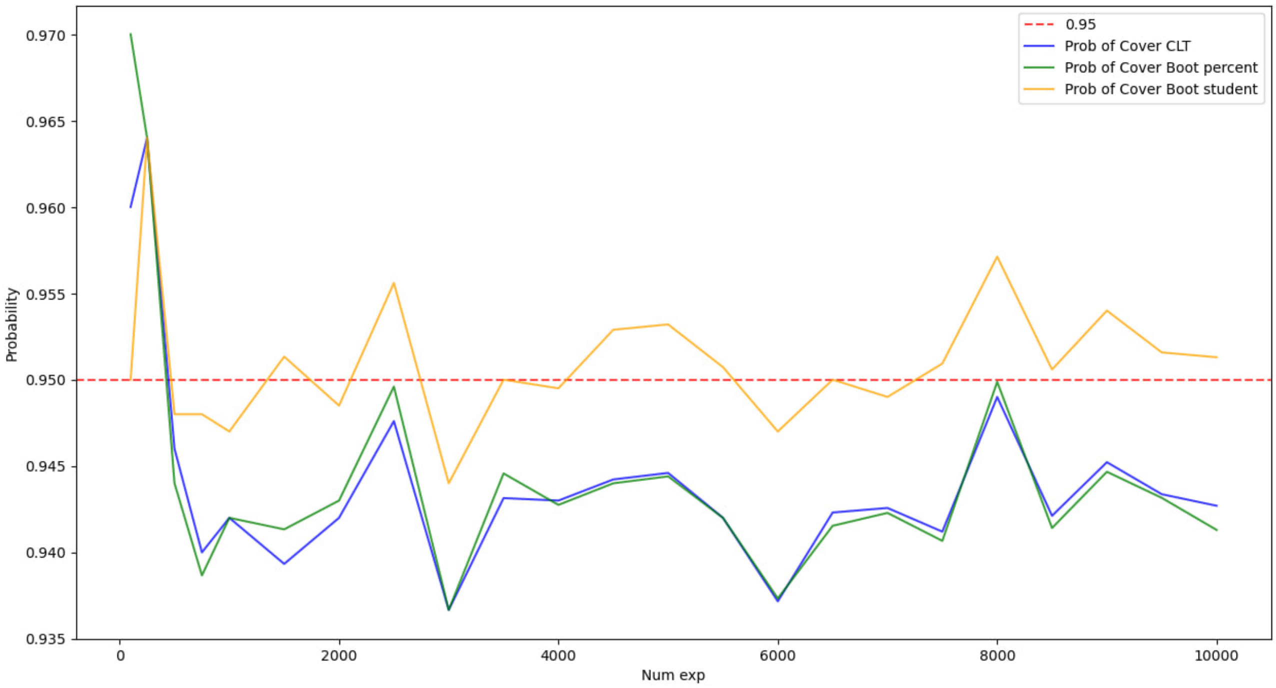

Рисунок 2.4. Зависимость оцениваемой вероятности накрытия от числа экспериментов, нормальное распределение, 50 наблюдений

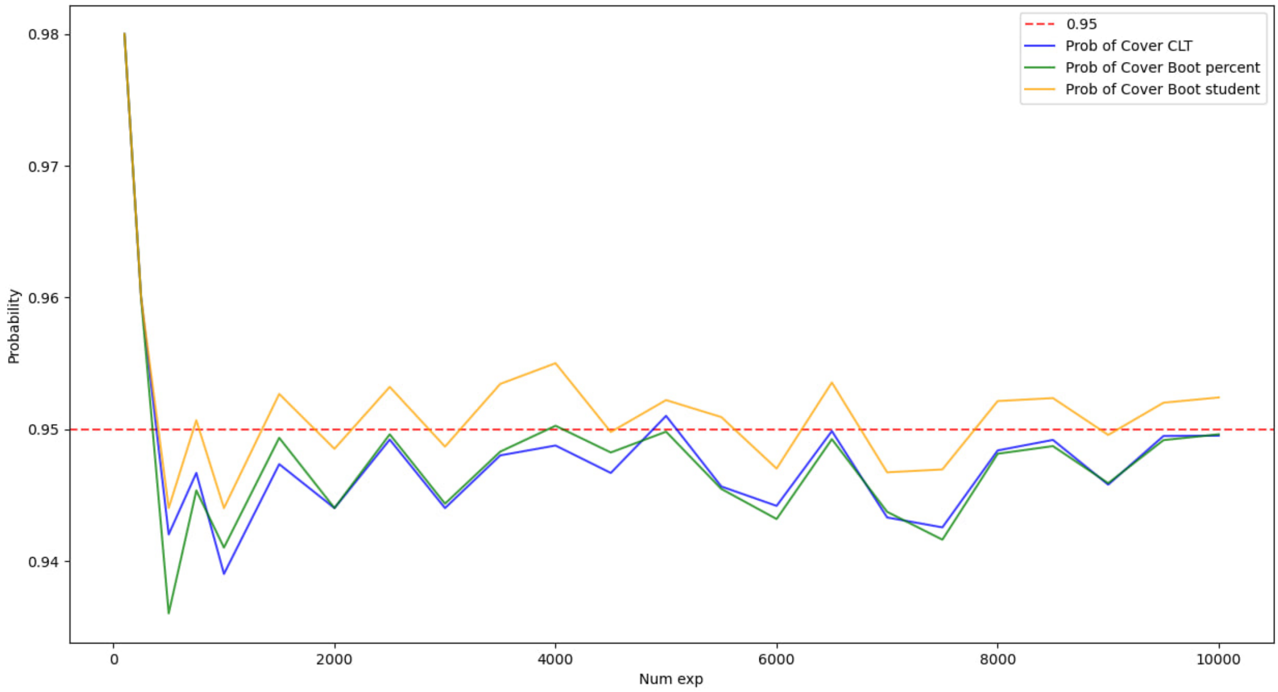

Рисунок 2.5. Зависимость оцениваемой вероятности накрытия от числа экспериментов, нормальное распределение, 100 наблюдений

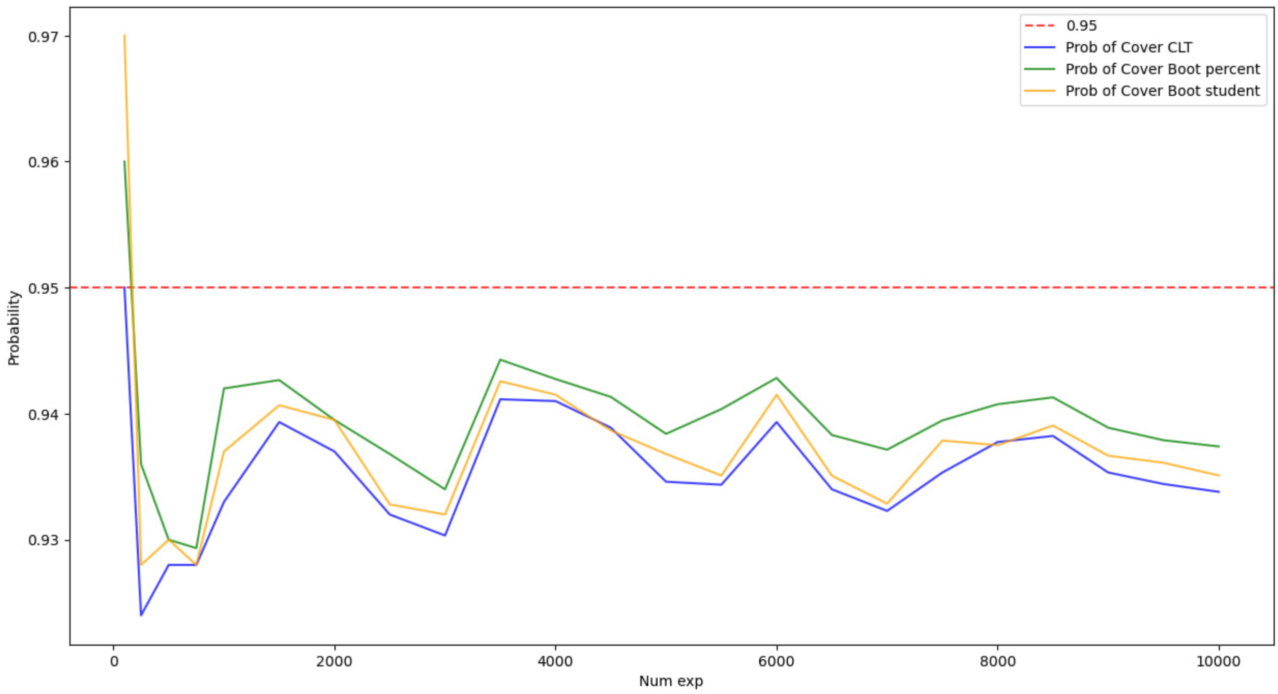

Рисунок 2.6. Зависимость оцениваемой вероятности накрытия от числа экспериментов, равномерное распределение, 30 наблюдений

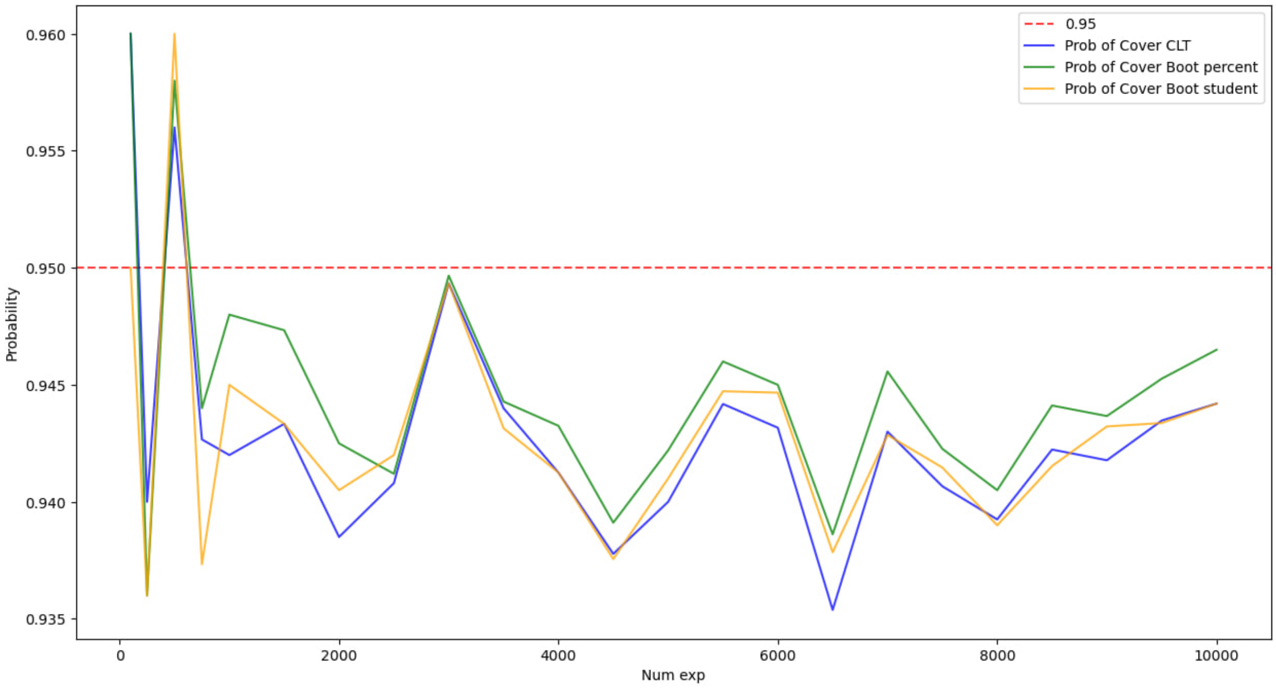

Рисунок 2.7. Зависимость оцениваемой вероятности накрытия от числа экспериментов, равномерное распределение, 50 наблюдений

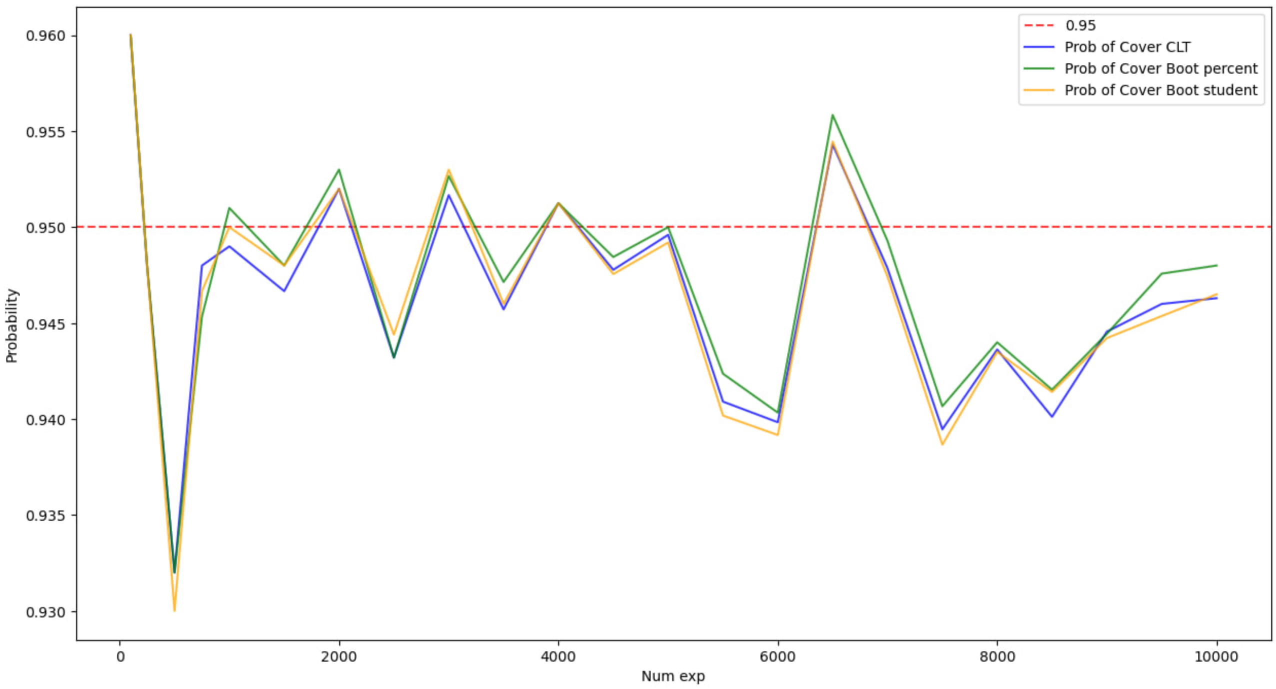

Рисунок 2.8. Зависимость оцениваемой вероятности накрытия от числа экспериментов, равномерное распределение, 100 наблюдений

2.6. Бутстрэп с коррекцией смещения и ускорением (Bias-corrected and accelerated bootstrap, BCa)

Методом, модифицирующим перцентильный бутстрэп и позволяющий корректировать асимметрию выборочного распределения, является BCa-бутстрэп. В нем дополнительно оцениваются два параметра: $\hat{a}$ — параметр «ускорения» и $\hat{z}_0$ — фактор коррекции смещения. Алгоритм BCa-бутстрэпа выглядит следующим образом:

1. Повторяются шаги 1, 2 из метода обычного перцентильного бутстрэпа;
2. Оценивается фактор коррекции смещения:
$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\{\hat{\theta}^*_b < \hat{\theta}\}\right)$$
где $\Phi^{-1}(\cdot)$ — обратная функция от функции распределения стандартной нормальной случайной величины, а ее аргументом является доля бутстрэпированных оценок, меньших выборочной оценки.
3. Далее небходимо оценить параметр «ускорения». Введем обозначения: $\hat{\theta}_{(i)}$ — оценка параметра θ без учета наблюдения X_i, $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$ — среднее таких оценок (jackknife оценка). Тогда параметр «ускорения» оценивается следующим образом:
$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^2\right]^{3/2}}$$
4. Следующим шагом вычисляются квантили бутстрэпированного распределения $\hat{\theta}$, которые станут границами доверительного интервала:
$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_\alpha}{1 - \hat{a}\left(\hat{z}_0 + z_\alpha\right)}\right)$$
$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\alpha}\right)}\right)$$
где $\Phi(\cdot)$ — функция распределения стандартного нормального распределения, z_α — квантиль стандартного нормального распределения порядка α.
5. Оцененным $(1-\alpha)$-процентным доверительным интервалом истинного параметра θ будет отрезок: $CI_{BCa} = \left[\hat{\theta}^*_{\alpha_1}; \hat{\theta}^*_{\alpha_2}\right]$.

3. Бутстрэп в линейной регрессии

Рассмотрим случай классической множественной линейной регрессии

$$y_i = x_i^\top \beta + \varepsilon_i,$$

где $\varepsilon_i \overset{i.i.d}{\sim} F_\varepsilon(\cdot)$ и $\mathbb{E}(\varepsilon_i|x_i) = 0$.

Алгоритмы бутстрэпа для линейной регрессии могут быть использованы те же, что и для простой выборки. Различаться будут скорее способы построения псевдовыборок для пар $\{(x_i, y_i)\}_{i=1}^n$. Рассмотрим два основных метода.

3.1. Парный бутстрэп

В парном бутстрэпе наблюдения для псевдовыборок равновероятно генерируются в виде пар $\{(x_i^*, y_i^*)\}_{i=1}^n$ из исходной выборки $\{x_i, y_i\}_{i=1}^n$:

$$\mathbb{P}\{x_l^* = x_i, y_l^* = y_i\} = \frac{1}{n}, \forall\ i = 1, ..., n,\ l = 1, ..., n$$

В итоге имеем B псевдовыборок $\{\mathbf{X}_b^*, \mathbf{y}_b^*\}_{b=1}^B$ с матрицей объекты-признаки $\mathbf{X}_b^*$ и вектором целевой переменной $\mathbf{y}_b^*$. Для каждой псевдовыборки оценивается вектор параметров $\{\hat{\beta}_b^*\}_{b=1}^B$. Далее для построения доверительных интервалов оценок коэффициентов может быть использован один из методов бутстрэпа, описанных в предыдущей главе.

3.2. Параметрический бутстрэп (бутстрэп по остаткам)

Парный бутстрэп может давать менее корректные результаты в случае наличия влиятельных наблюдений или выбросов с большим абсолютным значением остатков e_i. Для учета таких наблюдений может быть использован параметрический бутстрэп.

Первым шагом оценивается исходная модель и находится состоятельная оценка $\hat{\beta}$. Далее вычисляются остатки $e_i = y_i - x_i^\top \hat{\beta}$. В данном случае равновероятно генерируются пары $\{(x_i^*, e_i^*)\}_{i=1}^n$. Зависимая переменная восстанавливается как $y_i^* = x_i^{*\top} \hat{\beta} + e_i^*$. Однако в данном случае метод построения псевдовыборок идентичен парному бутстрэпу.

Повысить эффективность параметрического метода можно при соблюдении определенных предпосылок.

Если известно, что в рассматриваемой модели регрессоры и случайные ошибки независимы, генерировать $\{x_i^*\}_{i=1}^n$ из $\{x_i\}_{i=1}^n$ и $\{e_i^*\}_{i=1}^n$ из $\{e_i\}_{i=1}^n$ можно генерировать по отдельности, независимо.

Если помимо независимости регрессоров и случайных ошибок выполнено предположение о нормальности распределения случайных ошибок, т. е. $e_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, то еще более эффективным способом генерации псевдовыборок будет извлечение псевдоостатков $\{e_i^*\}_{i=1}^n$ из распределения $\mathcal{N}(0, \hat{\sigma}^2)$, где

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-k}$$

4. Бутстрэп в моделях бинарного выбора

Использование рассмотренных ранее параметрических методов генерирования псевдо-выборок будет некорректным для моделей бинарного выбора, поскольку независимая переменная y_i в данном случае принимает значения 0 и 1. Поэтому рассмотрим метод, близкий к парному бутстрэпу.

Вспомним, что модель модель бинарного выбора имеет следующий вид:

$$\mathbb{P}(y_i = 1 | x_i) = F(x_i^\top \beta),$$

где $F(\cdot)$ — функция распределения. Чаще всего используется стандартное нормальное распределение (пробит-модель) и стандартное логистическое распределение (логит-модель).

Первым шагом находятся оценки параметров для исходной модели $\hat{\beta}$. Затем генерируется вектор объясняющих переменных x_i^* равновероятно из исходных векторов $\{x_i\}_{i=1}^n$. Для каждого набора регрессоров независимая переменная генерируется следующим образом

$$y_i^* = \begin{cases} 1, & \text{с вероятностью } F(x_i^{*\top} \hat{\beta}) \\ 0, & \text{с вероятностью } 1 - F(x_i^{*\top} \hat{\beta}), \end{cases} \quad i = 1, ..., n$$