РАНДОМИЗИРОВАННАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ИМИТАЦИОННЫХ ЭКСПЕРИМЕНТОВ

Л. В. Гаев (Липецк)

Бутстреп-метод, предложенный в 1977 году Б. Эфроном [1], является одним из методов рандомизированной обработки данных. Сущность данного метода состоит в том, что по имеющимся наблюдениям за случайной величиной моделируется процесс их получения следующим образом. Предполагается, что имеющиеся N значений образуют генеральную совокупность, из которой извлекаются выборки с возвращением объема N с равными вероятностями (1/N) извлечения каждого значения. Всего извлекается B выборок, по каждой из них строится оценка интересующего параметра исходной случайной величины, а затем полученные оценки усредняются.

Если данные, по которым производится статистическая оценка, получены с точностью d, то результат стандартной оценки, например, математического ожидания будет иметь точность d/N. Оценка бутстреп-методом позволит получить точность d/BN. Поэтому для получения оценки результатов моделирования с требуемой точностью достаточно будет производить меньшее количество имитационных экспериментов, оставляя точность результатов моделирования прежней.

Данный метод находит широкое применение в различных областях, о чем свидетельствует обширный список литературы, доступный, например, в INTERNET по адресу http://www.resample.com. В то же время, строгое обоснование свойств бутстреп-оценок отсутствует, имеются лишь асимптотические оценки их поведения. Они не позволяют определить процедуру применения бутстреп-метода и его параметры для получения оценок требуемой точности. Поэтому встает проблема обоснования данного метода.

При проведенном экспериментальном исследовании свойств бутстреп-метода сначала были рассмотрены вопросы, связанные с влиянием количества бутстреп-повторений на результаты оценивания. Было выявлено [2], что при малых значениях B (количества бутстреп-повторений) получающиеся оценки могут располагаться в значительном интервале вокруг истинного значения оцениваемого параметр. С ростом числа B этот интервал уменьшается. Очевидно, что данное свойство определено связью между дисперсией бутстреп-оценки и количеством бутстреп-повторений.

Следующий этап исследований касался влияния интенсивности аддитивного белого шума, накладываемого на истинные значения исследуемого процесса, на результаты бутстреп-анализа. При моделировании шум возникает из-за учета не всех факторов, влияющих на исследуемый процесс. Полученные здесь данные показывают, как шум влияет на дисперсию бутстреп-оценки, и найден эмпирический закон влияния [3].

Наконец, было рассмотрено влияние наличия в выборке данных, не принадлежащих исследуемой генеральной совокупности (выбросов в наблюдениях). Подобные выбросы могут быть связаны с получением значений, находящихся в той области, где модель уже становится неадекватной исследуемому процессу. Получающиеся при этом значения бутстреп-оценок были смещенными, но, по сравнению со стандартными методами, это смещение было значительно меньше [4].

Проведенные экспериментальные исследования позволяют прийти к выводу, что применение бутстреп-метода анализа данных в ряде случаев может иметь преимущество по сравнению с традиционными при наличии шумов и выбросов в наблюдениях. Однако для корректности использования бутстреп-метода требуется решить ряд вопросов, определяющих саму процедуру бутстреп-анализа. Прежде всего, это вопрос о количестве бутстреп-повторений, поскольку он определяет саму процедуру обработки данных. Для решения указанного вопроса следует определить статистические характеристики бутстреп-оценок.

При изучении свойств бутстреп-метода к настоящему моменту исследовано поведение бутстреп-оценки вероятности успеха в одном испытании Бернулли [4,5] и бутстреп-оценки математического ожидания биномиальной случайной величины.

Пусть имеются результаты N испытаний Бернулли. Стандартной оценкой вероятности успеха p является величина

$$\hat{p} = \frac{k}{N},$$

где k – количество успехов среди N наблюдений.

Математическое ожидание и дисперсия оценки равны

$$M(\hat{p}) = p, D(\hat{p}) = \frac{pq}{N}.$$

Для бутстреп-оценки $\breve{p}$ показано [5], что:

$$M(\breve{p}) = p, \quad D(\breve{p}) = \frac{pq}{N}\left(1 + \frac{1}{B} - \frac{1}{BN}\right),$$

где B – количество бутстреп-выборок.

Для биномиальной случайной величины стандартной оценкой математического ожидания является

$$\tilde{m} = \frac{1}{N}\sum_{i=0}^{n} i x_i ,$$

где x_i – количество результатов наблюдений со значением i (i=0,1,2, … , n), при этом $\sum_{i=0}^{n} x_i = N$. Ее математическое ожидание и дисперсия равны

$$M\left(\tilde{m}\right) = np , \; D\left(\tilde{m}\right) = \frac{npq}{N}$$

Рассмотрим бутстреп-выборку $u_b=(k_0, k_1, \ldots, k_n)$ ($\sum_{i=0}^{n} k_i = N$), полученную из исходного набора наблюдений. Вероятность ее генерации равна

$$\frac{N!}{k_0!\,k_1!\ldots k_n!}\left(\frac{x_0}{N}\right)^{k_0}\left(\frac{x_1}{N}\right)^{k_1}\ldots\left(\frac{x_n}{N}\right)^{k_n} =$$

$$C_N^{k_0}\left(\frac{x_0}{N}\right)^{k_0} C_{N-k_0}^{k_1}\left(\frac{x_1}{N}\right)^{k_1}\ldots C_{N-k_0-\ldots-k_{n-2}}^{k_{n-1}}\left(\frac{x_{n-1}}{N}\right)^{k_{n-1}}\left(\frac{x_n}{N}\right)^{k_n}$$

«Частная бутстреп-оценка» строится аналогично стандартной

$$\tilde{m}_b = \frac{1}{N}\sum_{i=0}^{n} i k_i .$$

Перейдем к итоговой бутстреп-оценке интересующей нас величины по имеющимся результатам наблюдений Ф

$$\breve{m}_{\Phi} = \frac{1}{B}\sum_{b=1}^{B}\tilde{m}_b ,$$

где B – количество бутстреп-повторений, а $\tilde{m}_b$ – «частные бутстреп-оценки».

Ее математическое ожидание равно стандартной оценке. Найдем дисперсию

$$D\left(\breve{m}_{\Phi}\right) = M\left(\breve{m}_{\Phi}^2\right) - \left(M\left(\breve{m}_{\Phi}\right)\right)^2 = \left(\frac{1}{N^2} - \frac{1}{BN^3}\right)\left(\sum_{i=0}^{n} ix_i\right)^2 + \frac{1}{BN^2}\sum_{i=0}^{n} i^2 x_i - \left(\frac{1}{N}\sum_{i=0}^{n} ix_i\right)^2$$

Теперь можно найти безусловную дисперсию бутстреп-оценки $\breve{m}$ математического ожидания биномиальной случайной величины. Обозначим $p_i = C_n^i p^i (1-p)^{n-i}$ – вероятность получить значение i при очередном наблюдении (i=0,1,…,n). Тогда

$$M(\breve{m}^2) = \sum_{x_0=0}^{N} C_N^{x_0}\left(p_0\right)^{x_0} \ldots \sum_{x_{n-1}=0}^{N-x_0-\ldots-x_{n-2}} C_{N-\ldots-x_{n-2}}^{x_{n-1}}\left(p_{n-1}\right)^{x_{n-1}}\left(p_n\right)^{x_n}$$

$$\left(\left(\frac{1}{N^2} - \frac{1}{BN^3}\right)\left(\sum_{i=0}^{n} ix_i\right)^2 + \frac{1}{BN^2}\sum_{i=0}^{n} i^2 x_i\right) =$$

$$\left(\frac{1}{N^2} - \frac{1}{BN^3}\right)\left\{N\sum_{i=0}^{n} i^2 p_i + N(N-1)\left(\sum_{i=0}^{n} ip_i\right)^2\right\} + \frac{1}{BN^2}N\sum_{i=0}^{n} i^2 p_i =$$

$$\frac{npq}{N}\left(1 + \frac{1}{B} - \frac{1}{BN}\right) + (np)^2$$

Таким образом, дисперсия бутстреп-оценки математического ожидания биномиальной случайной величины

$$D(\breve{m}) = \frac{npq}{N}\left(1 + \frac{1}{B} - \frac{1}{BN}\right).$$

Как следует из приведенных расчетов, бутстреп-оценки вероятности успеха в одном испытании Бернулли и математического ожидания биномиальной случайной величины являются несмещенными и эффективными. Поскольку большое количество разнообразных законов распределения (как дискретных так и непрерывных случайных величин) могут быть получены при некоторых асимптотических допущениях из многократных применений испутаний Бернулли, то найденные результаты можно считать приближенно верными и для них. Это позволяет применять данный метод при оценивании результатов имитационных экспериментов. С ростом числа бутстреп-повторений их дисперсии приближаются к дисперсиям соответствующих стандартных оценок, и при этом точность получаемых оценок будет большей. При одинаковых значениях количества бутстреп-повторений, дисперсия будет меньшей на выборке меньшего объема. Теоретические и экспериментальные исследования показывают качественно одинаковые результаты: при малых значениях B, получающиеся оценки могут располагаться в значительном интервале вокруг истинного значения оцениваемого параметр. С ростом числа B этот интервал уменьшается, но, начиная с некоторого значения, уменьшение прекращается. Очевидно, данное свойство определено связью между дисперсией бутстреп-оценки, количеством бутстреп-повторений и точностью получаемых результатов.

В настоящий момент остается открытым вопрос о количестве бутстреп-повторений B, требуемом для получения достаточно хороших оценок. Его решение, скорее всего, должно определяться зависимостью между дисперсией оценки и точностью получаемых результатов вычислений.

Литература

1. **Эфрон Б.** Нетрадиционные методы многомерного статистического анализа.- М.: Финансы и статистика, 1988.- 263 с.
2. **Гаев Л.В., Шмарион М.Ю.** Компьютерное исследование бутстреп-моделирования//Современные проблемы информатизации. Тезисы докладов второй электронной научной конференции.- Воронеж: ВГПУ, 1997.- С.176.
3. **Гаев Л.В., Шмарион М.Ю.** Исследование степени влияния интенсивности шума на бутстреп-оценку//Современные проблемы информатизации. Тезисы докладов третьей электронной научной конференции.- Воронеж: ВГПУ, 1998.
4. **Блюмин С.Л., Гаев Л.В., Шмарион М.Ю**. Характеристика бутстреп-оценки математического ожидания бернуллиевской случайной величины//Вестник ЛГТУ-ЛЭГИ, 2001, – № 1.
5. **Гаев Л.В.** О поведении бутстреп-оценки вероятности успеха в одном испытании Бернулли//Современные проблемы математики и естествознания. Материалы пятой Всероссийской научно-технической конференции.- Н.Новгород: МВВО АТН РФ, 2003.- С. 6–7.