УДК 677.024:519.876.5
DOI 10.47367/0021-3497_2021_1_50

ОЦЕНКА БУТСТРЕП-МЕТОДА ДЛЯ ОБРАБОТКИ РЕЗУЛЬТАТОВ ТЕХНОЛОГИЧЕСКИХ ИЗМЕРЕНИЙ

EVALUATION OF THE BOOTSTRAP METHOD FOR PROCESSING RESULTS OF TECHNOLOGICAL MEASUREMENTS

П.А. СЕВОСТЬЯНОВ, Т.А. САМОЙЛОВА

P.A. SEVOSTYANOV, T.A. SAMOYLOVA

(Российский государственный университет имени А.Н. Косыгина (Технологии. Дизайн. Искусство))

(Russian State University named after A.N. Kosygin (Technologies. Design. Art))

E-mail: petrsev46@yandex.ru

Выполнено сравнение результатов оценивания коэффициента парной корреляции методами, основанными на нормальной и асимптотической теории статистики, и бутстреп-методами. Сравнение выполнено методами статистического моделирования с высокой надежностью получаемых выводов, которая обеспечивается большими объемами повторных испытаний. Исследование проведено на примере коэффициента парной корреляции для выборочных данных в широком диапазоне размеров выборки для нормальных и равномерно распределенных случайных величин.

The results of evaluating the pair correlation coefficient using methods based on normal and asymptotic statistical theory and bootstrap methods are compared. The comparison was performed using statistical modeling methods with high reliability of the obtained conclusions, which is provided by large volumes of repeated tests. The study is based on the example of the pair correlation coefficient for sample data in a wide range of sample sizes for normal and uniformly distributed random variables.

Ключевые слова: текстильные материалы, показатели качества, выборка, бутстреп, точность оценки, надежность оценки.

Keywords: textile materials, quality indicators, sampling, bootstrap, accuracy of evaluation, reliability of evaluation.

Измерения и обработка данных при оценке качественных показателей текстильных материалов и изделий, как правило, ввиду больших трудо- и временных затрат вынуждены обходиться выборками малого объема [1]. Бутстреп (bootstrap)-методы (БМ) уже сорок лет привлекают заманчивой перспективой извлечь большой

объем информации из выборочных данных при малом их объеме n [2], [3]. Результат достигается за счет многократного использования выборочных данных в повторных выборках. Повторные выборки с возвратом данных отбираются из исходной выборки случайным образом. Общее количество повторных выборок одинакового объема и отличающихся друг от друга хотя бы одним элементом равно n^n. Даже для выборок умеренного объема это количество может быть огромным и позволяет надеяться на получение не только точечных, но и интервальных оценок целевых показателей. С позиций теории информации интерпретация эффективности БМ такова: классические статистические методы, основанные на нормальной, асимптотической или робастной теории (АН - оценки), не позволяют извлечь "всю" информацию, содержащуюся в исходной выборке. БМ позволяют это сделать, и поэтому можно обойтись выборками меньшего объема. К сожалению, теоретические исследования, посвященные изучению информационно-статистических свойств БМ - оценок, не дают однозначного ответа на вопрос об их преимуществах. Литература на тему БМ, в основном, сводится к описанию разных вариаций БМ и алгоритмов их реализации. Вместе с тем о популярности БМ говорит включение инструментария для реализации БМ в распространенные программные пакеты.

Наибольший интерес представляет задача о зависимости точности БМ-оценки целевого показателя от объема выборки и сравнения этой оценки с АН - оценкой. Рассмотрим эту задачу на примере исследования корреляционных показателей. В качестве такого целевого показателя возьмем коэффициент парной корреляции между двумя нормально распределенными случайными величинами. Коэффициент корреляции является одним из наиболее часто оцениваемых показателей во многих исследованиях регрессий и корреляций при контроле и управлении технологическими процессами и изучении свойств текстильных материалов. Поэтому результаты оценивания представляют общий интерес. Процедура получения как точечной, так и интервальной оценок коэффициента парной корреляции для нормально распределенных случайных величин известна [4], [5]. Распределение точечной выборочной оценки коэффициента корреляции с ростом объема выборки весьма медленно сходится к нормальному распределению. Поэтому при умеренных объемах выборки (до 100 пар значений) для применения нормальной теории и получения интервальных оценок используют так называемое преобразование Фишера оценки коэффициента корреляции [6], [7]. Именно этот общепризнанный подход реализован в функции corr системы Matlab. БМ-оценки коэффициента корреляции получим методом компьютерного статистического моделирования [8…10].

В качестве источника данных – парной выборки – используем простейшую линейную корреляционную модель Y = X + Z. Здесь X и Z, а соответственно, и Y – нормально распределенные случайные величины. Если X и Z не коррелированы, то коэффициент парной корреляции между X и Y равен

$$R = 1 \Big/ \sqrt{1 + S_z^2 / S_x^2} \ ,$$

где S_x и S_z – среднеквадратические отклонения (СКО) X и Z. При имитации выборки выберем значения СКО $S_x = 1$ и $S_z = \sqrt{3}$. Тем самым задаем определенный коэффициент корреляции между X и Y , равный R = 0,5.

Реализуем "классическую" АН - оценку коэффициента корреляции средствами Matlab с помощью следующих строк

```
>> n = 10000; x = normrnd (0, 1, n, 1); z =
= normrnd (0, sqrt(3), n, 1);  y = x + z;
>> [r, p, rlo, rup] = corrcoef (x, y);
```

При точном значении коэффициента корреляции 0,5 его точечная оценка по парной выборке объемом n = 10000 пар значений X и Y оказалась равной r = 0,5078, и интервальная оценка (rlo = 0,4931; rup = = 0,5222). Как видим, при больших объемах выборки n точность оценки весьма высока с высокой надежностью 0,95. Для выборок малого объема n = 5 вычисленные тем же методом точечная и интервальная АН-оценки являются абсолютно неприемле-

мыми. Например, одна выборка дает точечную оценку r = +0,1744 и интервальную оценку (rℓo = -0,8366; rup = +0,9158), вторая выборка r = +0,8603 и (rℓo = -0,0911; rup = = +0,9907), а третья выборка r = - 0,0066 и (rℓo = -0,8837; rup = 0,8808), что лишний раз доказывает непригодность методов асимптотической или нормальной теории для оценки коэффициента корреляции при малых выборках.

Рассмотрим возможности БМ для решения той же задачи точечной и интервальной оценки коэффициента корреляции для разных объемов парной выборки. Число бутстреп-выборок, генерируемых на основе исходной парной выборки, обозначено nb. Это число в эксперименте принималось равным nb = 500. Для получения оценок использованы следующие команды Matlab:

>> n = 10000; x = normrnd(0, 1, n, 1); z = = normrnd(0, sqrt(3), n, 1); y = x + z;

>> nb = 500; [bootstat, bootsam] = bootstrap (nb, @corr, y, x) ;

>> rb = mean(bootstat); [rbℓo, rbup] = = bootci(nb, @corr, y, x);

и далее повторение для nb = 100; 25; 10; 5.

Результаты модельного эксперимента приведены в табл. 1 (сравнение оценок парного коэффициента корреляции нормальным/асимптотическим методом и бутстреп-методом). В 1-й строке приведен объем модельной парной выборки n. Во 2-й строке - точечная АН-оценка. В 3-й и 4-й строках приведены нижняя и верхняя границы интервальной АН-оценки коэффициента корреляции. В 5-й, 6-й и 7-й строках – точечная и интервальная БМ-оценки. В предпоследних строках таблицы приведены длины оцененных доверительных интервалов для доверительной вероятности 0,95, равные D = = rup – rℓo и Db = rbup – rbℓo, а в последней строке – их отношение Ebc = Db / Dc.

Таблица 1

1	n	10000	100	25	10	5
2	r	0,5078	0,4455	0,4518	0,5109	0,4459
3	rℓo	0,4931	0,2730	0,0689	-0,1750	-0,7193
4	rup	0,5222	0,5903	0,7186	0,8629	0,9532
5	rb	0,4900	0,5008	0,5728	0,5502	0,4872
6	rbℓo	0,4722	0,3255	0,1933	-0,0910	-1,0000
7	rbup	0,5027	0,6420	0,7785	0,8878	1,0000
8	D	0,0291	0,3173	0,6497	1,0379	1,6725
9	Db	0,0305	0,3165	0,5852	0,9788	2,0000
10	Ebc	1,0481	0,9975	0,9007	0,9431	1,1958

На рис. 1 (r = 0,4789; rℓo = -0,6985; rup = = 0,9569; rb = 0,4815; rbℓo = -1; rbup = +1. Объем выборки n = 5) и рис. 2 (r = 0,6819; rℓo = 0,0915; rup = 0,9176; rb = 0,6719; rbℓo= =0,2226; rbup=0,9327. Объем выборки n=10) показаны примеры гистограмм значений точечной оценки коэффициента корреляции rb для R, полученной БМ-методом для n = 5 и n = 10, построенные по nb = 500 бутстреп-выборкам. Видно, что БМ-оценки так же, как и АН-оценки коэффициента корреляции, имеют существенно асимметричные распределения с настолько широким диапазоном значений, что это делает оценки неприемлемыми.

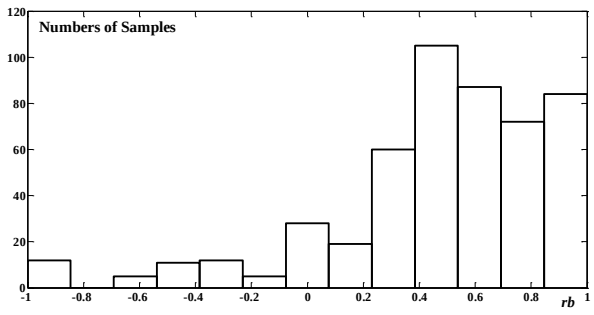

Рис. 1

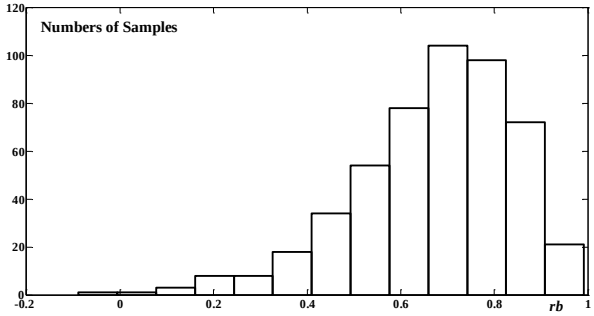

Рис. 2

Сравним полученные результаты АН и БМ-оценивания парного коэффициента корреляции. Оба подхода дают близкие точечные оценки и практически одинаковую точность (длины доверительных интервалов), хорошие при очень больших выборках, удовлетворительные при выборках среднего объема и неприемлемые для малых выборок n = 10 и 5. Значения критерия Ebc наглядно показывают, что в широком диапазоне объемов выборок точность оценки коэффициента корреляции сравниваемыми методами практически одинакова.

Чтобы убедиться в робастности этого вывода, эксперимент был повторен с равномерно распределенными случайными величинами X и Z : X ~ Uniform(-1;1); Z ~ Uniform(-√3; √3). В табл. 2 для разных объемов парных выборок приведено по четыре повторных значения показателя Ebc, вычисленные на разных множествах случайных чисел. Из данных таблицы следует, что на всем исследованном диапазоне объемов выборок от n =5 до n = 10000 длина доверительных интервалов для оценки коэффициента корреляции, полученной с применением формул асимптотической теории оценивания, практически такая же, как и оценка, полученная с применением БМ. Заметим, что при малых объемах выборки n = 5 оба класса методов оценивания дают одинаково большой диапазон разброса длины доверительного интервала, то есть точность оценивания недопустимо мала и ненадежна, каков бы ни был метод, и как бы ни были распределены исследуемые случайные величины.

Таблица 2

n	5	10	25	100	10000
Ebc	2,3310	1,1346	1,0929	1,0810	1,0116
	0,5821	1,0214	1,1466	0,8966	0,8914
	1,1272	1,1478	0,8198	0,9408	0,9030
	1,1337	0,8452	0,9903	0,9748	0,9412

Следовательно, по крайней мере, в данной задаче затраты на более сложную и длительную компьютерную обработку выборочных данных при использовании БМ оказываются неоправданными повышением точности и достоверности результатов.

ВЫВОДЫ

Бутстреп-методы дают приемлемые оценки измеряемого показателя на выборках среднего (> 30) и большого объема. При этом их точность сопоставима с точностью оценок, получаемых на основе нормальной или асимптотической теории статистики, то есть бутстреп-методы, будучи более трудоемкими, не дают выигрыша в точности и надежности оценивания. Для малых выборок нормальные и бутстреп-методы дают одинаково неточные результаты с низкой надежностью.

ЛИТЕРАТУРА

1. *Соловьев А.Н., Кирюхин С.М.* Оценка и прогнозирование качества текстильных материалов. – М.: Легкая и пищевая промышленность, 1984.

2. *Anatolyev S.A.* Robustness of residual-base bootstrap to composition of serially correlated errors // Journ. Statistical Computation and Simulation. – Vol. 79, №.3. P.315…320.

3. *Орлов А.И.* Прикладная статистика. – М.: Изд-во "Экзамен", 2004.

4. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. – М.: Физматлит, 2006. ISBN 5-9221-0707-0

5. *Севостьянов П.А., Ордов К.В*. Основы анализа и моделирования данных в технике и экономике. – М.: "Тисо Принт", 2015.

6. *Севостьянов П.А*. Математические методы обработки данных. – М.: МГТУ им. А.Н. Косыгина, 2004. ISBN 5-8196-0056-9

7. *Митропольский А.К.* Техника статистических вычислений. – 2-е изд., перераб. и доп. – М.: Наука, 1971.

8. *Севостьянов П.А., Самойлова Т.А., Монахов В.И., Тихомирова М.Л., Забродин Д.А.* Современные информационные технологии в исследованиях и оптимизации процессов рыхления и очистки экологических волокнистых материалов // Сб. научн. тр. Междунар. научн.-технич. симпозиума: Современные инженерные проблемы в производстве товаров народного потребления Международного Косыгинского Форума: Современные задачи инженерных наук: Современные инженерные проблемы в производстве товаров народного потребления: сборник научных трудов (29-30 октября 2019 г.). – М.: РГУ имени А.Н. Косыгина, 2019. Часть 2. С. 88...93.

9. *Севостьянов П.А., Самойлова Т.А., Монахов В.В.* Моделирование удлинения основной нити в ткани // Изв. вузов. Технология текстильной промышленности. – 2019, № 2. С. 199...202.

10. *Севостьянов П.А., Самойлова Т.А., Монахов В.В., Воробьев И.Н*. Планирование экспериментов и обработка данных моделирования процессов старения полимерных материалов // Сб. мат. Междунар. научн.-техн. конф.: Дизайн, технологии и инновации в текстильной и легкой промышленности (ИННОВАЦИИ-2018). Часть 2. – М.: РГУ имени А.Н. Косыгина, 2018. С. 246...249.

R E F E R E N C E S

1. Solov'ev A.N., Kiryukhin S.M. Otsenka i prognozirovanie kachestva tekstil'nykh materialov. – M.: Legkaya i pishchevaya promyshlennost', 1984.

2. Anatolyev S.A. Robustness of residual-base bootstrap to composition of serially correlated errors // Journ. Statistical Computation and Simulation. – Vol.79, №.3. P.315…320.

3. Orlov A.I. Prikladnaya statistika. – M.: Izd-vo "Ekzamen", 2004.

4. Kobzar' A.I. Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov. – M.: Fizmatlit, 2006. ISBN 5-9221-0707-0

5. Sevost'yanov P.A., Ordov K.V. Osnovy analiza i modelirovaniya dannykh v tekhnike i ekonomike. – M.: "Tiso Print", 2015.

6. Sevost'yanov P.A. Matematicheskie metody obrabotki dannykh. – M.: MGTU im. A.N. Kosygina, 2004. ISBN 5-8196-0056-9

7. Mitropol'skiy A.K. Tekhnika statisticheskikh vychisleniy. – 2-e izd., pererab. i dop. – M.: Nauka, 1971.

8. Sevost'yanov P.A., Samoylova T.A., Monakhov V.I., Tikhomirova M.L., Zabrodin D.A. Sovremennye informatsionnye tekhnologii v issledovaniyakh i optimizatsii protsessov rykhleniya i ochistki ekologicheskikh voloknistykh materialov // Sb. nauchn. tr. Mezhdunar. nauchn.-tekhnich. simpoziuma: Sovremennye inzhenernye problemy v proizvodstve tovarov narodnogo potrebleniya Mezhdunarodnogo Kosyginskogo Foruma: Sovremennye zadachi inzhenernykh nauk: Sovremennye inzhenernye problemy v proizvodstve tovarov narodnogo potrebleniya: sbornik nauchnykh trudov (29-30 oktyabrya 2019 g.). – M.: RGU imeni A.N. Kosygina, 2019. Chast' 2. S. 88...93.

9. Sevost'yanov P.A., Samoylova T.A., Monakhov V.V. Modelirovanie udlineniya osnovnoy niti v tkani // Izv. vuzov. Tekhnologiya tekstil'noy promyshlennosti. – 2019, № 2. S. 199...202.

10. Sevost'yanov P.A., Samoylova T.A., Monakhov V.V., Vorob'ev I.N. Planirovanie eksperimentov i obrabotka dannykh modelirovaniya protsessov stareniya polimernykh materialov // Sb. mat. Mezhdunar. nauchn.-tekhn. konf.: Dizayn, tekhnologii i innovatsii v tekstil'noy i legkoy promyshlennosti (INNOVATsII-2018). Chast' 2. – M.: RGU imeni A.N. Kosygina, 2018. S. 246...249.

Рекомендована кафедрой автоматизированных систем обработки информации и управления. Поступила 14.09.20.