УДК 311

М.Ю. Карышев

ПРИМЕНЕНИЕ ЧИСЛЕННЫХ МЕТОДОВ В АНАЛИЗЕ ОТДЕЛЬНЫХ ПОКАЗАТЕЛЕЙ СТАТИСТИКИ ТРАНСПОРТА

Статистические методы оценивания параметров и проверки гипотез традиционно основываются на ряде предположений о свойствах изучаемой выборочной совокупности и требований к ним, часто невыполнимых (например, достаточно большой объем выборки, нормальность ее распределения). В статье рассматривается альтернативный подход, позволяющий нивелировать эту проблему посредством использования статистического метода бутстреп и рандомизации, относящихся к области численных методов. Целью работы является демонстрация прикладной ценности и действенности этого метода в анализе статистических показателей железнодорожного и автомобильного транспорта на примере регионов Поволжья.

Ключевые слова: транспорт, статистика, численные методы, бутстреп, рандомизационный тест

Введение. Статистическая наука рассматривает любой набор данных как результат реализации некоторой гипотетической генеральной совокупности, истинные параметры которой неизвестны сейчас и вряд ли будут известны когда-либо в будущем. Все, что может статистика, это оценить интересуемые параметры по имеющейся выборке полученных эмпирическим путем числовых или нечисловых значений. Такие оценки, помимо прочих аспектов классификации, подразделяются на точечные, выраженные единым числом, и интервальные, ограничивающие с определенной вероятностью область, накрывающую истинное значение параметра генеральной совокупности. Тонким моментом здесь является то обстоятельство, что предпосылки к построению этих доверительных интервалов основаны на уверенном предположении о знании закона распределения, которому подчиняется генеральная совокупность. Однако во многих случаях получить такую уверенность весьма проблематично в силу специфики имеющихся данных, например, их малочисленности.

В этой связи решением может стать применение численных методов, в целом, не требующих подобной априорной информации. В 1979 г. профессор Стэнфордского университета Б. Эфрон опубликовал статью «Компьютеры и статистика: подумаем о невероятном» [1], где обосновал развитие нового класса «альтернативных компьютерно-интенсивных (computer-intensive) технологий, включающих рандомизацию, бутстреп и методы Монте-Карло» [2]. Особенностью этих методов явилось то, что они могли выполнять многократную обработку исходной выборочной совокупности путем извлечения из нее подвыборок и таким «магическим» образом генерировать новые данные, казалось бы, «из ничего». Как показала практика, наиболее ценным из этих методов в плане анализа малых выборок является бутстреп (bootstrap) – случайный повторный отбор. Он, конечно же, не создает новые данные и не компенсирует малый размер выборки. Его смысл в том, чтобы показать, как поведут себя многочисленные подвыборки, извлеченные из исходной выборки, полагая следующее: для подвыборок исходная выборка то же, что для исходной выборки – генеральная совокупность.

Материалы и методы. В процессе анализа с целью оценивания выборочных характеристик использовался типичный алгоритм метода бутстреп, который, например, для оценки среднего значения, включает следующие этапы [3]:

1. Извлечение значения x_i из исходной выборки $\{x_1, x_2, \dots, x_n\}$, его регистрация и возвращение обратно в выборку;
2. Повторение пункта первого n раз;
3. Определение среднего для n повторно отобранных значений;

4. Повторение b раз этапов 1–3;
5. На основе полученного распределения совокупности из b средних вычисление для их выборочного среднего: а) стандартной ошибки и б) границ доверительного интервала.

В общем случае этот алгоритм легко обобщается на любую иную выборочную характеристику θ^*: медиану, стандартное отклонение и пр. Стандартная ошибка такой характеристики для b выборок бутстрепа является ее стандартным отклонением [4]:

$$se_{boot} = [\frac{1}{b-1}\sum_{j=1}^{b}[\theta^{*j} - \theta^*(\cdot)]^2]^{1/2}, \text{ где } \theta^*(\cdot) = \frac{1}{b}\sum_{j=1}^{b}\theta^{*j}.$$

В качестве методов определения границ доверительных интервалов (например, в случае с оцениванием среднего значения) при 5 %-ном уровнем значимости применялись:

а) метод процентилей, как самым простой и интуитивно понятный, использующий в качестве границ доверительного интервала квантили бутстреп-распределения $[q_{\alpha}^*,\ q_{(1-\alpha)}^*]$;

б) метод основных интервалов, полезный в случае наличия асимметрии распределения статистик, полученных в результате применения бутстрепа:

$$[2\bar{x}_{boot} - q_{(1-\alpha)}^*;\ 2\bar{x}_{boot} - q_{\alpha}^*];$$

в) метод, основанный на использовании t-критерия:

$$\bar{x}_{boot} \pm t_{\alpha} se_{boot},$$

где t_{α} – критическое значение α-го квантиля распределения Стьюдента $\mathrm{t}(\alpha,\ n-1)$.

Логическим продолжением оценивания параметров распределения является статистическая проверка гипотез, поскольку «если в ходе эксперимента изучаются свойства объекта, то по результатам измерений можно сформулировать некоторые содержательные предположения (*научные гипотезы*) о природе наблюдаемых закономерностей» [2]. В рамках проведенного анализа проверке была подвергнута гипотеза об однородности двух выборок объемом n_1и n_2 путем сравнения их средних (нулевая гипотеза H_0: различия случайны и, значит, выборки извлечены из одной и той же генеральной совокупности, альтернативная H_1: отличия носят неслучайный характер). Для этого использовался реализующий процесс Монте-Карло рандомизационный тест, имеющий такой алгоритм (подробно описанный в [2]):

1. С целью оценки значимости различий выборочных средних $\bar{X}_1$и $\bar{X}_2$ двух групп данных со стандартным отклонением $S_{\bar{X}}$ выбирается некоторый статистический критерий со статистикой T (например, t-статистика Стьюдента: $(\bar{X}_1 - \bar{X}_2)/S_{\bar{X}}$);
2. Исчисляется наблюдаемое значение этой статистики t_{obs} для исходных сравниваемых выборок;
3. Некоторое число раз (B раз >1000) выполняются в цикле такие действия:

- объединение данных из обеих выборок и перемешивание их случайным образом

- определение первых n_1 наблюдений в первую группу, а остальных n_2 – во вторую

- вычисление тестовой статистики t_{ran} для полученных рандомизированных данных

- увеличение переменной-счетчика b в том случае, если $|t_{ran}| > |t_{obs}|$

4. Производится расчет относительной частоты (b/B), с которой величина t_{ran} превышает значение t_{obs}, что соответствует оценке вероятности p того, что случайная величина T примет значение, большее чем t_{obs}. При $p > 0.05$ принимается нулевая гипотеза H_0: $\mu_1 = \mu_2$ о равенстве средних, в ином случае альтернативная – H_1.

В состав анализируемой совокупности вошли регионы Приволжского федерального округа (14 единиц), характеризуемые по ряду статистических показателей в разрезе видов

транспорта [5]. В частности, для железнодорожного транспорта: отправлено грузов (млн. т.) и отправлено пассажиров (тыс. чел.), плотность железнодорожных путей (км путей на 10000 кв. км территории); для автомобильного транспорта: перевозки грузов (млн. т.) и перевозки пассажиров (млн. чел.), плотность автомобильных дорог общего пользования с твердым покрытием (1 км пути на 1000 кв. км территории). Перечень анализируемых индикаторов мог бы быть более емким, но, к сожалению, публикуемая Росстатом региональная статистика в этом аспекте очень лаконична (так, в ней отсутствует информация о грузообороте и пассажирообороте железнодорожного транспорта, что весьма и весьма удивительно). Подобные сведения, конечно же, имеют место в специализированных статистических сборниках, но там они, как правило, представлены в целом по стране.

Результаты и обсуждение. Существует выработанный статистической практикой ценз, согласно которому совокупности, объем которых (иначе говоря, число входящих в их состав единиц) не превышает 30, считаются малыми. Соответственно, на этом основании корректируются предположения о характеристиках такой совокупности, а также выбираются иные критерии для оценивания ее параметров. Например, в такой ситуации нередко осуществляется переход от методов параметрической оценки, основанных на знании законов распределения исследуемой совокупности, к непараметрическим методам, такого знания не требующим.

В настоящем случае исследуемые регионы Поволжья, в общем счете не превышающие полутора десятка субъектов, могут быть отнесены к разряду малых совокупностей. Для демонстрации того, что представляют собой исходные данные такой совокупности (в качестве примера был выбран показатель отправления грузов железнодорожным транспортом) – их подчиненность нормальному закону и распределение по величине значений – построены, соответственно, графики а) и б) рис. 1. Совершенно очевидно, что эмпирические величины не согласуются с теоретическими ожиданиями. Так, график а), характеризующий соотношения наблюдаемых значений и соответствующих им квантилей нормального распределения (часто называемый графиком квантиль-квантиль), не выполняет требования нахождения этих величин на прямой; график б) показывает заметное расхождение между столбцами распределения эмпирических данных и кривой нормального распределения. Такое положение вещей означает, что характеристики, которые могут быть получены в результате анализа этой совокупности, будут вызывать определенные сомнения в их надежности и правдивости. Решение обозначенной проблемы может быть найдено по-разному. Например, путем замены характеристик, чувствительных к объему совокупности (прежде всего, это среднее) на так называемые робастные (по-другому, устойчивые) статистики (наиболее широко используемой из них является медиана – центр распределения, делящий совокупность на две равные части, единицы одной из которых имеют значения признака, не большие медианного, а единицы другой – не меньшие). Однако в контексте настоящей работы логично будет привлечь для анализа вычислительные возможности метода бутстреп.

Применение метода бутстреп привело к созданию 1000 выборок (как правило, это наименьшее число воссоздаваемых копий), подобных исходной, и позволило получить выборочное распределение средних значений. Визуальное представление о распределении этих средних дают графики в) и г) рис. 1. Здесь явно налицо отличие распределения как квантилей, так и частот полученной «растиражированной» совокупности от исходных данных графиков а) и б). Этот факт вполне объясняется положениями Центральной предельной теоремы о том, что сумма достаточно большого количества слабо зависящих случайных величин имеет распределение, близкое к нормальному. Таким образом, бутстрепирование исходной эмпирической совокупности позволяет осуществить ее статистический анализ корректным образом и получить обоснованные оценки ее параметров.

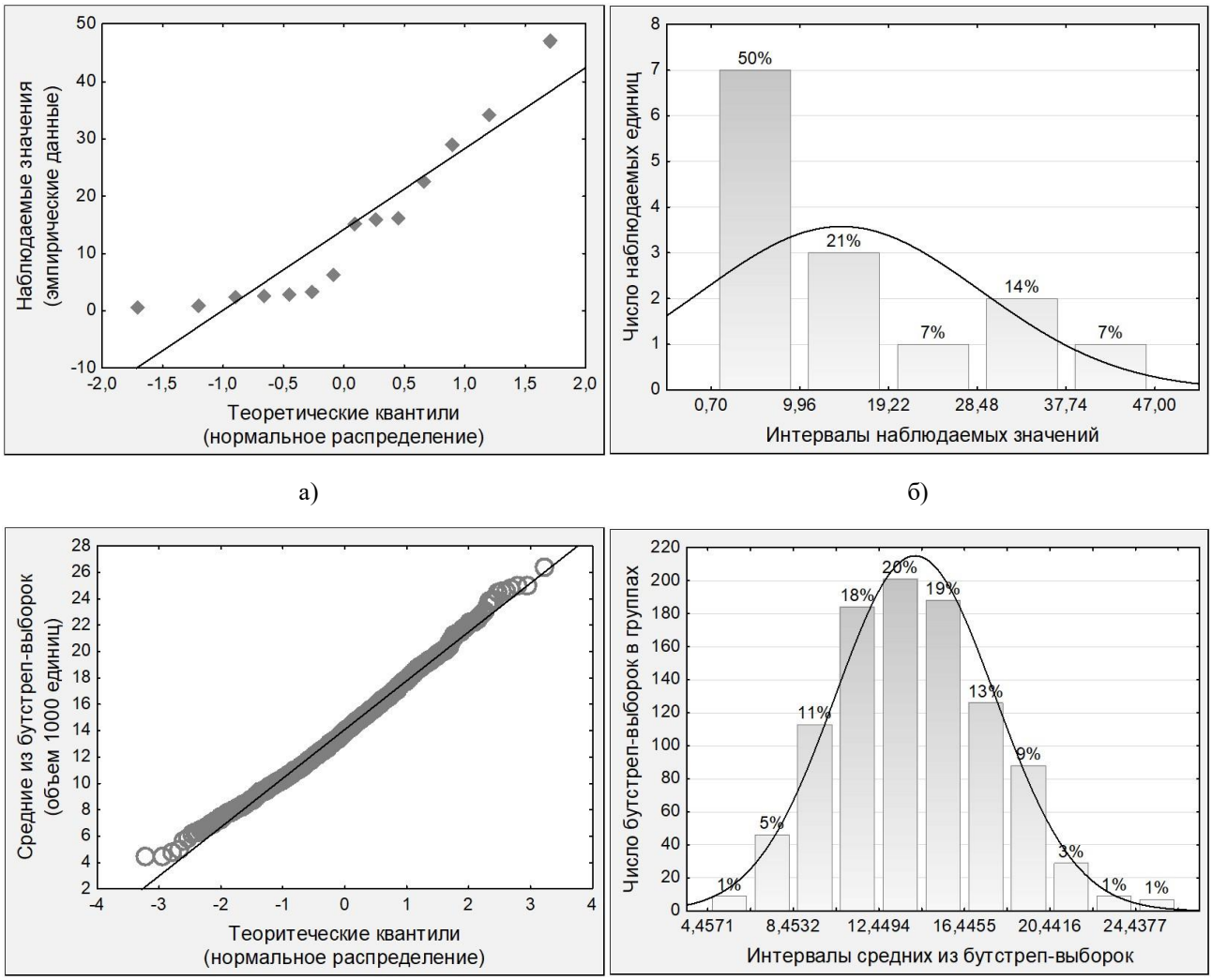

Рис. 1. Графики исходных значений и их бутстреп-выборок на примере показателя объема отправленных грузов железнодорожным транспортом общего пользования, млн. т.

Статистики исходных данных и доверительные интервалы средних, полученные посредством применения метода бутстреп (табл. 1.) для отдельных показателей транспорта, требуют некоторого пояснения. Прежде всего, следует дать интерпретацию такому важному статистическому показателю меры рассеяния как коэффициент вариации. Он представляет собой отношение среднеквадратического (или, как принято выражаться в математической статистике, стандартного) отклонения к среднему арифметическому и имеет определенное пороговое значение: если оно ниже 33 процентов, то анализируемая совокупность считается однородной (т. е. отдельные ее единицы кардинально не отличаются между собой по значениям статистического признака), а полученные характеристики (главным образом, средняя) вполне адекватными, внушающими доверие оценками; и, разумеется, наоборот, если порог в 33 процента превышен, то оценки вызывают известные сомнения в их правдивости.

Ни один из анализируемых показателей статистики транспорта не отмечен значением коэффициента ниже порогового, и, следовательно, исходные средние уровни мало пригодны для того, чтобы им верить. Это обстоятельство диктует необходимость перехода от ненадежных точечных оценок к оценкам интервальным, полученным, в настоящем случае, на основе методов, использующих результаты применения метода бутстреп.

Таблица 1

Средние уровни и бутстреп-характеристики показателей статистики транспорта

Показатели	усредненное значение	коэффициент вариации, %	метод основных интервалов
Отправлено грузов железнодорожным транспортом общего пользования, млн. т.	14,214	101,7	6,271 – 21,108
Отправлено пассажиров транспортом общего пользования, тыс. чел.	5028,2	100,2	2243,0 – 7302,2
Плотность железнодорожных путей на конец года, км путей на 10000 кв. км территории	160,3	37,4	129,9 – 190,0
Перевозки грузов автомобильным транспортом, млн. т.	22,314	74,5	13,078 – 29,875
Перевозки пассажиров автомобильным транспортом, млн. чел.	175,5	69,2	111,9 – 234,4
Плотность автомобильных дорог общего пользования с твердым покрытием, на конец года 1 км пути на 1000 кв. км территории	266,2	37,6	214,3 – 317,3

В процессе анализа были исчислены границы доверительных интервалов по всем трем рассмотренным ранее методам (процентилей, основных интервалов и с использованием t-критерия). Наиболее внятные результаты дал метод основных интервалов (его числовые значения представлена в табл. 1.). Эти интервалы строились с 5 %-ным уровнем значимости, т. е. наложенные ограничения квантилей бутстреп-распределения составляли 0,025 и 0,975 ранжированного ряда из единиц совокупности средних, полученной методом бутстреп. При изучении интервалов легко заметить, что их границы весьма и весьма широки. Это обстоятельство, в целом, достаточно однозначно характеризующее аналитическую ценность таких интервалов, объясняется как высоким уровнем уверенности в их надежности (вероятность накрытия полученными интервалами истинных значений параметров составляет 0,950), так и малым объемом исходных данных (поскольку число единиц выборочной совокупности обратно пропорционально величине меры рассеяния ее значений). Именно применение метода бутстреп дало возможность получить интервальную оценку неизвестных параметров (в настоящем случае, значения средней) и составить определенное представление о пределах их вариации.

Продолжением анализа стало статистическое оценивание гипотез о тождественности выборок исходных данных путем сравнения их средних. Целью заключалась в том, чтобы выяснить, существенны ли различия в уровнях аналогичных показателей деятельности и инфраструктуры в разрезе видов транспорта: железнодорожного и автомобильного. В частности, сопоставлению были попарно подвергнуты выборки по следующим индикаторам: а) отправлено грузов железнодорожным транспортом и перевозки грузов автомобильным транспортом; б) отправлено пассажиров железнодорожным транспортом и перевозки пассажиров автомобильным транспортом; в) плотность железнодорожных путей и плотность автодорог общего пользования с твердым покрытием.

Реализация проверки таких гипотез потребовала применения рандомизированного теста, алгоритм которого был описан выше. Роль статистики, критическое значение которой выступало бы критерием сходства или различия, исполнил ряд показателей, основанных как на традиционной t-статистике Стьюдента, так и на абсолютных или относительных разностях базовых статистик: суммы значений, их средней и медианы. В целом, результаты, полученные этими методами, дали сходные результаты. Нулевая гипотеза о несущественности значений средних не была отвергнута в пользу альтернативной гипотезы при 5 %-ном уровне значимости. Проверка на основе доверительных интервалов подтвердила сделанные выводы: границы построенных интервалов включали нулевое значение.

Заключение. Применение в статистическом анализе численных методов, к которым, в частности, относят бутстреп, рандомизацию, методы Монте-Карло, позволяют решить традиционную проблему нехватки эмпирических данных и, как следствие, невозможности оценить параметры и проверить гипотезы посредством традиционных статистических методов. Проведенный анализ отдельных показателей статистики транспорта убедительно продемонстрировал эффективность использования алгоритмов бутстрепирования и рандомизационного тестирования, позволил сделать на основе малой выборки объективные научно обоснованные выводы. В перспективе эти методы, в силу своей универсальности, могут быть распространены на анализ других направлений предметной области, в частности, для описания и прогнозирования технических и технологических процессов, связанных с организацией перевозочного процесса, автоматикой и телеметрией, путевым хозяйством и пр. (материалы статьи: текст, графики, исходные данные и скрипты программ доступны по ссылке https://github.com/karyshev63rus/vestnik_transporta_bootstrap).

СПИСОК ЛИТЕРАТУРЫ

1. Efron B. Computers and the theory of statistics: thinking the unthinkable // SIAM Review. 1979a. V. 21, № 4. P. 460-480.
2. Шитиков В.К., Розенберг Г.С. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. – Тольятти: Кассандра, 2013. – 314 с.
3. Практическая статистика для специалистов Data Science: Пер. с англ. / П. Брюс, Э. Брюс, П. Гедек. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2021. – 352с.: ил.
4. Efron B., Tibshirani R.J. An introduction to the bootstrap. N.Y.: Chapman & Hall, 1993. 436 p.
5. Регионы России. Социально-экономические показатели. 2021: Стат. сб. / Росстат. – М., 2021. – 1112 с.