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 новых данных 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 из исходной выборки , его регистрация и возвращение обратно в выборку;
2. Повторение пункта первого раз;
3. Определение среднего для повторно отобранных значений;
4. Повторение раз этапов 1–3;
5. На основе полученного распределения совокупности из 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 : медиану, стандартное отклонение и пр. Стандартная ошибка такой характеристики для выборок бутстрепа является ее стандартное отклонением [4]:

, где .

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, берущий в качестве границ доверительного интервала квантили бутстреп-распределения ;

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

;

в) метод, основанный на использовании -критерия:

,

где – критическое значение -го квантиля распределения Стьюдента .

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 и путем сравнения их средних (нулевая гипотеза : различия случайны и, значит, выборки извлечены из одной и той же генеральной совокупности, альтернативная : отличия носят неслучайный характер). Для этого использовался реализующий процесс Монте-Карло рандомизационный тест, имеющий такой алгоритм (подробно описанный в [2]):

1. C целью оценки значимость различий выборочных средних и двух групп данных со стандартным отклонением выбирается некоторый статистический критерий со статистикой (например, -статистика Стьюдента: ;
2. Исчисляется наблюдаемое значение этой статистики для исходных сравниваемых выборок;
3. Некоторое число раз ( раз >1000) выполняются в цикле:

- объединение данных из обеих выборок и перемешивание их случайным образом

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

- вычисление тестовой статистики для полученных рандомизированных данных

- увеличение переменной-счетчика в том случае, если

1. Производится расчет относительной частоты (), с которой величина превышает значение , что соответствует оценке вероятности того, что случайная величина примет значение, большее чем . При принимается нулевая гипотеза о равенстве средних, в ином случае альтернативная – .

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). Подобная сведения, конечно же, имеют место в специализированных статистических сборниках, но там они представлены в целом по стране.

**Результаты и обсуждение.**

**Заключение.**

**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.