

Курочкин Александр Владимирович

*канд. физ.-мат. наук, доцент, ОУП ВО «Академия труда и социальных
отношений», РФ, г. Москва*

ОЦЕНКА ДОСТОВЕРНОСТИ ВЫЧИСЛЕНИЯ ХАРАКТЕРИСТИК СТАТИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ В ТАБЛИЧНЫХ ПРОЦЕССОРАХ

Аннотация. В статье рассматриваются возможности определения уровня достоверности вычисления характеристик статистических распределений при обработке данных табличными процессорами в условиях недостаточно высокой выборки и сложной формы распределений. Приведены результаты статистического моделирования с применением метода классического (непараметрического) бутстрапа.

Abstract. This article examines the feasibility of determining the reliability of statistical distribution characteristics when processing data using spreadsheets under conditions of insufficient sample size and complex distribution shapes. The results of statistical modeling using the classical (nonparametric) bootstrap method are presented.

Ключевые слова: табличный процессор, бутстрап, статистическое распределение, медиана, уровень достоверности.

Keywords: spreadsheet, bootstrap, statistical distribution, median, level of confidence.

Современные табличные процессоры – MS Excel, а также некоторые альтернативные программы, например, Libre Office Calc, WPS Office Sheets, имеют обширный набор статистических функций и позволяют определять многие характеристики распределений – среднее значение, медиану, моду, различные виды дисперсии и среднеквадратичного отклонения, эксцесс, асимметричность, коэффициент корреляции и др., но не дают в общем случае ответа на вопрос корректной оценки погрешности вычисления той или иной характеристики, что связано с несколькими основными причинами:

1. Отличие статистики от нормального закона распределения при недостаточном объеме выборки;
2. Наличие корреляции между отдельными характеристиками (например, между средним значением и медианой);
3. Сложная форма распределений (асимметрия, многомодальные распределения).

В данной работе рассматривается оценка достоверности определения статистических характеристик распределений на примере соотношения медианы и среднего значения для асимметричных статистических распределений в условиях недостаточно высокой исходной выборки, хотя используемый метод может быть применен и к другим статистическим характеристикам. Среднее значение и медиана распределений (а также их разность и отношение) широко используются в самых различных прикладных задачах — в финансово-экономической сфере (распределения зарплат, доходность инвестиций, ценообразование), контроле качества на производстве, в медицинской статистике, социальных опросах и маркетинге и в других областях (см, например, [1-4]).

Для иллюстрации постановки задачи рассмотрим модельный пример исходного распределения (предположим, что это распределение доходов сотрудников некоторой гипотетической фирмы за определенный период), см. рис. 1. Объем выборки 150, минимальные и максимальные значения по выборке 62,5 и 307,3 тыс. руб. соответственно.



Рис. 1. Пример исходного распределения

Использование в MS Excel (или в альтернативных программах, указанных выше) статистических функций: СРЗНАЧ, СТАНДОТКЛОН.В, МЕДИАНА а также СКОС (асимметрия распределения) дают для рассматриваемого распределения значения, приведенные в таблице 1.

Таблица 1

Объем выборки	150
Среднее значение	143,3
Выборочное стандартное отклонение среднего	47,5
Ошибка определения среднего значения	3,9
Медиана	135,9
Разность среднего и медианы	7,4
Асимметрия исходного распределения	0,85

Принято считать, что для одномодальных распределений с положительной асимметрией медиана часто (но не всегда) меньше среднего значения. Но на основе только данных из таблицы 1 это не может быть признано статистически достоверным уже даже исходя из значения ошибки среднего, определенной как

$$\frac{s}{\sqrt{n}} = 3,9 \quad (1)$$

где $s=47,5$ – выборочное стандартное отклонение, $n=150$ – объем выборки. Важно также, что еще существует ошибка определения медианы, и табличные

процессоры не дают на нее ответа. В работах по математической статистике показано, что ошибка определения медианы равна

$$\sqrt{\frac{\pi}{2}} \frac{s}{\sqrt{n}}, \quad (2)$$

т.е., приблизительно в 1,253 раза превышает ошибку среднего значения (1), но это только в предположении нормального распределения (см, например, [5]). Далее будет показано заметное отличие для нашего модельного примера. Следует также учитывать, что между значениями выборочного среднего и медианы имеется корреляция, что затрудняет вычисление суммарной ошибки определения разности среднего и медианы.

Таким образом, учитывая только ошибку определения среднего значения и принимая во внимание перечисленные выше факторы неопределенности, мы не можем средствами табличного процессора ответить на вопрос о достоверности превышения среднего значения над медианой и, в общем случае, получить уровень достоверности определения статистических характеристик. Особенно актуально это будет для распределений с существенно меньшей выборкой и (или) более сложной формой, чем на рис. 1.

В данной работе для корректной оценки достоверности характеристик статистических распределений в условиях малой или средней выборки будет использован метод бутстрапа [6] (от англ. bootstrapping – в данном контексте букв. «самонастройка» или «самообеспечение»). Метод бутстрапа использует технологии статистического моделирования и имеет несколько разновидностей, среди которых наиболее распространены две:

- классический (непараметрический) бутстрап или метод многократной генерации «выборок с возвращением», при этом не предполагается какое-либо теоретическое распределение исходных данных;
- параметрический бутстрап, который основан на предположении о виде распределения исходных данных (например, при детектировании количества частиц от радиоактивного источника в единицу времени

следует предположить распределение Пуассона или нормальное распределение, в зависимости от интенсивности потока частиц).

Актуальность применения метода бутстрапа будет продемонстрирована на примере рассмотренной выше выборки из 150 значений (рис. 1). Для данного примера мы не можем предположить какое-либо определенное теоретическое распределение и поэтому используем метод классического (непараметрического) бутстрапа.

Приведем алгоритм вычислений для данного метода, основанный на статистическом моделировании.

1. Задается количество реплик N , т.е., повторений (для данного примера выбрано $N=50000$, так как реализованная программа работает достаточно быстро, но можно использовать и меньшее значение количества реплик).
2. Затем для каждой реплики создается новая псевдовыборка при помощи извлечения случайного элемента из исходного массива данных «с возвращением», то есть, после каждого извлечения элемент возвращается обратно в исходный массив и может быть выбран снова в следующих репликах (один и тот же элемент исходных данных может попасть в выборку несколько раз и сохраняется как элемент нового массива псевдовыборки).
3. Для массива из N псевдовыборок определяется необходимая статистическая характеристика (среднее значение и медиана, их разность, асимметрия, при необходимости и другие характеристики).

Алгоритм может быть реализован на различных языках программирования. В данной работе была использована система программирования VBA (Visual Basic for Application) [7], что вполне логично, поскольку данная система встроена в MS Excel а также и в альтернативные табличные процессоры, указанные в начале статьи. Время вычислений для рассматриваемого в статье примера на обычном персональном компьютере составило несколько секунд для 50000 реплик (в зависимости от числа анализируемых характеристик).

На рис. 2 приведены распределения среднего и медианы для рассматриваемого примера (в тыс. руб), полученные в результате работы программы для 50000 реплик, для наглядности распределения приведены в виде графиков. Видно, что распределения имеют существенно различную форму, но только из этих данных нельзя сделать вывод об уровне достоверности превышения среднего значения над медианой (как указывалось выше, имеет значение корреляция значений медианы и среднего и форма распределений). Коэффициент корреляции между медианой и средним также был получен (см. в конце статьи).

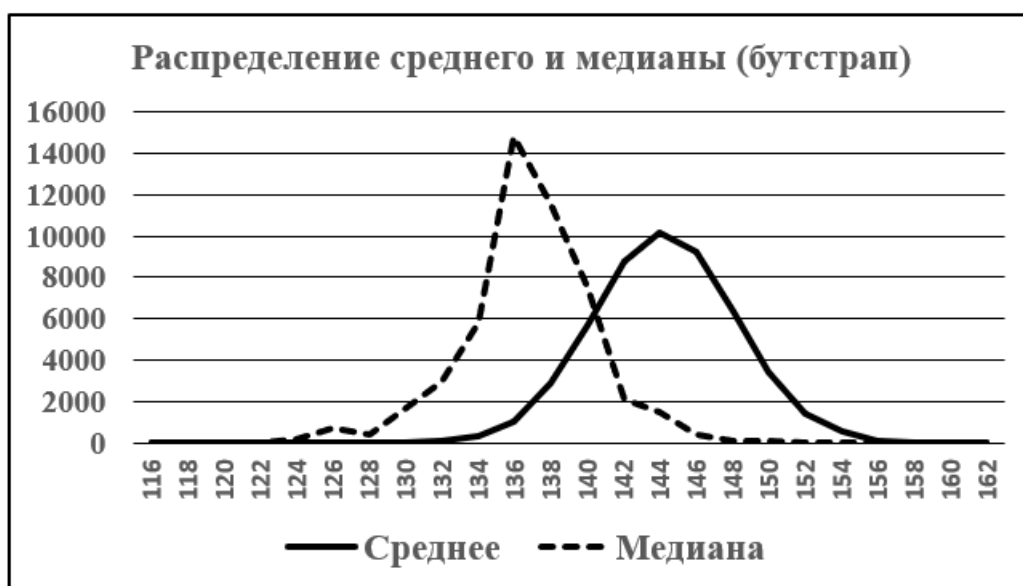


Рис. 2. Распределение среднего и медианы для 50000 реплик

Рис. 3 иллюстрирует распределение разностей среднего и медианы для 50000 реплик, полученные в результате применения бутстрапа. Распределение близко к нормальному, но, тем не менее, имеет отличия. Вычисление асимметрии дает значение $0,14 \pm 0,01$, что связано с «хвостами» распределения, поэтому уровни достоверности следует определять с помощью функции ПРОЦЕНТИЛЬ из полученного массива данных (50000 реплик). Достоверность положительного значения разности среднего и медианы была определена на уровне 99,8 %. Этот результат нельзя было получить только на основании данных из таблицы 1 для исходного распределения (рис. 1), не применяя технологию бутстрапа. Также следует отметить, что в результате применения

бутстрапа отношение ошибки медианы к ошибке среднего равно 0,92, а не 1,253, что было бы в предположении нормального распределения, см. формулу (2).

Рис. 3. Распределение разностей среднего и медианы для 50000 реплик

В результате полученных по бутстрапу распределений среднего, медианы и разностей среднего и медианы было определено для примера исходной



выборки, показанной на рис.1:

- среднее значение по выборке: $143,3 \pm 3,9$;
- среднее значение медианы: 135,9 (выборочная стандартная ошибка 3,6);
- коэффициент корреляции медианы и среднего по бутстрапу: $0,750 \pm 0,001$;
- среднее значение разности среднего и медианы по бутстрапу: $7,33 \pm 2,74$;
- асимметрия исходного распределения $0,85 \pm 0,19$;
- достоверность превышения среднего над медианным значением: 99,8 %;
- с вероятностью 68% разность среднего значения и медианы находится в диапазоне от 4,7 до 10,0, а с вероятностью 95% в диапазоне от 2,2 до 12,5.

Таким образом, в данной работе показано, что в табличных процессорах с помощью компьютерной реализации метода классического (непараметрического) бутстрапа можно определять уровни достоверности различных статистических характеристик при недостаточно высокой выборке и сложной форме исходного распределения.

Список литературы

1. Xu Y., Sarinova, Al. Sarinova As., Lozhnikova A. V. Approaches to measuring average and median wages in Russia and abroad: data sources and users // Вестник Томского государственного университета. Экономика. - 2023. № 62. – С. 19–31.
2. Шредер О.В., Горячев Д.В., Меркулов В.А. Основные принципы расчета необходимой численности участников клинических исследований // Regulatory Research and Medicine Evaluation. – 2025. Vol. 15, No. 1, – С. 92–103.
3. Смирнова Н.А., Копылов Г.М., Борисенко С.В. Применение статистических методов управления качеством в производстве пищевых продуктов // Евразийский Союз Ученых (ЕСУ) № 6, Технические науки, – 2014. – С. 120–122.
4. Курочкин А.В. Компьютерное моделирование характеристик бимодальных статистических распределений в экономических задачах // Вестник Московского финансово–юридического университета. – 2021, № 4. – С. 216–225.
5. Miller Steven J. The Probability Lifesaver: Order Statistics and the Median Theorem. [Электронный ресурс] - URL: https://web.williams.edu/Mathematics/sjmillier/public_html/probabilitylifesaver/supplementalchap_mediantheoremmandorderstatistics.pdf (дата обращения: 26.07.2026).
6. Chernick Michael R. Bootstrap Methods: A Guide for Practitioners and Researchers. Second Edition. A John Wiley & sons, inc., publication, 2008. – 369 p.
7. Комолова Н. В., Яковлева Е. С. Программирование на VBA в Excel.– СПб.: БХВ-Петербург, 2024. – 512 с.