УДК 519.2; 331.2

СПОСОБ УСТОЙЧИВОГО ОЦЕНИВАНИЯ, ИСПОЛЬЗУЮЩИЙ НЕРАВЕНСТВО ЧЕБЫШЕВА, И ЕГО ПРИЛОЖЕНИЕ К АНАЛИЗУ ДОХОДОВ В РОССИИ

В.Л. Чечулин,
Пермский государственный национальный исследовательский университет

В.И. Грацилев,
Пермский государственный национальный исследовательский университет

Описывается реализация метода устойчивого оценивания, основанного на неравенстве Чебышева. Произведено сравнение стандартных методов оценивания, в том числе устойчивых, требующих знания о типе функции распределения анализируемых данных, с предложенным методом, не требующим знания о виде функции распределения. Подчеркивается, что при стремлении извлечь максимум информации из выборки наблюдений априорные предположения о типе распределения наблюдаемой выборки направляют анализ данных по ложному пути, в связи с чем необходимо пользоваться методами, не зависящими от распределений. Дается описание математических процедур метода устойчивого оценивания, использующего неравенство Чебышева для вычисления весов наблюдений выборки. Проведено сравнение оценок положения и масштаба, полученных как с помощью стандартных методов, так и предлагаемого, для одномерных данных. На основе вычислительных экспериментов показано, что по устойчивости предлагаемый способ устойчивого оценивания сравним с медианным оцениванием. В качестве прикладного примера рассмотрено приложение метода устойчивого оценивания для анализа экономических данных о среднедушевом денежном доходе в РФ с 2001–2011 гг.

Ключевые слова: *статистическое оценивание; устойчивые оценки; робастность; методы, не зависящие от распределений; неравенство Чебышева; взвешивание наблюдений с использованием неравенства Чебышева; анализ доходов населения.*

ВВЕДЕНИЕ

В процессе получения данных возникают ошибки, шумы, а также всякого рода помехи, которые вносят возмущения в оценки параметров выборки. Требуется такие шумы фильтровать. В связи с этим возрастает роль устойчивых методов оценивания параметров выборки.

Как писал Хампель, «Статистика – это одновременно искусство и наука извлечения полезной информации из данных, полученных в результате наблюдений» [5]. Один из эффективных способов добыва-

ния такой информации связан с использованием параметрических стохастических моделей. В основе «классического подхода» лежат строгие стохастические модели, в отношении которых было установлено, что используемые ими предположения показывают действительность приукрашенной. Кроме того, обоснованность применения процедур, связанных со стохастическими моделями, и их корректность гарантируются только при условии полного соответствия сделанным предположениям. Позже возникла новая область исследований – непараметрическая статистика (критерии сравнения выборок), методы которой стали популярны при решении прикладных задач. Непараметрические статистики предназначены для сравнения выборок между собой, но не позволяют оценивать положение и масштаб выборки. Часть задач в подобной постановке находила приемлемое решение, но за параметрическими моделями сохранялась их особая роль. Это объясняется тем, что с помощью параметрических моделей наиболее полно учитывается информация, содержащаяся в наборе данных, а также тем, что диапазон применимости параметрических моделей в приложениях шире, особенно в сложных ситуациях [6].

Устойчивое оценивание

Устойчивая (робастная) статистика соединила в себе достоинства параметрических и непараметрических подходов. В ней, как средство представления информации, используются параметрические модели и применяются те процедуры, зависимость которых от допущений, вызванных этими моделями, не критична.

Как отмечается у Хампеля, ученые-статистики опасались получить ошибочные результаты, применяя неадекватные предположения. До анализа данных, до использования классических статистических процедур, ученые-статистики исправляли резко выделяющиеся наблюдения или даже исключали их. «Формальные критерии, созданные для выделяющихся наблюдений, дали толчок развитию теории обнаружения ошибок, причем ее методы как формальные, так и неформальные, основываются именно на таком подходе. Вместе с тем, наряду с комбинацией обнаружения, исправления и классической обработки, представляющей собой робастную процедуру, существуют и другие методы, позволяющие достигать лучших результатов» [5].

Впервые теоретический подход к проблеме робастности в статистике был предложен Хьюбером в 1964 году. «Робастность означает нечувствительность к малым отклонениям от предположений» [6]. Отклонениями от предположений могут быть ошибки детектора, регистрирующего наблюдения, попытки «подогнать» выборку до того, как она попадет в статистику, ошибки оформления, опечатки и др. Например, наиболее робастной оценкой параметра сдвига закона распределения является медиана [5]. «Помимо непосредственно «бракованных» наблюдений также может присутствовать некоторое количество наблюдений, подчиняющихся другому распределению. Ввиду условности законов распределений (так как это модели описания) сама по себе выборка может содержать некоторые расхождения с идеалом» [5].

Однако параметрический подход настолько доказал свою простоту и целесообразность использования, что не следует от него отказываться. Поэтому возникла необходимость изменения старых моделей так, чтобы они подходили к новым задачам.

Следует учитывать, что отбракованные наблюдения нуждаются в отдельном, более пристальном внимании. Наблюдения, которые кажутся «плохими» для одной гипотезы, могут вполне соответствовать другой гипотезе. Резко выделяющиеся наблюдения не всегда являются «браком». «Одно такое наблюдение для генной инженерии, к примеру, стоит миллионов других, мало отличающихся друг от друга» [5].

Для ограничения влияния неоднородностей, либо для их исключения, существует множество различных подходов. Среди них выделяются четыре основных

пути решения проблемы:

– сгруппировать данные, не отбраковывая отдельные наблюдения, таким образом значительно снизив возможность порчи выборки отдельными выпадами. После чего с достаточной степенью уверенности пользоваться классическими методами статистики:

– отслеживать выбросы непосредственно в процессе анализа;

– использовать подход, основанный на функции влияния;

– фильтровать выделяющиеся наблюдения.

В работе описан еще один способ минимизации влияния шумов – взаимное взвешивание наблюдений, – способ устойчивого оценивания, основанный на неравенстве Чебышева. Этот метод не требует знаний о типе функции распределения анализируемых случайных величин и применим, если существуют математическое ожидание и дисперсия выборки.

ПРИМЕРЫ НЕКОТОРЫХ СПОСОБОВ СТАТИСТИЧЕСКОГО ОЦЕНИВАНИЯ

а) Среднее арифметическое. Оценка среднего является стандартной статистической оценкой. В математике и статистике среднее арифметическое $\overline{x}$ набора данных $x_1, x_2, \ldots, x_n$ – это сумма всех чисел в этом наборе, деленная на их количество:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + \cdots + x_n\right).$$

Хотя среднее арифметическое используется в качестве средних значений или центральных тенденций, это понятие не относится к устойчивой статистике, что означает, что среднее арифметическое подвержено сильному влиянию «больших отклонений». Примечательно, что для распределений с большим коэффициентом асимметрии среднее арифметическое может не соответствовать понятию «среднего», а значения среднего из робастной статистики (например медиана) может лучше описывать центральную тенденцию.

Пример 1. В фирме работает 20 человек, зарплата 19 из них составляет 10 000 рублей, а зарплата 20-го, руководителя, – 1 000 000 рублей. Тогда средняя зарплата – 59 500 руб., что является неадекватной оценкой средней заработной платы.

б) Медиана. Медиана x_{med} исследуемого признака определяется как его срединевероятное значение, т.е. такое значение, которое обладает следующим свойством: вероятность того, что анализируемая случайная величина окажется больше x_{med}, равна вероятности того, что она окажется меньше x_{med} [1]. Медиана определяет положение середины распределения. Медиана более устойчива и поэтому может быть более предпочтительной для распределений с тяжелыми хвостами.

Медиана определяется для широкого класса распределений (например, для всех непрерывных), а в случае неопределенности естественным образом доопределяется (см. ниже).

Пример 2. В условиях примера 1 медиана равна 10 000 (полусумма десятого и одиннадцатого, срединных значений вариационного ряда). Это означает, что, разделив работников на две равные группы по 10 человек, можно утверждать, что в первой группе каждый имеет зарплату не больше 10 000, во второй же не меньше 10 000. Эта оценка достаточно адекватна.

в) **Усеченное среднее.** Отбрасывание крайних наблюдений выборки позволяет вычислять усеченное среднее – среднее значение, рассчитанное для выборки, усеченной с помощью наперед заданного процента усечения.

Например, если процент усечения равен 10, то из заданного распределения удаляется 10 % значений, наиболее сильно отличающихся по абсолютному значению от среднего арифметического заданного распределения. Причем удаляются 5 % наименьших и 5 % наибольших значений.

Пример 3. В условиях примера 1 при усечении 10 % наблюдений (отбрасывании минимального и максимального из 20 наблюдений) получается, что средняя зарплата на предприятии 10 000 руб. Такое значение зарплаты совпадает с медианой.

Усеченная оценка является устойчивой, однако удаление части выборки не всегда оправданно, поскольку в некоторых случаях сильно выделяющиеся данные не являются ошибочными и их следует учитывать при построении средних оценок.

СПОСОБ УСЕЧЕНИЯ ВЫБОРКИ ПО МИНИМИЗАЦИИ ДИСПЕРСИИ

Для сравнения с рассматриваемой далее устойчивой оценкой, использующей неравенство Чебышева, кроме стандартных способов оценивания применялся метод усечений, минимизирующий дисперсию.

Пусть имеется массив значений, в котором присутствует шум. Необходимо устранить шум в выборке, удаляя некоторые элементы массива. Определение элементов, которые надо удалить, происходит следующим образом:

1. Строится массив дисперсий S, такой что:

$$S = \{S_i(X_i) | X_i = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\}, \quad i = \overline{1,n}.$$

2. В массиве дисперсий S ищется минимальная $S_j = \min\{S_i(X_i)\}$, ее номер j соответствует элементу исходного массива, который вносит наибольший шум – это первый кандидат на удаление. Далее процедура повторяется до тех пор, пока не найдется точка перегиба.

Точка перегиба может определяться автоматически, с помощью численного дифференцирования:

$$y''(x_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2},$$

где h – шаг (количество удаляемых элементов, в данном случае $h = 1$), а y – это дисперсия выборки, после удаления очередной точки. Точка перегиба будет найдена, когда $y''(x_i)$ сменит знак.

Пример 4. Истинное модельное распределение – стандартное нормальное $N(0,1)$, вручную были внесены шумы: значения 3 и 4 (внешний шум подчеркнут в табл. 1).

Среднее квадратичное отклонение для этой выборки составляет 1,947, а среднее арифметическое – 0,135. Удаляя по вышеописанному алгоритму элементы, подозрительные на шум, получаем понижение отклонения (рис. 1). Нулевой точке на оси x соответствует среднеквадратичное отклонение для всей выборки. Анализируя график, можно заметить, что точкой перегиба является 3, т.е. можно удалить три значения, именно элементы, соответствующие номерам 1, 2 и 3 на рис. 1

Таблица 1

Исходные данные

Номер значения	Значение
1	0,399
2	2,893
3	–0,876
4	0,610
5	–0,743
6	–0,673
7	3,000
8	0,077
9	4,000
10	–1,551

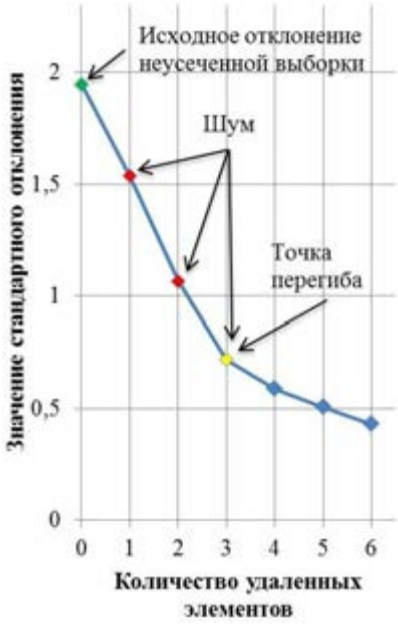

Рис. 1. Визуальное определение точки перегиба

(со значениями 4, 3 и 2,89 соответственно). После удаления этих элементов среднее квадратичное отклонение падает до 1,066, а среднее становится равным –0,706.

УСТОЙЧИВЫЕ ОЦЕНКИ, ИСПОЛЬЗУЮЩИЕ ФУНКЦИИ ВЛИЯНИЯ

При построении устойчивых оценок в западной научной школе значимыми стали предположения о типе распределения наблюдаемой величины, которая может быть зашумлена некоторыми отдельными отклоняющимися наблюдениями. Предположения об известности типа распределения наблюдаемой величины являются произвольными, т.к. в действительности, а не при моделировании, тип распределения наблюдаемых значений остается неизвестным. При таком искусственном предположении об априорной известности функции распределения рассматривается ситуация, когда имеется набор числовых наблюдений $x_1, x_2,\ldots, x_n$. В этом наборе резко выделяется одно наблюдение, для определенности x_n. Наблюдения $x_1, x_2,\ldots, x_n$ рассматриваются как реализации независимых одинаково распределенных случайных величин $X_1, X_2,\ldots, X_n$, где $X_1, X_2,\ldots, X_{n-1}$ имеют распределение $F(x)$, а X_n – распределение $G(x)$, которое «существенно сдвинуто вправо» относительно $F(x)$, например, $G(x) = F(x–A)$, где A достаточно велико [3].

В работе Хампеля [5] была введена функция влияния, которая отслеживает реакцию исследуемой статистики T на выделяющееся наблюдение, показывая зависимость от вклада x_n на оценку по всей совокупности данных, где T – это функция от некоторой выборки $X = X_1, X_2, \ldots, X_n \in X$ из распределения F с параметром $\theta \in \Theta$, т.е. T является функцией от закона F и от параметра θ. Функция влияния $IF(x, F, T)$ – это первая производная статистики T (плотности вероятности) при истинном распределении F, где x играет роль координатной оси (рис. 2).

Для того чтобы построить функцию влияния, требуется знать истинное распределение, но изначально оно не известно. Также истинное распределение может сильно отличаться от известных распределений, используемых в статистике. Поэтому выбор известного распределения в качестве истинного будет являться сильным допущением.

Обобщенно эти допущения выглядят так: постулируется, что наблюдения – это значения, принимаемые случайными величинами, которые подчиняются совместному распределению вероятностей P, принадлежащему некоторому известному классу **P**, то есть наблюдения еще не произведены, а функция распределения вероятностей предполагается уже определенной. Распределения индексируются параметром θ, принимающим значения в множестве Ω так, что

$$\boldsymbol{P} = \{P_\theta, \theta \in \Omega\}.$$

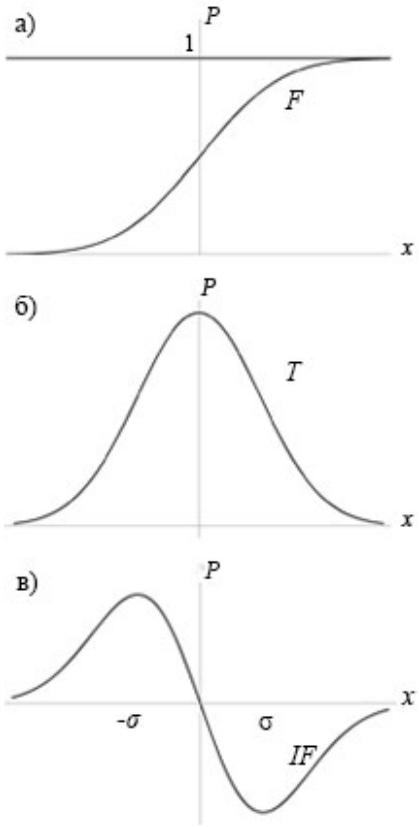

Рис. 2. Пример для нормального распределения:
а) функция распределения;
б) функция плотности вероятности;
в) функция влияния

Цель анализа состоит тогда в том, чтобы указать правдоподобное значение параметра Ω (это есть задача точечного оценивания) либо, по меньшей мере, определить подмножество Ω, о котором с известной долей уверенности утверждается, что оно содержит или не содержит истинное θ (оценивание доверительными множествами или проверка гипотез). Подобное утверждение относительно θ рассматривается как результат извлечения информации, содержащейся в данных, и используется затем как руководство к действию.

Этот подход с уточнением предположений позволяет делать сильные выводы, но делается это ценой предположений, которые становятся соответственно все более подробными и, стало быть, все менее надежными [2].

Такой поход приписывает наблюдаемой выборке «излишнюю» информацию о виде распределения, т.е. вносит «информационное возмущение» в наблюдаемую выборку, которое и является в данном случае неустранимым препятствием для получения подлинной информации о свойствах выборки. В этом заключается методологическая некорректность данного подхода в целом.

В частности, функция влияния есть вторая производная от функции распределения, и чтобы задать эту функцию влияния, необходимо знать функцию распределения, а в общем случае тип распределения неизвестен. Поэтому при исследовании совокупностей с неизвестными законами распределения, а таковыми являются практически все наблюдаемые (немоделируемые) случайные величины, необходимо использовать методы, свободные от распределения (непараметрические).

ИСПОЛЬЗОВАНИЕ НЕРАВЕНСТВА ЧЕБЫШЕВА ДЛЯ ВЗВЕШИВАНИЯ НАБЛЮДЕНИЙ

В реальных данных функция распределения неизвестна, факторы, действующие на объект измерений, в общем виде не являются аддитивными, поэтому центральную предельную теорему о сходимости к нормальному распределению применять нельзя. Это справедливо как для одномерных, так и для многомерных распределений.

В отличие от параметрических методов оценивания, неравенство Чебышева свободно от вида распределения случайной величины, поэтому далее оно используется для построения оценки, использующей взвешивание наблюдений, которые фильтруют отклоняющиеся наблюдения.

Каждому единичному измерению сопоставляется некая плотность вероятности, ему соответствующая, если известна точность измерительного инструмента. Затем, когда делается второе измерение, эта плотность вероятности корректируется. Используя такое представление о характере вероятностных закономерностей, для оценки плотности вероятности целесообразно применять неравенство Чебышева, благодаря которому становится возможным конструировать функцию взвешивания. Эта функция дает аппроксимацию суммарной плотности вероятности в виде веса наблюдения. Вес – это аналог вероятности (если сумма весов равна единице). Данный метод был предложен в начале 90-х годов и рассмотрен в статье [10].

Обоснование метода

По неравенству Чебышева, для случайной величины $X : \Omega \to R$, определенной на вероятностном пространстве (Ω, F, P), с конечным математическим ожиданием μ и конечной дисперсией σ^2 имеет место соотношение:

$$P\left(\left|X-\mu\right| \geq k\sigma\right) \leq \frac{1}{k^2}. \qquad (1)$$

В первом приближении фильтрация сильно отклоняющихся наблюдений сводится к тому, что при известной дисперсии выборки (или ее оценке) можно оценить верхнюю границу вероятности сильно отклоняющихся наблюдений и присвоить им

эту оценку в качестве веса (меньшего единицы), наблюдениям же, для которых величина правой части неравенства (1) больше единицы, оставить единичный вес.

Сумма вероятностей наблюдений в выборке равна единице, поэтому веса, полученные на предыдущем шаге, следует перенормировать так, чтобы сумма их была единичной. Далее производить обычное оценивание с использованием весов наблюдений. Но это было бы возможно при некоторой предварительной известной оценке положения (μ) и масштаба (σ), которые позволили бы вычислить веса наблюдений. В качестве таких предварительных оценок (1-го приближения) для μ и σ^2 подходят обычные (неустойчивые) оценки математического ожидания и дисперсии $E(X)$ и $D(X)$, затем, после получения устойчивых оценок, организуется итеративная процедура.

В качестве первого приближения для оценки σ может быть использовано значение точности измерительного инструмента, посредством которого получается выборка наблюдений.

Интерпретация метода

Другой вариант построения неравенства Чебышева предполагает, что каждое наблюдение X_i в выборке (мощностью n) есть некоторая реализация математического ожидания μ, тогда для X_1 и оценки масштаба (например, в виде точности измерительного инструмента) выполняется оценивание весов остальных наблюдений по неравенству (1). То есть если бы математическое ожидание выборки равнялось X_1, то (при некоторой оценке σ) наибольшая вероятность появления других наблюдений была бы оценена по неравенству (1) и им были бы присвоены соответствующие веса. Поскольку наблюдения в выборке предполагаются независимыми (реализациями случайной величины), то это рассуждение повторяется для всех X_i $\left(i=\overline{1,n}\right)$, а веса каждого j-го наблюдения при i–x рассмотрениях суммируемы. Затем веса нормируются так, чтобы сумма весов равнялась единице. Для первого приближения оценка масштаба σ используется как неустойчивая оценка, если точность измерительного инструмента неизвестна. Итерационная процедура использует оценки σ, полученные на предыдущем шаге.

Построение устойчивых оценок с использованием взвешивания

Используя интерпретацию неравенства Чебышева и подход с использованием функций влияния [10], вводится некоторая «взвешивающая» функция f_0

$$f_0(x_1;x_2)= \begin{cases} f_0(x_1-x_2)=h_1^2/(x_1-x_2)^2, & |x_1-x_2|>h_1; \\ f_0(x_1-x_2)=(x_1-x_2)/h_1, & |x_1-x_2|\le h_1, \end{cases} \quad (2)$$

обладающая свойствами:

1) симметричности;

2) ограниченности;

3) убывания на бесконечности до 0 (рис. 3). В (2) h_1 интерпретируется как точность измерительного инструмента. Посредством f_0 определяются веса наблюдений выборки.

Другой вариант взвешивающей функции g_0:

$$g_0(x_1;x_2)= \begin{cases} g_0(x_1-x_2)=h_1^2/(x_1-x_2)^2, & |x_1-x_2|>h_1; \\ g_0(x_1-x_2)=1, & |x_1-x_2|\le h_1. \end{cases}$$

При использовании g_0 у всех точек будет единичный вес при учете диагонали перебора. Если при переборе исключить диагонали $g_0(x_k;x_k)$, то функция g_0 является применимой.

Для получения оценок положения и масштаба используется функция влияния (2). Для каждого наблюдения выборки x_j определяется его вес ω_i как сумма влияний $f_0(x_i;x_j)$ на наблюдение x_j наблюдений x_j:

$$\omega_i=\sum_{j=1}^{n} f_0(x_i;x_j).$$

Затем для выборки строятся обычные оценки среднего с весовыми коэффициентами:

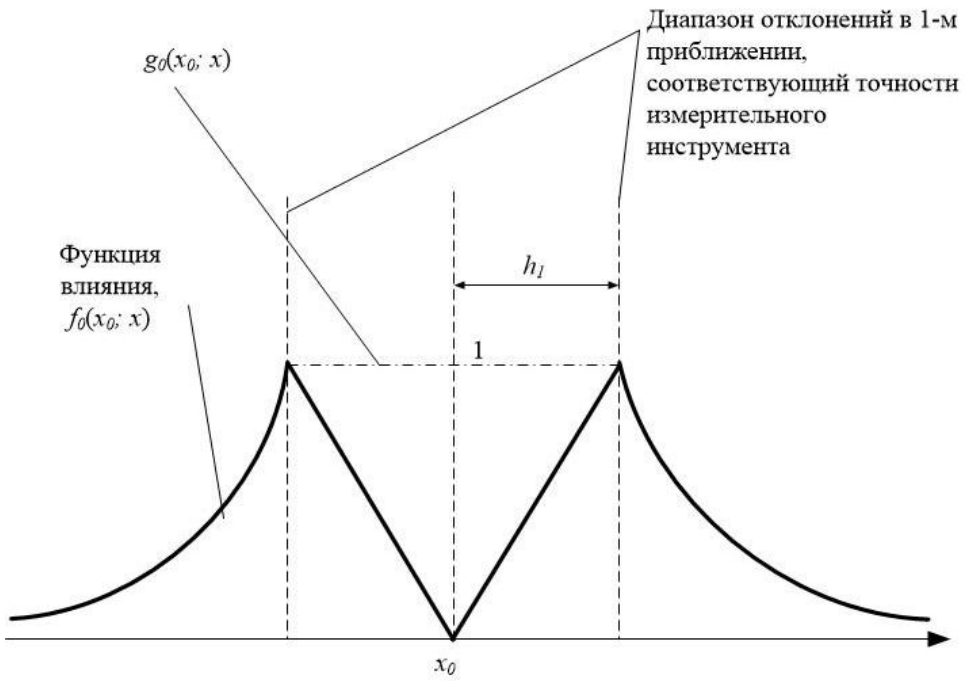

Рис. 3. Функция влияния оценки положения, 1-е приближение

$$Mu(X) = \frac{1}{\sum_{i=1}^{n} \omega_i} \sum_{i=1}^{n} x_i \cdot \omega_i.$$

Следует отметить, что выражение для стандартного отклонения дает новую оценку точности измерений h_2:

$$h_2 = Su(X) = \frac{1}{\sum_{i=1}^{n} \omega_i} \sum_{i=1}^{n} \left(x_i - Mu(X)\right)^2 \cdot \omega_i.$$

При этом, по результатам вычислительных экспериментов, последовательность h_i сходится к некоторой величине h_0, являющейся некоторым выражением точности произведенного набора измерений.

В случае применения этой процедуры взвешивания при аналогичных рассуждениях разность между любыми двумя наблюдениями из выборки $(x_i - x_j)$ есть некоторая реализация «разброса» наблюдений, тогда оценка масштаба получается независимой от оценки положения. Мера масштаба (рассеяния) выборки, приближенно совпадающая с обычной оценкой стандартного отклонения, есть сумма квадратов разностей наблюдений, умноженных на веса обоих наблюдений, деленная на сумму произведений весов [9]:

$$SRu(\mathrm{X}) = \frac{1}{2\mathrm{K}} \sum_{i}^{n} \sum_{j}^{n} \left(x_i - x_j\right)^2 \cdot \omega_i \omega_j,$$

где $\mathrm{K} = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j$, $\sum_{i=1}^{n} \omega_i = 1.$

Пример 5. Для имитационного исследования сгенерирована выборка размерностью 1 000 элементов, имеющая тяжелые хвосты, по формуле

$$\xi = tg\left(\frac{\varphi \cdot \pi}{2}\right),$$

где φ равномерно распределено на интервале (–1, 1)[1]. К полученным данным применены алгоритмы усечения (использовано усечение в 5 % наиболее выделяющихся наблюдений) и устойчивого оценивания, основанного на неравенстве Чебышева (табл. 2).

Оценки среднего в этом случае совпадают, но оценки стандартного отклонения очень сильно отличаются, даже в том случае, когда процент усечения совпадает с процентом шума. Поэтому для определения оценок необходимо использовать более устойчивые методы, например метод, основанный на неравенстве Чебышева.

[1] В этом случае генерируется распределение, близкое к нормальному, но имеющее «толстые хвосты», медиана этого распределения 0, а квартили (нижний, верхний): –1, 1.

Таблица 2

Сравнение оценок положения и масштаба

Оценка	Значение
Медиана	–0,017
Среднее	0,030
Усеченное среднее (5 % усечения)	0,015
Устойчивое среднее	0,007
Стандартное отклонение	13,520
Усеченное стандартное отклонение (5 % усечения)	3,353
Устойчивое стандартное отклонение	1,034

СРАВНИТЕЛЬНОЕ ТЕСТИРОВАНИЕ СПОСОБОВ ОЦЕНИВАНИЯ

Сравнительное тестирование способов оценивания проводилось посредством вычислительных экспериментов с моделируемыми данными. Моделирование использовало основную выборку данных, наблюдения которой заменялись на возмущающий шум [7]. Использовалась выборка объемом 1 000 наблюдений, исходные данные имеют стандартное нормальное распределение N (0,1), шум – равномерное распределение R (0,10). Шум односторонний, с размахом, в несколько раз превышающим стандартное отклонение исходной выборки. Доля шума в выборке при экспериментах увеличивается от 0 до 100 % с шагом 10 %.

По модельным данным были рассчитаны следующие оценки: среднее, медиана, среднее усеченное, среднее устойчивое (рис. 4), среднее квадратичное отклонение, среднее квадратичное отклонение по усеченным данным и устойчивое отклонение по методу Чебышева (рис. 5).

На рис. 4 видно, что среднее арифметическое растет линейно на уровне зашумления в 10 %, в то время как медиана,

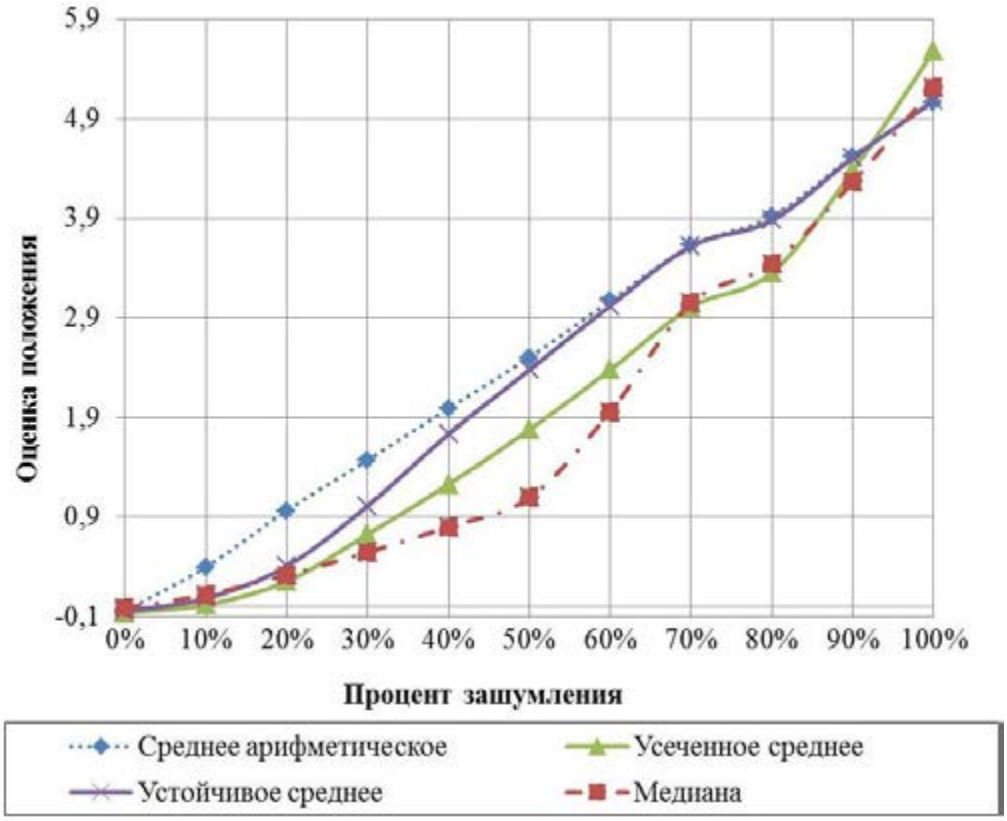

Рис. 4. Оценки положения, полученные на одномерной выборке с малым шумом

усеченное среднее и устойчивое среднее сравнимы по устойчивости с зашумляющими наблюдениями. Устойчивое среднее проигрывает усеченному среднему в силу того, что зашумленные значения не отбрасываются, они остаются в выборке, но с малым весом.

На рис. 5 видно, что наибольшее значение имеет среднеквадратичное отклонение, устойчивое отклонение показывает более реалистичную картину, в то время как наилучшим значением обладает усеченное отклонение. На уровне 10 % шумов усеченное и устойчивое отклонения примерно одинаковы.

На основании вычислительных экспериментов сделан вывод о том, что при уровне шума до 20 % на одномерной выборке устойчивая оценка положения сравнима с медианой и усеченными оценками.

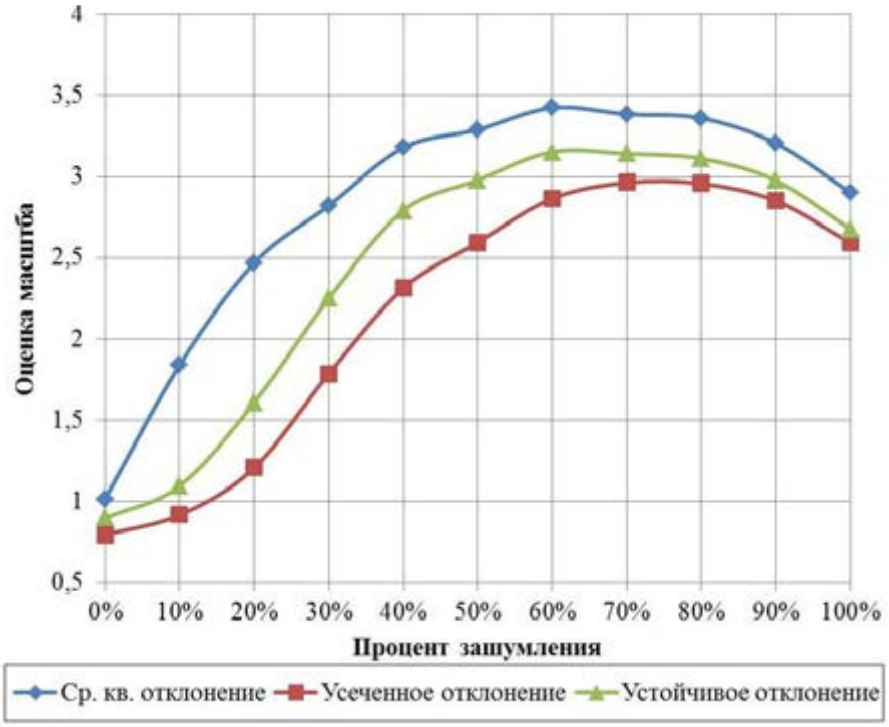

Рис. 5. Оценки масштаба, полученные на одномерной выборке с малым шумом

ПРИЛОЖЕНИЕ СПОСОБА УСТОЙЧИВОГО ОЦЕНИВАНИЯ К АНАЛИЗУ СРЕДНЕДУШЕВЫХ ДОХОДОВ

Показательные искусственные примеры анализа доходов различными оценками приводились в п. 2 (примеры 1–3). Далее для анализа доходов использован описанный выше способ устойчивого оценивания.

В качестве оценки экономических данных были рассмотрены различные оценки величины относительного среднедушевого дохода: среднее арифметическое, медиана, устойчивое среднее и мода, вычисляемая по взвешиванию наблюдений. Данный набор оценок позволяет адекватно оценить состояние уровня среднедушевого дохода населения России в период с 2001 по 2011 г.

Для исследования были использованы данные с сайта Федеральной службы государственной статистики из сборника «Россия в цифрах» [4], а именно данные по уровню жизни населения: распределение населения по величине среднедушевых денежных доходов, минимальный прожиточный минимум, величина ВВП на душу населения. В сборнике среднедушевые денежные доходы (в месяц) исчисляются делением годового объема денежных доходов на 12 и на численность населения.

Способы оценивания

В качестве исходных весов наблюдений взяты веса, соответствующие доле людей с определенным доходом (табл. 3). Веса искусственно нормировались таким образом, чтобы их сумма была равна 1:

$$\sum_{i=1}^{n} \omega_i = 1.$$

а. Среднее арифметическое

Для вычисления среднедушевого дохода необходимо просуммировать произведение средних заработных плат P_i на соответствующие веса:

$$\overline{P} = \sum_{i=1}^{n} P_i \omega_i .$$

б. Медиана

Для нахождения медианы необходимо определить вероятность получения каждого среднедушевого дохода. Далее ищутся две смежные величины дохода, между которыми заключена вероятность 0,5. Разности Δ_1 и Δ_2 (по модулю) между соответствующими вероятностями и 0,5 суммируются, далее происходит вычисление медианы по формуле

$$P_{0,5} = \frac{P_1 \cdot \Delta_2 + P_2 \cdot \Delta_1}{\Delta_1 + \Delta_2}.$$

в. Метод устойчивого оценивания посредством неравенства Чебышева

Данный метод описан выше. С помощью алгоритма, по имеющимся долям людей, получающих определенный доход, устанавливаются новые веса, которые используются в дальнейшем для определения моды и устойчивой оценки среднедушевого дохода.

г. Мода

Для нахождения моды необходимо вычислить устойчивую оценку и новые веса с помощью неравенства Чебышева. Величина среднедушевого дохода, имеющая новый наибольший вес, является модой. Веса аппроксимируют плотность вероятности, поэтому моду можно вычислять данным образом.

Анализ данных о доходах

Связь среднедушевого дохода с демографическими показателями была рассмотрена в работе [11], в частности, было показано, что низкая рождаемость обусловлена низкими доходами населения; поэтому целесообразно оценивать наиболее вероятный доход методом устойчивого оценивания, фильтрующим сверхдоходы незначительной части населения.

Алгоритмы устойчивого оценивания были применены к реальным данным. Была проведена оценка благосостояния населения России в период с 2001 по 2011 годы (см. табл. 3).

С помощью методов, описанных выше, по этим данным были проведены расчеты и получены следующие результаты: среднее арифметическое: 22 953 руб.; медиана: 12 863 руб.; устойчивое среднее: 9 995 руб.; мода: 8 500 руб. (см. рис. 5).

При этом 72 % людей имеют доходы меньше, чем средняя величина, таким образом, средний среднедушевой доход не является адекватной оценкой. Даже медиана, доходы ниже которой имеют 50 % населения, не совсем адекватна. Более адекватными являются мода (наиболее вероятный доход) и устойчивое среднее.

Таблица 3

Распределение населения по среднедушевым денежным доходам за 2011 год

Группы среднедушевых доходов, руб.	Средний доход по группе, руб.	Процент людей с таким доходом	Вычисленный вес
До 3 500,0	1 750	2,8	0,031562
3 500,1–5 000,0	4 250	4,6	0,126385
5 000,1–7 000,0	6 000	8,1	0,232175
7 000,1–10 000,0	8 500	13,5	0,267504
10 000,1–15 000,0	12 500	19,8	0,206205
15 000,1–25 000,0	20 000	24,8	0,103996
25 000,1–35 000,0	30 000	12,1	0,029338
Свыше 35 000,0	70 000	14,3	0,002835

При использовании устойчивого метода оценивания по данному распределению дохода была построена аппроксимация плотности вероятности и определена мода. Для 2011 года аппроксимация плотности вероятности и мода приведены на рис. 6 и в табл. 3.

Из рис. 7 видно, что все графики показывают абсолютный рост среднедушевых доходов, причем среднее арифметическое намного опережает медиану, моду и устойчивую оценку. Устойчивая оценка показывает результат, сравнимый с медианой.

Далее сравниваются уровни среднедушевых доходов по относительному доходу. Величина среднедушевого дохода, полученная каждым из методов, поделена на величину прожиточного минимума того же года (рис. 8).

Как видно из рис. 8, в 2005 году произошел провал в доходах населения, это связано с резким увеличением в том же году величины прожиточного минимума.

Начиная с 2007 года все графики показывают, что рост средней заработной платы прекратился, начинается период, когда среднедушевой доход значительно не увеличивается, а обесценивание денег продолжается, с качественной стороны это было показано в [8].

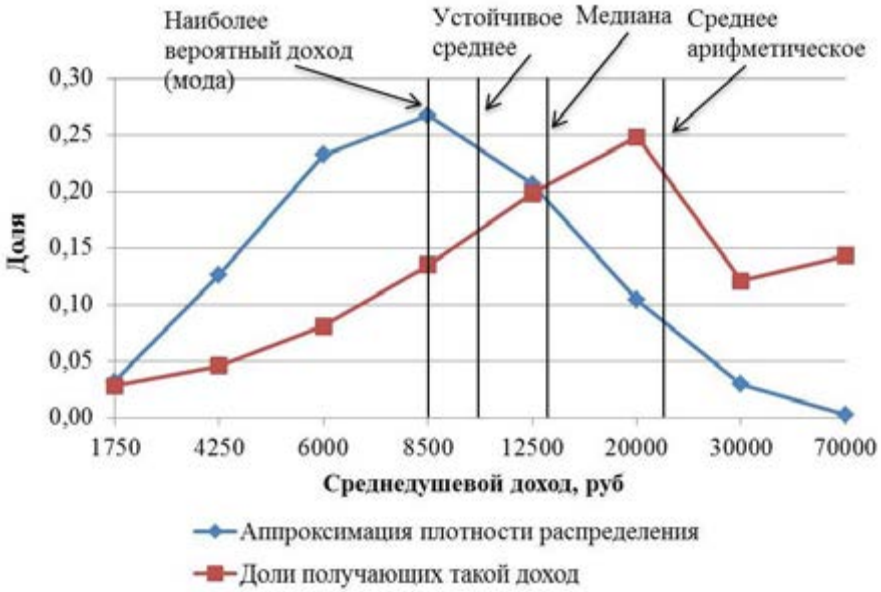

Рис. 6. График относительных долей зарплат и оценки весов (плотности вероятности, см. табл. 3)

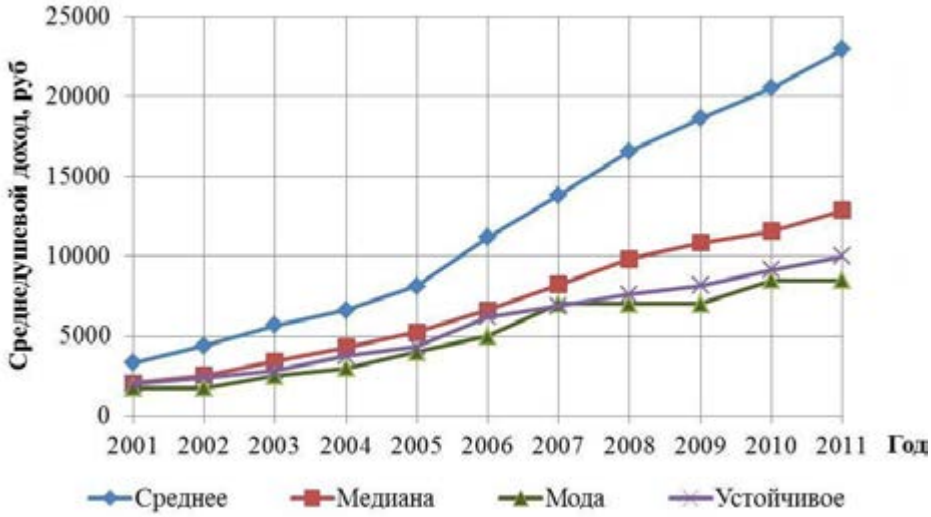

Рис. 7. Сравнение уровня среднедушевых доходов по абсолютной величине

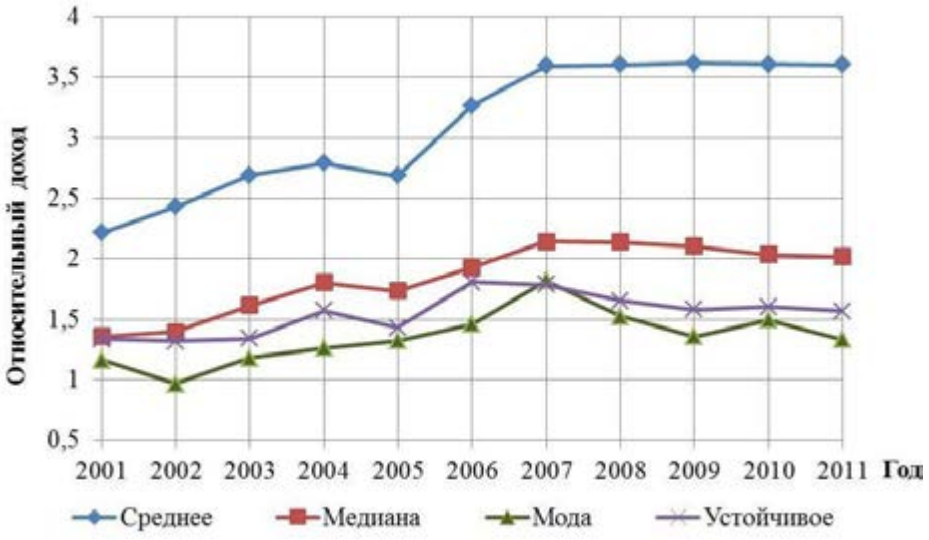

Рис. 8. Сравнение уровня среднедушевых доходов по относительному доходу (в прожиточных минимумах)

Значение моды, как и значение устойчивого среднего, находится ниже, чем значение арифметического среднего, и даже ниже, чем медиана. Это означает, что наиболее вероятные доходы для массы трудящихся низки (графики моды и устойчивого среднего представлены на рис. 6 и 7, значения – в табл. 3). Это связано с тем, что доля людей, получающих высокую зарплату, смещает среднюю арифметическую оценку вверх, ввиду ее неустойчивости.

На рис. 8 видно, что при устойчивом оценивании и по оценке моды, среднедушевой доход составляет 1,5 прожиточных минимума на человека, что, в среднем дает одного ребенка на семью. По медиане – около двух прожиточных минимумов, но медиана здесь не применима, т.к. она отражает величину доходов, ниже и выше которой по 50 % населения. Доход малоимущих учитывается в более полной мере устойчивым оцениванием, потому что оно оценивает наиболее вероятное состояние, по определению весов наблюдений. Исходя из полученных результатов получен вывод, что оценка благосостояния населения в виде среднедушевых доходов неадекватна реальному состоянию экономики. Поэтому следует использовать иные оценки: моду, устойчивое оценивание, принимая во внимание нижнюю границу доходов, определяемую минимальной оплатой труда.

Относительное измерение экономики

Экономика измерима в относительных среднедушевых доходах, отнесенных к валовому внутреннему продукту (ВВП).

В качестве относительной меры оценки обеспеченности, на фоне которой были построены графики оценок среднедушевых доходов (рис. 9–11), приведены относительные величины валового внутреннего продукта, валового накопления и расходов государственного бюджета относительно прожиточных минимумов. По этим графикам видно, что с 2008 по 2012 год экономического роста практически не наблюдалось. Это объясняется уровнем среднедушевых доходов, которые за 2007–2009 годы даже снижались вследствие роста дифференциации доходов.

Из анализа графиков на рис. 9–11 видно, что Россия имеет достаточные ресурсы для содержания и расширенного воспроизводства населения, однако цель расширенного производства не достигается в связи с дифференциацией доходов и заниженными доходами большинства населения, что было показано в предыдущих рассуждениях (см. рис. 8).

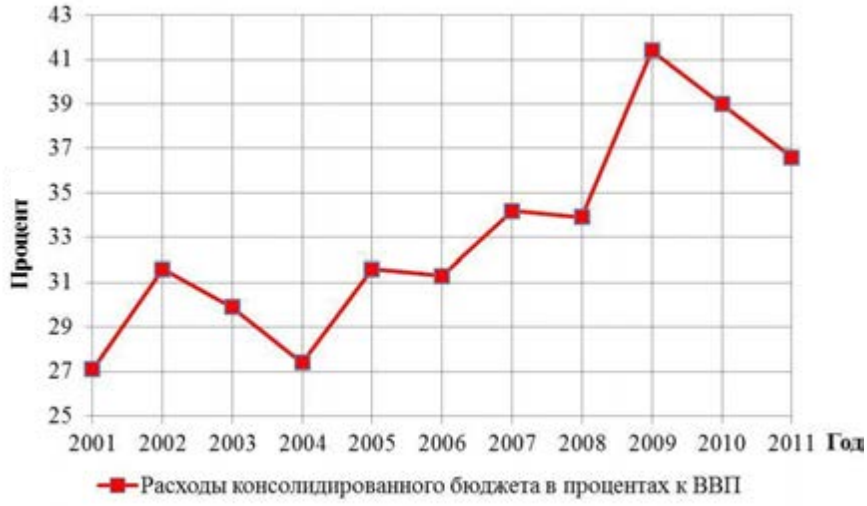

Рис. 9. Расходы консолидированного бюджета государства в процентах к ВВП

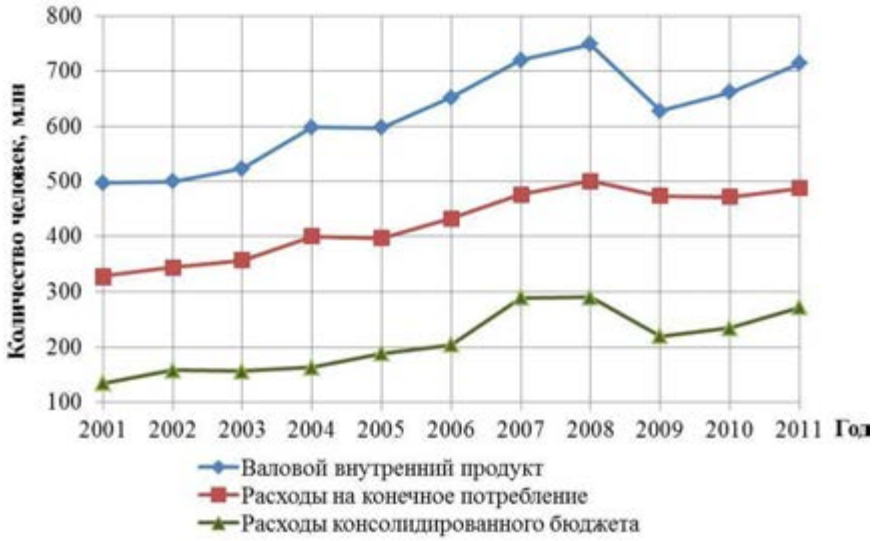

Рис. 10. Индикатор уровня экономических величин: ВВП, накопление валовое, расход государственного бюджета, выраженные в прожиточных минимумах за год (это дает количество населения, которое может жить при одинаковом доходе)

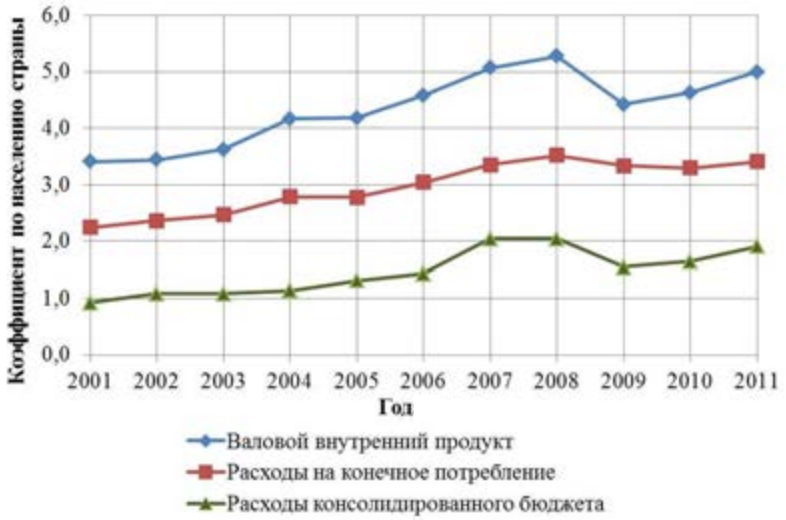

Рис. 3. Индикатор количества текущих населений страны, которые могли бы жить в России, при среднедушевом доходе, равном прожиточному минимуму

ЗАКЛЮЧЕНИЕ

Разработанный способ устойчивого оценивания основывается на неравенстве Чебышева, не зависящем от типа распределения, поэтому при его применении в оценивании параметров выборки не требуются предположения о типе распределения наблюдений, что является методологически корректным. Посредством вычислительных экспериментов показано, что устойчивое оценивание (использующее неравенство Чебышева) при односторонних шумах по устойчивости сравнимо с медианой. В приложении к анализу данных о доходах в РФ показано, что средняя величина малоадекватна ввиду того, что 72 % людей имеют доходы меньше, чем средняя величина дохода. Более адекватной является устойчивая оценка.

Таким образом, устойчивое оценивание необходимо применять в случаях, когда данные асимметричны и их исходное распределение неизвестно.

Библиографический список

1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
2. *Леман Э.* Теория точечного оценивания: пер. с англ. – М.: Наука, 1991. – С. 9-15.
3. *Лемешко Б.Ю.* Робастные методы оценивания и отбраковка аномальных измерений // Заводская лаборатория. – 1997. – Т. 63. – № 5. – С. 43–49.
4. Федеральная служба государственной статистики. Сборник «Россия в цифрах». Данные по распределению населения по уровню заработной платы, величины прожиточного минимума, ВВП на душу населения.
   URL: http://www.gks.ru/wps/wcm/connect/rosstat/rosstatsite/main/publishing/catalog/
5. *Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В.* Робастность в статистике. Подход на основе функций влияния. – М.: Мир, 1989. – 512 с.
6. *Хьюбер Дж.П.* Робастность в статистике. – М.: Мир, 1984. – 304 с.
7. *Чечулин В.Л., Грацилев В.И.* Качественное сравнение способов устойчивого оценивания // Университетские исследования, 2012 (раздел: математика)
8. URL: www.uresearch.psu.ru/files/articles/634_52153.doc
9. *Чечулин В.Л.* Об основаниях потребностного подхода к обеспечению социальной защищенности граждан // Человеческий капитал. – 2011. – № 4. – С. 72–76.
10. *Чечулин В.Л.* Об оценке масштаба (дисперсии) выборки, не использующей оценку положения (среднего) // Университетские исследования, 2011 (раздел: математика) URL: http://www.uresearch.psu.ru/files/articles/553_26764.doc
11. *Чечулин В.Л.* К обоснованию метода устойчивого оценивания посредством неравенства Чебышева // Вестник Пермского ун-та. Сер. Математика. Механика. Информатика. – 2010. – Вып. 2 (2). – С. 29–32.
12. *Чечулин В.Л., Федосов А.Ю.* Обоснование повышения минимальной заработной платы для сохранения численности населения России до 2050 года // Университетские исследования, 2012 (раздел: демография) URL: http://www.uresearch.psu.ru/files/articles/589_9685.doc

METHOD OF ROBUST EVALUATION THAT USES CHEBYSHEV'S INEQUALITY, AND ITS APPLICATION TO ANALYSIS OF INCOMES IN RUSSIA

V.L. Chechulin, V.I. Gratsilev

Perm State National Research University

The article describes implementation of the robust evaluation method, based on the Chebyshev's inequality. Standard methods of estimation, including robust ones, that require knowledge about the type of the distribution function of the analyzed data are compared with the proposed distribution-free method provided. The emphasis is laid on the fact that if the goal is to extract maximum information from the sample of observations, a priori prediction about the type of the distribution function can lead to false conclusions, so distribution-free methods are required. The description of mathematical procedures of robust estimation method using Chebyshev's inequality to calculate observations weights, is given.

Comparison of location and scale estimations, provided by standard methods vs the proposed method, is made for univariate data. Computational experiments showed that in terms of robustness the proposed method is comparable to median estimation. As an application example, robust estimation method is applied to analysis of economic data about per capita income in Russia between 2001–2011 is considered.

Keywords: statistical estimation, robust estimates, robustness, distribution-free methods, Chebyshev's inequality, weighing observations using Chebyshev's inequality, analysis of per capita income.

Сведения об авторах

Чечулин Виктор Львович, старший преподаватель кафедры прикладной математики и информатики, Пермский государственный национальный исследовательский университет (ПГНИУ), 614990, г. Пермь, ул. Букирева, 15; e-mail: chechulinvl@mail.ru

Грацилев Вадим Игоревич, студент 2-го курса магистратуры кафедры прикладной математики и информатики, ПГНИУ; e-mail: Vadim.Gratsilev@gmail.com

Материал поступил в редакцию 25.05.2015 г.