УДК 519.234.7

РОБАСТНЫЕ МЕТОДЫ ОЦЕНИВАНИЯ ЧИСЛОВЫХ ХАРАКТЕРИСТИК ВЫБОРКИ

В.Л. Шергин[1], Т.А. Мирошниченко[2]

[1] ХНУРЕ, г.Харьков, Украина, shergin@kture.kharkov.ua
[2] ХИНЭМ, г.Харьков, Украина.

Проведён анализ устойчивости числовых характеристик статистической выборки по отношению к выбросам. Предложено использовать робастные характеристики, основанные на порядковых статистиках – медиану выборки и медиану абсолютных отклонений. Проведено моделирование методом Монте-Карло, которое проиллюстрировало и подтвердило целесообразность предложенного подхода.

ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ, РОБАСТНОСТЬ, МЕДИАНА АБСОЛЮТНЫХ ОТКЛОНЕНИЙ

Введение

Одними из основных методов количественного анализа данных являются статистические методы. Основной предпосылкой применения классических методов математической статистики является гипотеза об однородности выборки. При использовании параметрических методов дополнительно предполагается, что элементы выборки следуют некоторому известному закону распределения. Так, если выборка получена из генеральной совокупности, подчиняющейся нормальному закону $N(\mu,\sigma)$, то эффективными оценками параметров сдвига (m) и масштаба (σ) являются среднеарифметическое значение и стандартное отклонение

$$m_x = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2} \qquad (1)$$

соответственно, а для оценки меры корреляционной связи применяют выборочный коэффициент корреляции Пирсона:

$$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2} \cdot \sqrt{\sum (y_i - m_y)^2}} = \frac{\sum (x_i - m_x)(y_i - m_y)}{(n-1)\cdot s_x \cdot s_y}. \qquad (2)$$

В силу сложившейся традиции при проведении количественного анализа данных параметры сдвига, масштаба и силу связи оценивают именно по формулам (1)-(2). При этом гипотезы, лежащие в их основе, либо игнорируются, либо проверяются апостериорно с помощью критериев согласия, которые на коротких выборках имеют малую мощность. В результате получаемые оценки обладают низкой устойчивостью по отношению к выбросам в выборке и к отклонению закона распределения от нормального.

1. Анализ устойчивости числовых характеристик выборки

Традиционный подход, использующий оценки вида (1)-(2), имеет ряд существенных методологических недостатков. Во-первых, как отмечают специалисты в области прикладной статистики [1], данные реальных измерений или наблюдений редко бывают распределены по нормальному закону. Гораздо чаще встречаются выборки с «тяжёлыми хвостами», то есть с существенно бульшими, чем при нормальном законе, частотами появления наблюдений, заметно отклоняющихся от средних. Для моделирования таких выборок целесообразно использовать другие законы распределения (например, распределение Стьюдента, Коши, логистическое), а для оценивания параметров (θ) применять метод максимального правдоподобия, согласно которому их находят из условия

$$\sum_i \frac{\partial \ln(P_i)}{\partial \theta} = 0 \ , \qquad (3)$$

где $P_i = f(x_i,\theta)$ – плотность вероятности закона распределения $f(x,\theta)$ при выборочном значении $x = x_i$.

Недостатками метода максимального правдоподобия являются как сложность получения ОМП-оценок (3), так и их жесткая «привязка» к выбранному закону, а также к гипотезе об однородности.

Во-вторых, выборки реальных данных зачастую бывают «засорены» как вследствие ошибок наблюдения, так и в силу внутренней неоднородности, то есть наличия наблюдений, подчиняющихся другому распределению. При этом наличие в выборке даже небольшого числа резко выделяющихся наблюдений (выбросов) способно фатально повлиять на результат статистического исследования, и значения оценок, подобных (1)-(3), могут потерять какую-либо практическую ценность.

В-третьих, существует ряд задач анализа данных, сама постановка которых отвергает гипотезу об однородности выборки. К таким задачам относятся, например, задачи классификации и распознавания, решаемые методами кластерного и дискриминантного анализа. При этом неотъемлемой частью большинства из этих методов являются процедуры стандартизации данных, то есть центрирования и нормирования:

$$z_i = \frac{x_i - m_x}{s_x} \ . \qquad (4)$$

По инерции для этих целей чаще всего используют оценки (1), хотя их применение может быть обусловлено разве что простотой получения.

При проведении количественного анализа данных экономического характера исследователь сталкивается со всеми из перечисленных ситуаций: во-первых, реальные выборки могут быть коротки и обладать «тяжёлыми хвостами», во-вторых, выборки зачастую содержат не вполне достоверные элементы, и в-третьих, обычно нет оснований считать выборки однородными.

Таким образом, основой при проведении статистического анализа выборочных данных должны быть *робастные* методы оценивания. Под *робастностью* в статистике понимают нечувствительность к различным отклонениям и неоднородностям в выборке, связанным с теми или иными, в общем случае неизвестными, причинами [2].

Основными характеристиками робастности оценки являются точка разрыва *BP* (*breakdown point*) и функция влияния *IF* (*influence function*). *BP* определяется как максимальная доля в выборке экстремальных выбросов (равных бесконечности), при которых оценка параметра остаётся конечной. Легко видеть, что оценки (1) не являются робастными, так как для них $BP = 0$.

2. Построение робастных числовых характеристик выборки

К основным методам получения робастных оценок относятся: группировка данных, усечение выборки, взвешивание наблюдений. Усечение заключается в отбрасывании некоторой доли наблюдений. Очевидно, что если отбросить p процентов наблюдений, то $BP = p/100$. При группировке данных отдельные наблюдения не отбрасываются. Разбиение на интервалы не представляет особых трудностей и даёт весьма ощутимый результат в случае, когда выборка не очень короткая, причиной неоднородности являются немногочисленные резкие выбросы, а остальные элементы выборки действительно соответствуют выбранному закону распределения. Недостатками данного метода являются неработоспособность на коротких выборках и параметрический характер метода, то есть использование закона распределения. Так, если причиной отклонения выборки от нормального является «тяжелый хвост», то группировка данных не позволяет улучшить качество оценок параметров [3]. Подобными же свойствами в смысле робастности обладает и метод усечения.

Взвешивание наблюдений заключается в преобразовании данных выборки x_i с помощью некоторой весовой функции $\psi(z)$, где z_i является стандартизованным значением i-го элемента выборки, вычисляемым согласно (4). Целью взвешивания является установление барьера на влияние резко выделяющихся (далеко отстоящих от предполагаемых значений параметров) наблюдений. В качестве весовой функции чаще всего используются биквадратная функция Тьюки $\psi(z) = \frac{z}{z^2+1}$, и функция гиперболического тангенса $\psi(z) = th(z)$. Усечение также можно рассматривать как взвешивание усечённой линейной функцией.

При использовании взвешивания эффективными (и в то же время робастными) оценками параметров являются М-оценки, являющиеся обобщёнными ОМП-оценками:

$$\sum_i \psi(x_i, \theta) = 0 . \qquad (5)$$

Для получения М-оценок, то есть для решения системы уравнений (5), необходимо применять численные методы. В этом заключается недостаток взвешивания. При этом необходимо задать начальные приближения этих оценок. В общем случае они, конечно, могут быть любыми, однако желательно подбирать их как можно точнее, то есть ближе к искомым конечным значениям.

Наиболее логичным будет взять такие начальные приближения оценок параметров, которые сами являются робастными оценками. По этой причине рекомендуется [1], [4] использовать в качестве начального приближения для параметра сдвига *медиану* выборки

$$h_x = median(X) . \qquad (6)$$

Медиана является *наилучшей* из возможных оценок параметра сдвига по критерию BP, поскольку $BP(h_x) = \frac{\lfloor (n+1)/2 \rfloor - 1}{n} = \frac{1}{2} - \varepsilon$, (где n – объём выборки), и, в то же время, эффективной М-оценкой. Для параметра масштаба такой робастной оценкой является значение, кратное медиане отклонений от медианы, называемое MAD – *median absolute deviation*:

$$MAD_x = k \cdot median|x_i - h_x| . \qquad (7)$$

Для удобства практического применения оценок (7) коэффициент пропорциональности (k) обычно выбирают таким, чтобы оценка параметра масштаба для выборки X, следующей нормальному закону $N(\mu, \sigma)$, равнялась σ. В этом случае числовое значение коэффициента равняется

$$k = 1/u_{0,75} \approx 1{,}4826 , \qquad (8)$$

где $u_{0,75}$ – квантиль нормального распределения на уровне $p = 0{,}75$.

Дополнительным преимуществом использования медианной оценки является наличие у неё содержательной (а не только статистической) интерпретации, что особенно полезно при анализе экономических данных. Так, медиана выборки (то есть значение, находящееся посредине отсортиро-

ванной выборки) равна значению показателя для среднего наблюдения: расход ресурсов среднего по частоте встречаемости процесса; доход средней семьи; рентабельность среднего предприятия и т.п. Очевидно, что такая оценка не только робастна, но и гораздо более информативна, чем среднее арифметическое тех же величин.

3. Численное моделирование робастности числовых характеристик выборки

Для проверки робастности различных оценок сдвига и разброса было проведено численное моделирование методом Монте-Карло. Исследовались две оценки меры сдвига: среднеарифметическое значение и медиана (m_x и h_x), вычисляемые согласно (1) и (6) и обозначенные как mean и median соответственно, и две оценки меры разброса: стандартное отклонение s_x и медиана абсолютных отклонений MAD_x (1) и (7), обозначенные как stdev (standard deviation) и MAD.

Рассматривались пять различных законов распределения, а именно:

(A) – стандартный нормальный закон $N(0,1)$;

(B) – вероятностная смесь двух нормальных законов $N(0,1)$ и $N(0,3)$ с весами 0,9 и 0,1 соответственно;

(C) – логистическое распределение $L(0,1)$;

(D) – вероятностная смесь двух логистических законов $L(0,1)$ и $L(0,3)$ с весами 0,9 и 0,1 соответственно;

(E) – распределение Коши $C(0,1)$.

Плотности логистического распределения $L(\alpha,\beta)$ и распределения Коши $C(\alpha,\beta)$ имеют вид

$$f(x)=\frac{e^{(x-\alpha)/\beta}}{\left(1+e^{(x-\alpha)/\beta}\right)^2} \quad \text{и} \quad f(x)=\frac{1}{\pi}\frac{\beta}{\beta^2+(x-\alpha)^2}$$

соответственно.

Распределения (B) и (D) можно трактовать как “зашумлённые”, в которых 90% наблюдений следуют основному закону ($N(0,1)$ и $L(0,1)$ соответственно), а 10% наблюдений являются выбросами.

Для каждого из законов распределения методом Монте-Карло генерировалось по 100 выборок объёмом $n=10,20,30,50,100,200,500,1000,2000,5000$ каждая. Вычислялись оценки числовых характеристик выборки $\theta=\{m_x, h_x, s_x, MAD_x\}$ и дисперсии этих оценок $D(\theta_n,\theta^*)$ относительно математических ожиданий θ^*. Для всех пяти законов $M(m_x)=M(h_x)=0$. Математические ожидания от s_x для рассматриваемых законов равны соответственно 1, $\sqrt{1.8}$, $\pi/\sqrt{3}$, $\pi\sqrt{0,6}$ и $+\infty$. Математические ожидания от MAD_x определяются из условий $k\cdot F^{-1}\left(\frac{3}{4}\right)$, где $F(x)$ – функция распределения, а $F^{-1}(p)$ – квантиль.

Поскольку $D(\theta_n,\theta^*)$ убывает обратно пропорционально n, то мерой качества полученных оценок может служить статистика [3]

$$T(n,\theta_n)=\sqrt{n\cdot D(\theta_n,\theta^*)}\,. \qquad (9)$$

Для примера на рис. 1-2 приведены графики статистики (9), полученные для оценок меры сдвига в зависимости от объёма выборки (n) в случае порождающих распределений (A) и (B). По оси абсцисс используется логарифмический масштаб.

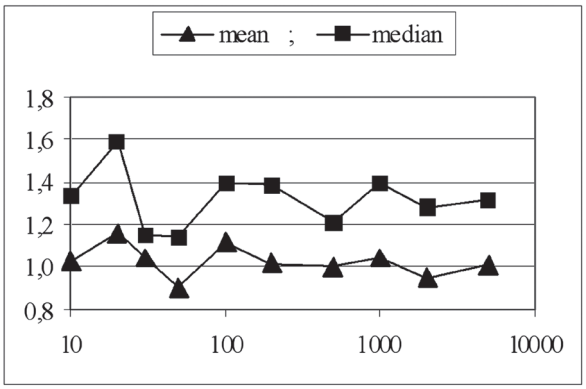

Рис. 1. Зависимость статистики (9) от размера выборки при порождающем распределении (A)

Итоговые значения этой статистики для параметров сдвига и разброса, полученные путём усреднения 100 выборок объёмом n каждая, приведены в табл.1-2 соответственно. Для сравнения в табл.1 представлены асимптотические оценки этой статистики при $n=\infty$, полученные теоретически.

Анализ полученных результатов показывает, что робастные оценки параметров сдвига и разброса (6)-(7), как и следует из теории, уступают по эффективности классическим оценкам (1) в случае, когда выборка соответствует “чистым” нормальному или логистическому законам (A) и (C). В то же время, в случае наличия выбросов, т.е. при законах (B) и (D) робастные оценки оказываются существенно эффективнее. Кроме того, робастные оценки могут успешно применяться и для тех законов, для которых моменты распределения не определены, таких, как закон Коши.

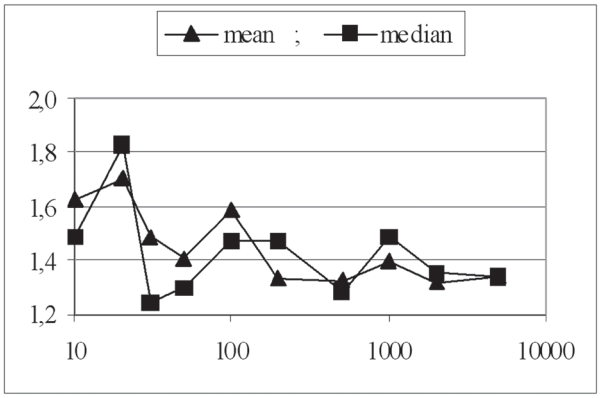

Рис. 2. Зависимость статистики (9) от размера выборки при порождающем распределении (B)

Таблица 1

Значения статистики (9) для оценок параметра сдвига

Порождающий закон распределения	Значения статистики (9) при $n = 20$		Значения статистики (9) при $n = 200$	
	mean	median	mean	median
A	1,159	1,593	1,015	1,383
B	1,699	1,824	1,330	1,475
C	1,541	1,757	1,667	1,994
D	1,805	1,859	2,376	2,105
E	297,528	1,468	415,320	1,577

Порождающий закон распределения	Значения статистики (9) при $n = 5000$		Асимптотические значения статистики (9)	
	mean	median	mean	median
A	1,005	1,316	1,000	1,253
B	1,338	1,339	1,342	1,343
C	1,770	2,051	1,814	2,000
D	2,490	2,218	2,433	2,143
E	1073,801	1,611	$+\infty$	1,571

Таблица 2

Значения статистики (9) для оценок параметра разброса

Порождающий закон распределения	Значения статистики (9) при $n = 20$		Значения статистики (9) при $n = 5000$	
	stdev	MAD	stdev	MAD
A	0,731	1,219	0,656	1,055
B	1,503	1,289	1,656	1,181
C	1,818	2,112	1,722	1,929
D	3,180	2,244	3,828	2,156
E	1323,9	2,734	75154,8	2,227

Выводы

Проведён анализ устойчивости числовых характеристик статистической выборки по отношению к выбросам. Выявлено, что среднее арифметическое значение и среднеквадратичное отклонение не являются устойчивыми. Предложено использовать устойчивые (робастные) характеристики сдвига и разброса – медиану выборки и медиану абсолютных отклонений (MAD). Проведено моделирование методом Монте-Карло, которое подтвердило целесообразность предложенного подхода.

Остаётся лишь удивляться, что такой полезный и широко применяемый для количественного анализа данных программный продукт как MS Excel, не содержит встроенной функции (6), позволившей бы получать робастные оценки параметра масштаба выборки. Более того, даже в таких пакетах для профессионального статистического анализа данных как Statistica и StatGraf хоть и реализованы методы робастного оценивания, но при стандартизации данных используются классические оценки (1), не являющиеся робастными.

Список литературы: 1. *Орлов, А. И.* Прикладная статистика [Текст] / А.И. Орлов. — М.: Экзамен, 2006. — 672 с. 2 Estimation of scale [Электронный ресурс]. — Режим доступа: http://en.wikipedia.org/wiki/Robust_statistics#Estimation_of_scale — 02.09.2010 г. — Загл. с экрана. 3. *Кокс, Д.* Теоретическая статистика [Текст] / Д. Кокс, Д. Хинкли. — М., Мир, 1978. — 640 с. 4. *Гайдышев, И.* Анализ и обработка данных [Текст]: специальный справочник / И. Гайдышев. — СПб: Питер, 2001. — 784 с.

Поступила в редколлегию 15.06.2010.

УДК 519.234.7

Робасні методи оцінювання числових характеристик вибірки / В.Л. Шергін, Т.О. Мірошниченко // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 3 (74). – С. 90–93.

Аналізується стійкість чисельних характеристик статистичної вибірки щодо викидів. Запропоновано використовувати робасні характеристики зсуву та розкиду вибірки — медіану вибірки та медіану абсолютних відхилень (MAD). Наведені результати чисельного моделювання, які підтверджують доцільність запропонованого підходу.

Іл. 2. Бібліогр.: 4 найм.

UDK 519.234.7

Robust methods for sampling characteristics estimation // V.L. Shergin, T.A. Miroshnichenko // Bionics of Intelligence: Sci. Mag. – 2010. – № 3 (74). – P. 90–93.

Stability of numerical characteristics of the sampling concerning to the outliers is analyzed. Robust characteristics such as median and median absolute deviation (MAD) are proposed to use instead mean and standard deviation as the characteristics of the shift and scale. The presented results of the numerical simulation confirms usability and appropriateness of the median and MAD.

Fig. 2. Ref.: 4 items.