## *11. Интервальная оценка моды в условиях асимметричного распределения данных*[1]

Достаточно часто на практике при анализе экономических показателей исследователи сталкиваются с сильной асимметрией. Именно это явление и было обнаружено в ходе проведённых нами исследований. Статистическая обработка таких данных оказывается затруднённой, так как ни средняя величина, ни дисперсия в этих данных уже не имеют особой ценности при анализе. В таком случае чаще всего прибегают к предварительному преобразованию исходных данных, например, к логарифмированию либо к возведению в степень. Все эти преобразования нужны для того, чтобы свести распределение к симметричному виду, а ещё лучше – наиболее близкого к нормальному. Такой подход с одной стороны удобен, так как позволяет после преобразования использовать все стандартные статистические характеристики, но с другой стороны – связан с трудностями подбора наилучшего механизма преобразования данных. В итоге поиск нужного преобразования оказывается связан скорее с попыткой подогнать те данные, которые имеются в распоряжении, под желаемые результаты исследователя, нежели к реальному исследованию.

Помимо этого, в тех случаях, если в данных встречаются не положительные значения, некоторые операции по преобразованию осуществить уже просто не получается. Например, вычислить логарифм нуля не представляется возможным (так как $\log(0) = -\infty$), а значит, и использовать описанный подход при работе с данными, содержащими нули и отрицательные числа, нельзя.

В этом случае вместо того, чтобы пытаться преобразовать данные, имеет смысл обратиться к другим статистическим характеристикам, лучше характеризующим имеющееся распределение. Одной из таких характеристик является мода.

Сложность расчёта моды по данным заключается в том, что чаще всего исследователь сталкивается с величинами, которые не повторяются либо повторяются редко. Например, затраты на лечение могут составлять и тысячу рублей, и 20 тысяч рублей. Вероятность того, что несколько человек понесут затраты одной и той же величины достаточно низка. Более того, совпадение затрат у нескольких человек ещё не говорит о том, что это и есть мода распределения. Для более точной оценки моды нужно имеющиеся данные представить в виде гистограммы и найти тот интервал, в который попало наибольшее число значений. То есть мода по сути своей в реальности скорее должна быть интервальной оценкой, нежели точечной. Сложность на данный момент заключается в том, что при выборе интервалов разной ширины могут получаться разные моды. Поэтому, несмотря на саму природу моды, для более точного анализа нужно всё-таки иметь в распоряжении точечную оценку. Процесс расчёта моды усложняется в связи с тем, что добавление или исключение наблюдений может приводить к изменению моды. То есть, по сути, мы имеем дело со случайной величиной, оценить которую оказывается достаточно сложно.

В случае если распределение случайной величины имеет незначительную асимметрию, выполняется равенство [75]:

$$\bar{y} - \hat{y} = 3 \cdot (\bar{y} - \tilde{y}), \quad (1)$$

где $\bar{y}$ - средняя арифметическая по ряду данных, $\hat{y}$ - мода, а $\tilde{y}$ - медиана.

В этом случае для оценки моды можно использовать равенство, выводящееся из (1):

$$\hat{y} = \bar{y} - 3 \cdot (\bar{y} - \tilde{y}). \quad (2)$$

Однако данная оценка оказывается неточной в случаях с большой асимметрией в данных и нужно обращаться к другим методам.

---

[1] Подготовлен совместно с И.С.Светуньковым

На данный момент существует несколько методов точечной оценки моды. Например, в [65] и [90] предлагается подход, состоящий из нескольких шагов:

1. Предварительное преобразование данных с помощью трансформации Бокса-Кокса таким образом, чтобы получить распределение величины, близкое к нормальному.
2. Расчёт средней величины и стандартного отклонения по полученным данным (либо медианы и стандартизированного медианного расстояния).
3. Преобразование полученных величин в исходные данные.

Такой подход позволяет получить оценку, близкую к моде в случае с унимодальным распределением. Сложность в данном случае вызывает преобразование Бокса-Кокса, которое подразумевает подбор параметра, дающего распределение величин, близкое к нормальному на первом шаге. Делается это путём максимизации функции правдоподобия.

Альтернативным данному является непараметрический подход с расчётом эмпирической функции плотности по ряду данных и выбору такого значения $y$, которое бы давало максимальное значение полученной функции плотности [134]. При использовании этого подхода сложности возникают в выборе параметров (например, ширины интервала для ядерной функции), но в целом этот подход позволяет получить оценку моды, близкую к реальной без предположений относительно закона распределения случайной величины.

Более простые методы оценки моды были предложены в статьях [171] и [75]. Общая идея этих методов заключается в том, что исследователь выбирает модальный интервал ($y_j$, $y_{j+k}$), в который попало наибольшее число наблюдений, после чего предполагается, что мода находится в середине интервала и рассчитывается соответствующая точка:

$$\hat{y} = \frac{y_j + y_{j+k}}{2}, \quad (3)$$

которая принимается точечной оценкой моды. Минусом этого подхода является отсутствие общего принципа определения ширины интервала: выбрав интервалы разной ширины, исследователь будет получать разные оценки моды.

Для того чтобы построить доверительный интервал для моды мы предлагаем обратиться к непараметрическому методу, идея которого заключается в расчёте моды по нескольким выборкам из имеющихся в распоряжении данных. Весь метод можно свести к следующим шагам:

1. Исследователь определяет число значений мод $M$, которое ему нужно для оценки дисперсии моды.
2. Из имеющегося ряда данных создаётся случайная выборка, состоящая из 75% наблюдений.

Число наблюдений может быть и другим. Заметим только, что оно не должно быть слишком маленьким (даже 50% может быть мало в некоторых случаях) или слишком большим (при 90% у моды может отсутствовать вариация).

3. По полученной выборке рассчитывается точечная оценка моды с помощью одного из методов, упомянутых выше.
4. Если число полученных значений мод меньше $M$, то повторяются шаги 2 – 4.
5. Получив $M$ значений мод, исследователь может по ним рассчитать интересующие его статистические характеристики.
    a. Например, если распределение мод оказалось симметричным, то в качестве итоговой точечной оценки моды $\hat{y}$ можно использовать среднюю арифметическую, а в качестве меры колеблемости – дисперсию по полученному ряду мод.
    b. Если же распределение оказалось не симметричным (что чаще встречается на практике), то для точечной оценки моды $\hat{y}$ можно использовать моду полученного распределения, а для оценки меры колеблемости имеет смысл обратиться к какому-нибудь методу

построения несимметричных доверительных интервалов.

Такой подход позволяет минимизировать субъективность при расчёте интервала для моды. В своих расчётах при построении доверительного интервала, мы использовали полупараметрический метод, который заключается в следующем.

Первым шагом мы оцениваем остатки относительно заданной нами оценки (в нашем случае - это мода по сгенерированному ряду) $\hat{y}_i$: $e_i = y_i - \hat{y}_i$ – и разделяем их на две группы:

1. Положительные остатки, $e_i > 0$,
2. Отрицательные остатки, $e_i < 0$.

Получить остатки, равные нулю, практически невозможно.

Следующим шагом нужно оценить количество положительных $n_r$ (то есть находящихся справа от оценки $\hat{y}_i$) и отрицательных $n_l$ (то есть находящихся слева от оценки $\hat{y}_i$) остатков, после чего вычисляем общее число наблюдений:

$$n_r + n_l = n \quad (4)$$

Затем нужно оценить число степеней свободы для каждой из этих групп. Учитывая то, что в общем случае в построенной модели $k$ коэффициентов, а наблюдений в каждой группе может быть не одинаковое количество, мы не можем рассчитать степени свободы по стандартной формуле:

$$df = n - k \quad (5)$$

Отнимать от $n_r$ и $n_l$ одинаковое число коэффициентов $k$ так же будет неправильно, в связи с тем, что сумма степеней свободы по разным остаткам не будет равна (5):

$$(n_r - k) + (n_l - k) = n - 2k$$

Значит, нам нужно оценить индивидуальный вклад каждой из групп в общее число степеней свободы. Для решения этой задачи могут быть предложены разные методы. Мы предлагаем число «коэффициентов» разделить между «левой» и «правой» частями остатков пропорционально числу наблюдений в этих частях:

$$k_l = \frac{n_l}{n} k, \quad k_r = \frac{n_r}{n} k \quad (6)$$

Тогда число степеней свободы для каждой части будет рассчитываться по формулам:

$$df_l = n_l - k_l = n_l - \frac{n_l}{n} k = n_l \left(1 - \frac{k}{n}\right) = n_l \frac{n-k}{n} = \frac{n_l}{n} df, \quad (7)$$

$$df_r = \frac{n_r}{n} df \quad (8)$$

Сумма степеней свободы (7) и (8), очевидно будет равна (5).

Следующим шагом требуется оценить дисперсии случайных величин относительно $\hat{y}_i$ для каждой из выделенных групп, которые рассчитывается по следующим формулам после упорядочивания ряда ошибок по возрастанию:

$$\begin{cases} \sigma^2_{e,l} = \dfrac{1}{df_l} \sum\limits_{i=1}^{n_l} e_i^2 \\ \sigma^2_{e,r} = \dfrac{1}{df_r} \sum\limits_{i=n_l+1}^{n} e_i^2 \end{cases} \quad (9)$$

Рассчитав отдельно СКО для положительных и для отрицательных остатков (как квадратный корень из (9)), мы получим меру колеблемости с разных сторон от нашего расчётного значения. Эти величины можно использовать для построения несимметричного доверительного интервала для моды по формуле:

$$\hat{y} - \pi_l \sqrt{\sigma_{e,l}^2} < \mu < \hat{y} + \pi_r \sqrt{\sigma_{e,r}^2}, \quad (10)$$

где $\mu$ – мода в генеральной совокупности, $\pi_l$ – статистика для отрицательных остатков, $\pi_r$ – статистика для положительных остатков.

В качестве статистики для расчёта интервалов мы для простоты будем использовать *t*-статистику. Данное решение может быть интерпретировано как то, что мы вводим предположение относительно независимого распределения случайных величин слева и справа от нашей точечной оценки – мы предполагаем, что распределение близко к нормальному, но с разными параметрами, что как раз и приводит к асимметрии. Это достаточно грубое предположение, и в некоторых случаях имеет смысл обратиться к другим статистикам (например, к статистике, выводимой из неравенства Чебышева). Однако для целей нашего исследования можно воспользоваться и этим допущением.

По данным о затратах на лечение больных ИБС до проведения операции на сердце, используя метод, описанный выше, мы рассчитали точечные оценки моды для двух групп больных: для тех, кто проходит реабилитацию и для тех, кто не проходит реабилитацию. При этом распределения, полученные в результате итеративной процедуры для получения интервальных оценок, оказались несимметричными, поэтому для оценки доверительных интервалов для моды мы воспользовались методом построения несимметричных интервалов, описанным выше.

В результате расчётов были получены следующие точечные и интервальные оценки мод по исходным данным (таблица 27).

Таблица 27.

**Точечные и интервальные оценки мод показателей**

| | I группа, без реабилитации | | | II группа, с реабилитацией | | |
|---|---|---|---|---|---|---|
| | Левая граница | Точечная оценка | Правая граница | Левая граница | Точечная оценка | Правая граница |
| Стоимость амбулаторных лечебно-диагностических мероприятий, руб./год | 17544,84 | 23075,93 | 35782,59 | 9917,47 | 22552,12 | 30155,42 |
| Стоимость вызовов бригад неотложной помощи, руб./год | 260,98 | 826,57 | 1144,26 | 282,10 | 397,81 | 648,25 |
| Стоимость стационарного лечения, руб./год | 11701,45 | 13803,42 | 25997,37 | 5998,62 | 8690,57 | 20191,82 |
| Стоимость лекарственной терапии в год, руб./год | 1491,04 | 2909,82 | 3590,09 | 1237,98 | 1637,77 | 2247,15 |
| Суммарные прямые затраты на лечение, руб./год | 37724,74 | 49111,77 | 80266,90 | 23938,42 | 30220,69 | 65966,37 |
| Выплаты пособий по временной нетрудоспособности (оплата больничных листов), руб./год | 3561,08 | 6744,08 | 10619,89 | 4432,21 | 5360,68 | 9651,68 |
| Денежные выплаты по стойкой нетрудоспособности (пособия по | Моду оценить невозможно. | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| инвалидности), руб./год | | | | | | |
| Суммарные непрямые (косвенные) экономические потери, руб./год | 19930,42 | 29105,01 | 48050,55 | 8075,78 | 11339,53 | 19814,82 |
| Суммарные экономические потери в связи с заболеванием, руб./год | 53807,31 | 67422,34 | 118994,20 | 42635,86 | 66093,43 | 83211,90 |

По исходным данным оценить моды денежных выплат по стойкой нетрудоспособности не представляется возможным, потому что данные в этой переменной по обеим категориям имеют мультимодальный вид: много наблюдений приходится на «0», много – на «17778,36» и примерно столько же – на «35556,84». Никаких промежуточных значений эта переменная не принимает, так как пособия по инвалидности имеют фиксированную величину.

По всем остальным данным явно видно, что затраты для второй группы (проходящей реабилитацию) оказываются ниже затрат первой группы. По некоторым показателям интервалы мод даже не пересекаются, что указывает на статистически значимое различие затрат для этих двух категорий больных. В большинстве случаев точечная оценка моды по затратам ниже для группы больных, проходивших реабилитацию. Только по выплате пособий по временной нетрудоспособности обе группы могут быть более-менее сопоставимыми. Практический вывод, следующий из этого анализа, заключается в том, что при прохождении реабилитации большая часть больных несёт значительно меньшие затраты, чем в случае с игнорированием реабилитации.

Как видим, предложенный метод позволяет проводить анализ статистических данных в условии асимметрии и получать более полную информацию о наиболее частом распределении затрат респондентов по различным показателям.