УДК 330.4

Пантелеева О.Б.,

к.э.н., доцент

кафедры бухгалтерского учета и анализа

Краснодарского филиала «РЭУ им. Г.В.Плеханова», РФ

НЕКОТОРЫЕ ПРИНЦИПЫ ОБРАБОТКИ ОДНОРОДНЫХ СОВОКУПНОСТЕЙ

SOME PRINCIPLES OF HOMOGENEOUS AGGREGATUALITY PROCESSING

Аннотация: *статья посвящена методам максимального правдоподобия при обработке однородных совокупностей. Методика и техника определения характеристик совокупностей рассмотрена на примере дискретных случайных величин. Предложенный прием может быть применим для оценки нескольких параметров распределения случайных величин экономических исследований.*

Abstract: *The article is devoted to the methods of maximum likelihood in the processing of homogeneous sets. The methods and technique for determining the aggregate characteristics is considered using the example of discrete random variables. The proposed method can be used to estimate several parameters of the distribution of random variables of economic research.*

Ключевые слова: *распределение случайных величин, однородные совокупности, метод максимального правдоподобия, дискретные случайные величины, состоятельная оценка.*

Key words: *distribution of random variables, homogeneous sets, maximum likelihood method, discrete random variables, consistent estimation.*

Приступая к статистической обработке некоторой совокупности, надо прежде всего представить ее как совокупность нескольких реализаций некоторой случайной величины, то есть как выборку из генеральной совокупности. Это означает ориентацию на приближенное, а не точное измерение характеристик данной совокупности. Такой подход не соответствует подходу, при котором некоторые конечные совокупности с самого начала рассматриваются как генеральные, а не выборочные. Например, изучая темпы роста объемов работ в какой-либо организации и рассматривая данные о величине этих объемов за весь период существования организации, экономист будет считать совокупность этих данных генеральной совокупностью (так как никаких других данных об организации в природе нет). При нашем подходе считается, что в анализируемом периоде объемы работ изменились под влиянием двух групп факторов - одни из них обусловили некоторый закономерный рост объемов, вторые - случайные отклонения от такой закономерности. Влияние факторов второй группы отражается в реализации нескольких конкретных случайных отклонений, являющихся выборкой из бесконечного множества возможных реализаций соответствующей случайной величины. В зависимости от этих случайных отклонений должен выбираться тот или иной метод экономического исследования.

Методологию и технику определения характеристик совокупности рассмотрим на примере, в котором случайные колебания ее элементов дискретны. Этому условию удовлетворяют совокупности элементов, принимающих целочисленные значения. Характерным примером таких совокупностей являются данные об интервалах поставки товарных запасов. Изучение и обработка этих данных необходимы для нормирования оборотных средств организаций.

Перенумеруем подряд, начиная с 1, календарные дни в отчетном периоде и обозначим через t_n номер дня, в котором имела место n-я по счету поставка товаров (n=1, ..., N+1), а через d_n - интервал между n-й и (n+1)-й поставками:

$$d_n = t_{n+1} - t_n \ (n=1, ..., N), \qquad (1)$$

Нас интересует средняя длительность интервала поставок d. Методы расчета этой величины будут различны в зависимости от характера случайных колебаний интервалов между поставками. При этом мы будем считать, что минимальный интервал между поставками равен одному дню, так что две поставки в один и тот же день будут считаться одной поставкой.

1. Наиболее простым предположением о законе распределения d_n является следующее: в каждый день рассматриваемого периода поставка производится с некоторой вероятностью p и не производится с вероятностью $(1- p)$ независимо от того, сколько поставок было произведено ранее и в какие дни. Поэтому вероятность того, что от одной поставки до другой прошло k дней, равна произведению вероятностей $(k-1)$-й "непоставки" на вероятность одной поставки одной поставки в последний, k-й день:

$$P(k)= p(1- p)^{k-1} \ (k =1,2,...), \qquad (2)$$

Средняя продолжительность интервала между поставками (т.е. математическое ожидание d_n) равна

$$d=\sum_{k=1}^{\infty} kP(k)=\sum_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p}, \qquad (3)$$

то есть, является обратной к вероятности поставки. Несмотря на целочисленность всех интервалов d_n, их среднее значение может быть и не целым, то есть в данном случае среднее не отражает это характерное общее свойство всех элементов совокупности.

Одним из наиболее общих и широко применяемых приемов оценки параметров случайных величин является метод максимального правдоподобия. В данном случае для его применения необходимо определить показатель L правдоподобия наблюдаемых фактических данных, исчисляемый как вероятность того, что случайно реализуются именно такие интервалы поставок $d_1,..., d_N$, которые наблюдались фактически. Учитывая,

что интервалы между различными поставками независимы, с помощью (2) находим показатель правдоподобия:

$$L = p(1-p)^{d-1}\ p(1-p)^{d-1} \dots p(1-p)^{d-1} = p^N (1-p)^{d+\dots+d-N}, \qquad (4)$$

"Наиболее правдоподобным" значением p считается такое, которому соответствует наибольшее возможное значение L. Другими словами, при определении p принимается, что при "истинном" p наблюдаемые значения интервалов поставок наиболее вероятны, поэтому-то и реализовались именно они, а не другие значения интервалов. В данном случае максимальное значение L будет иметь место при

$$p = \hat{p} = \frac{N}{d_1 + \dots + d_N} = \frac{N}{t_{N+1} - t_1},$$

и, следовательно, при "наиболее правдоподобном" значении

$$d = \hat{d} = \frac{d_1 + \dots + d_N}{N} = \frac{t_{N+1} - t_1}{N}, \qquad (5)$$

Полученной формуле можно дать следующее объяснение:

а) средний интервал поставок определяется как среднее арифметическое из отдельных интервалов;

б) средний интервал поставок определяется как интервал между первой и последней поставками, отнесенный к числу интервалов или к числу поставок без одной.

Оценки, получаемые по методу максимального правдоподобия, обладают многими хорошими свойствами. В частности, дисперсия такой оценки во многих случаях не может быть существенно улучшена. В данном случае эта дисперсия составляет:

$$S(\hat{d}) = \frac{1-p}{p^2 N} = \frac{d(d-1)}{N} \qquad (6)$$

2. Закон распределения интервалов поставок может быть и иным. Пусть, например, длительность каждого интервала независимо от других принимает с равной вероятностью любые значения от 1 до D:

$$P(k)=\begin{cases}\frac{1}{D} \text{ при } k = 1,\dots,D;\\ 0 \text{ при } k > D.\end{cases} \quad (7)$$

Математическое ожидание интервала поставок при этом будет:

$$d=\sum_{k=1}^{\infty} kP(k)=\sum_{k=1}^{D} k\frac{1}{D}=\frac{D+1}{2}\ , \quad (8)$$

Правдоподобие фактически наблюдаемых интервалов поставок в этом случае в силу (7) равно:

$$L=P(d_1)\dots P(d_N)=\begin{cases}\frac{1}{D^N}, \text{ если } d_1 \le D, d_2 \le D, \dots, d_N \le D,;\\ 0 \text{ в противном случае.}\end{cases} \quad (9)$$

Наиболее правдоподобным будет наименьшее возможное D, при котором еще будет $d_1 \le D, \dots, d_N \le D$, т.е.

$$\widetilde{D}=\max(d_1, \dots, d_N).$$

В этом случае среднее значение интервала поставок:

$$\widetilde{d}=\frac{1}{2}+\frac{1}{2}\max(d_1, \dots, d_N) \quad (10)$$

будет на 1/2 дня превышать половину максимального из наблюдавшихся интервалов между поставками. Для определения дисперсии этой оценки определим сначала математическое ожидание и дисперсию $\widetilde{D}$. Имеем:

$$M(\widetilde{D})=\sum_{1\le d_1,\dots,d_N\le D}\frac{\widetilde{D}}{D^N}=\sum_{k=1}^{D}\frac{k}{D^N}\sum_{\max(d_1,\dots,d_N)=k}1;$$

$$M(\widetilde{D^2})=\sum_{k=1}^{D}\frac{k^2}{D^N}\sum_{\max(d_1,\dots,d_N)=k}1.$$

Но

$$\sum_{\max(d_1,\dots,d_N)=k} 1 = \sum_{\max(d_1,\dots,d_N)\le k} 1 - \sum_{\max(d_1,\dots,d_N)\le k-1} 1 = k^N - (k-1)^N.$$

Отсюда

$$M(\tilde{D}) = \sum_{k=1}^{D} \frac{k}{D^N}\left(k^N - (k^N-1)^N\right) = D + \sum_{k=1}^{D-1} k^N \times \left(\frac{k}{D^N} - \frac{k+1}{D^N}\right) =$$

$$= D - \sum_{k=1}^{D-1} \frac{k^N}{D^N} = D - A_N,$$

где для m=1, 2, ..., *N*, *N*+1; $A_m = \sum_{k=1}^{D-1} \left(\frac{k}{D}\right)^m$;

$$M(\tilde{D^2}) = \sum_{k=1}^{D} \frac{k^2}{D^N}\left[k^N - (k-1)^N\right] = D^2 + \sum_{k=1}^{D-1} k^N \left(\frac{k^2}{D^N} - \frac{(k+1)^2}{D^N}\right) =$$

$$= D^2 - \sum_{k=1}^{D-1} \frac{k^N(2k+1)}{D^N} = D^2 - 2DA_{N+1} - A_N.$$

Поэтому

$$S(\tilde{D}) = M\left(\tilde{D^2}\right) - \left[M(\tilde{D})\right]^2 = 2D(A_N - A_{N+1}) - A_N - A_N^2.$$

Так как $0 < A_N < \frac{D}{N}$ и $S(\tilde{D}) < \frac{2D^2}{N^2}$,

То $D - \frac{D}{N} < M(\tilde{D}) < D$; $S(\tilde{D}) < \frac{2D^2}{N^2}$.

Следовательно,

$$0 < M(\tilde{d}) < d - \frac{d}{N}; \quad S(\tilde{d}) < \frac{2d^2}{N^2}. \qquad (11)$$

Сравнивая (11) с (8), получаем, что $\tilde{d}$ является довольно точной оценкой истинного значения *d* - как средняя, так и среднеквадратическое отклонение $\tilde{d}$ от *d* имеют порядок $\frac{d}{N}$. Это намного точнее, чем в предыдущем примере, где в силу (6) среднеквадратическое отклонение имело порядок $\frac{d}{\sqrt{N}}$.

Если бы, не зная характера закона распределения, мы применили для оценки *d* формулу (5), мы намного ухудшили бы точность оценки.

Интересно провести ее обратное сравнение - в условиях первого примера применить для оценки d формулу (10). Для нахождения математического ожидания и дисперсии $\tilde{D}$ и $\tilde{d}$ в этом случае удобно применить следующий прием. Поскольку вероятность события $d_i \le k$ равна

$$P(d_i \le k) = \sum_{n=1}^{k} p(1-p)^{n-1} = 1-(1-p)^k,$$

то вероятность того, что все $d_i \le k$, т.е. того, что максимум из них не превосходит k, равна

$$P(\tilde{D} \le k) = (1-(1-p)^k)^N.$$

Поэтому

$$P(\tilde{D} = k) = P(\tilde{D} \le k) - P(\tilde{D} \le k-1) = (1-(1-p)^k)^N - (1-(1-p)^{k-1})^N$$

Отсюда

$$M(\tilde{D}) = \sum_{k=1}^{\infty} k[(1-(1-p)^k)^N - (1-(1-p)^{k-1})^N] =$$

$$\sum_{k=0}^{\infty}[1-(1-(1-p)^k)^N].$$

Точное вычисление полученной суммы затруднительно, однако для наших целей достаточно довольно грубой оценки. Поскольку функция $\varphi(k) = 1-(1-(1-p)^k)^N$ с ростом k убывает, то

$$M(\tilde{D}) = \sum_{k=0}^{\infty} \varphi(k) > \sum_{k=0}^{\infty} \int_k^{k+1} \varphi(x)dx = \int_0^{\infty}[1-(1-(1-p)^x)^N]\,dx.$$

Сделав замену переменных $x = \frac{\ln(1-u)}{\ln(1-p)}$, получим

$$M(\tilde{D}) > -\int_0^1 [1-u^N]\frac{du}{(1-u)\ln(1-p)} = -\frac{1}{\ln(1-p)}\int_0^1 (1+u+\cdots+u^{N-1})du =$$

$$= -\frac{1}{\ln(1-p)}\left(1+\frac{1}{2}+\cdots+\frac{1}{N}\right).$$

При больших N и малых p эта величина имеет порядок $\frac{1}{p}\ln N$, поэтому $M(\tilde{D})$ будет асимптотически больше, чем $D\ln N$:

$$M(\tilde{d}) > d\ln N. \qquad (12)$$

Таким образом, оценка $\tilde{d}$ и d является значительно смещенной, причем тем больше, чем больше N. Это значит, что ее применение в данном случае приведет к грубым ошибкам.

Наши рассуждения показывают, что нельзя для оценки характеристик распределения пользоваться теми или иными оценками, не используя информацию о виде закона распределения.

Приведенные выше примеры показывают, что оценивание одного и того же параметра распределения (например, математического ожидания) целесообразно производить различными способами в зависимости от характера самого распределения. Однако достаточно подробной информации о характере распределения элементов совокупности может и не быть. Как же поступать, если мы не знаем, является ли закон распределения равномерным (7) или показательным (2). В этой ситуации нужно использовать метод оценивания. Тогда из двух изложенных методов необходимо выбрать первый. Если распределение равномерное, этот метод дает оценку, дисперсия которой с ростом N стремится к нулю, в то время как при использовании второго метода (10) для показательного распределения оценка получается "плохой" - ее дисперсия с ростом N неограниченно возрастает.

Такой подход, несмотря на свою внешнюю привлекательность, не самый лучший. Для выбора подходящего метода оценивания целесообразно использовать метод максимального правдоподобия. Действительно, если мы из многих возможных значений параметра d выбираем в качестве "наилучшего", то есть наиболее соответствующего фактическим данным, то, которому отвечает наибольшее значение правдоподобия, то этот же прием может быть применен для выбора "наилучшего" закона распределения. Для этого предположим сначала, что распределение d_i описывается законом (2). Выбрав наиболее правдоподобное значение $d = \hat{d}$, можем определить численное значение функции правдоподобия по формуле (4). Это значение будет равно:

$$\hat{L}=\left(\frac{1}{\hat{d}}\right)^N\left(1-\frac{1}{\hat{d}}\right)^{N\tilde{d}-N}.$$

Предположим теперь, что распределение d_i описывается формулой (7). Значение функции правдоподобия при наиболее правдоподобном значении $d = \hat{d}$ определяется по формуле (9):

$$\hat{L}=(\frac{1}{\hat{D}})^N = (\frac{1}{2\hat{d}-1})^N.$$

Сравним теперь полученные значения $\hat{L}$ и $\tilde{L}$. Если $\hat{L} > \tilde{L}$, то более правдоподобен первый закон распределения и лучше пользоваться оценкой $\hat{d}$, в противном случае фактические данные лучше соответствуют второму закону распределения и предпочтительнее оценка $\tilde{d}$.

Такой подход становится очевидным, если записать имеющуюся информацию о законе распределения d_i в следующей форме:

$$P(d_i=k)=\begin{cases}\frac{1}{z}(1-\frac{1}{d})^{k-1} \text{ при } \theta = 1, k = 1,\ 2, \ldots; \\ \frac{1}{2d-1} \text{ при } \theta = 2, k = 1,\ 2, \ldots, 2d-1; \\ 0 \text{ при } \theta = 2, k > 2d-1\end{cases} \qquad (13)$$

и поставить задачу оценки двух параметров распределения - математического ожидания d и параметра θ. Изложенный выше способ и является, по существу, максимализацией правдоподобия по этим двум параметрам. Приведем числовой пример, иллюстрирующий этот способ.

Имеем следующие величины интервалов между поставками кирпича на строительную площадку (в днях):

7, 5, 9, 2, 4, 11, 3, 2, 6, 9.

Необходимо определить средний интервал поставок.

Если распределение интервалов описывается законом (2), то согласно (5):

$$\hat{d}=\frac{7+5+9+2+4+11+3+2+6+9}{10} = \frac{58}{10}=5{,}8 \text{ дня}.$$

При этом,

$$\hat{L} = (\frac{1}{5.8})^{10}(1-\frac{1}{5.8})^{58-10} \approx 2{,}6\times 10^{-12}.$$

При равномерном распределении интервалов поставок (7):

$$\tilde{D}=max_i(d_i)=11;$$

$$\tilde{d}=\frac{1+D}{2}=6{,}0 \text{ дня;}$$

$$\tilde{L}=\left(\frac{1}{11}\right)^{10}\approx 3{,}9\times 10^{-11}.$$

Поскольку $\tilde{L} > \hat{L}$, то в данном случае более правдоподобен второй закон распределения и средний интервал поставок целесообразно оценить в 6 дней.

Теоретически возможен случай, когда метод максимального правдоподобия не позволяет определить параметры распределения однозначно или, более того, дает бесконечном множество оценок, не являющихся даже состоятельными. Например, пусть случайная величина равномерно распределена на отрезке единичной длины и ищется середина этого интервала θ. Математически это сводится к задаче определения параметра θ - величины, равномерно распределенной на отрезке $\left[\theta-\frac{1}{2},\theta+\frac{1}{2}\right]$, по ее реализациям x_1, ..., x_N. Наиболее правдоподобными будут любые значения $\hat{\theta}$ в интервале от $max_i x_i - \frac{1}{2}$ до $min_i x_i + \frac{1}{2}$. Состоятельными оценками этого параметра будут лишь некоторые из этих значений, например, середина указанного интервала или средняя арифметическая x_i.

В дальнейшем мы постараемся не рассматривать такие "патологические" случаи, редко встречающиеся в практике экономико-статистических расчетов. Как правило, мы будем иметь дело с такими случайными величинами, для которых при всех значениях реализаций функция правдоподобия достигает максимума в единственной точке, являющейся состоятельной оценкой неизвестного параметра распределения. Если же закон распределения не таков, будем предполагать, что имеется возможность выбора из нескольких "оптимальных" точек такой "наиболее оптимальной", которая дает состоятельную оценку этого параметра. Разумеется, тот же прием может быть применен и для оценки нескольких параметров распределения, и для выбора одного из нескольких возможных законов распределения.

Литература

1. Пантелеева М.А., Пантелеева О.Б. Применение информационных технологий для развития бизнес-коммерции//Актуальные проблемы экономической теории и практики. Кубанский государственный университет; Под ред. В.А. Сидорова. Краснодар, 2015. С.67-75.

2. Пантелеева М.А., Пантелеева О.Б. Анализ информационно-коммуникационных технологий в бизнес-коммерции//Сфера услуг: инновации и качество. 2016, № 21, с. 12

3. Пантелеева О.Б. Экономико-математический анализ альтернативных инвестиционных проектов // Сфера услуг: инновации и качество. 2012. №8. С. 23.

4. Пидяшова О.П., Кравченко Т.Е., Трещенко Т.А. Оценка инвестиционной активности организаций в современных условиях(региональный аспект) // Экономика и предпринимательство, 2016, №4-1 (69-1).с.433-439

5. Салий В.В., Кузьмина Э.В. Применение формализованных методов аналитико-синтетической переработки информации в библиотечно-библиографической деятельности // Культурная жизнь юга России. Приложение. 2015. № 1 (1). С. 102-105.

6. Терещенко Т.А., Гусакова М.А. Оценка налоговых доходов бюджета, формируемых субъектами малого бизнеса //Сфера услуг: инновации и качество, 2016, № 25, с.3

7. Углова И.А. Анализ эффективности использования основных средств предприятия на основе данных финансовой отчетности / И.А. Углова И.А., Е.А.Оксанич // В сборнике: Современная экономика: проблемы, перспективы, информационное обеспечение материалы VI международной научной конференции, посвященной 95-летию Кубанского ГАУ и 15-летию кафедры теории бухгалтерского учета. 2017. С. 489-494.