ОЦЕНКИ ПОКАЗАТЕЛЕЙ КАЧЕСТВА ПО МАЛЫМ ВЫБОРКАМ

Ломакин М.И.

Аннотация: предложен алгоритм нахождения оценок показателей качества, представленных вероятностными функционалами для случая малых выборок, основанный на результатах теории моментов; доказано, что для получения наиболее точных оценок следует использовать количество моментов равное объему выборки.

Ключевые слова: оценка, показатель, качество, вероятность, моменты распределения.

EVALUATION OF INDICATORS OF QUALITY FOR A SMALL SAMPLE

Lomakin M.I.

Abstract: We propose an algorithm for evaluation of the quality indicators presented probabilistic functional for the case of small samples, based on the results of the theory of moments, it is proved that for the most accurate estimates of the number of points you should use an equal volume of sample.

Keywords: assessment, rate, quality, chances are you moments of the distribution.

В ряде случаев для оценки качества тех или иных социально-экономических процессов используют показатели вида: $P(\xi \geq (<) \varepsilon)$; здесь $P(\ldots)$ вероятность того, что случайная величина ξ не меньше (меньше) заданной величины ε. Такой показатель используется для оценки вероятности того, что прибыль предприятия, доходность портфеля проектов будет не ниже заданной, величина риска не превысит заданный уровень и др.

В теории вероятностей [1] известна следующая оценка для произвольного распределения $F(t)$ с известным математическим ожиданием m и дисперсией σ^2

$$P(|\xi - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \qquad (1)$$

здесь ξ – случайная величина, распределенная в соответствии с неизвестным законом распределения $F(t)$;

ε – произвольная положительная величина.

Оценка (1) хотя и является универсальной, но весьма грубой.

Улучшениям и уточнениям оценки (1) посвящены были многие работы. Первой работой, где решена задача нахождения экстремальных или гарантированных (нижних и верхних) оценок вероятности $P(\xi \geq \varepsilon)$ при известном математическом ожидании и дисперсии была работа Гермейера Ю.Б., Иргер Д.С., Калабуховой Е.П. [2].

Нижняя оценка вероятности при этом равна

$$P_{н}(\xi \geq \varepsilon) = \frac{(m-\varepsilon)^2}{(m-\varepsilon)^2 + \sigma^2}; \quad \varepsilon \leq m. \qquad (2)$$

Верхняя оценка вероятности $P(\xi \geq \varepsilon)$ равна

$$P_{в}(\xi \geq \varepsilon) = \min\left\{1, \frac{m}{\varepsilon}, \frac{\sigma^2}{(m-\varepsilon)^2 + \sigma^2}\right\}. \qquad (3)$$

Однако, полученные в последней работе результаты не могут быть обобщены на произвольное число моментов распределения. Обобщение полученных результатов, на произвольное число моментов, получено автором [3].

Определим множество функций распределения с заданными моментами $m = (m_1, m_2,\ldots,m_k)$ в виде:

$$F_0 = \{F(t) : \int_{-\infty}^{\infty} t^i dF(t) = m_i; i = 1,2\ldots,k\}. \qquad (4)$$

Имеет место следующее утверждение [3].

Утверждение 1. Наибольшее (наименьшее) значение интеграла

$$J(F) = \int_{0}^{\tau+0} c(t)dF(t) \qquad (5)$$

при непрерывной подинтегральной функции $c(t)$, имеющей $k+1$ неотрицательную производную и $F(t) \in F_0$ достигается на единственном ступенчатом распределении $F(t)$, у которого среди точек роста $t_1, t_2,\ldots,t_v$ имеет точка τ; при нечетном k число точек роста v функции распределения F(t) определяется соотношением v = (k + 3)/2, причем $t_0 = 0 < t_1 < t_2 < \ldots < t_v < \infty$; при четном k число точек роста v функции распределения F(t) определяется соотношением v = k/2 + 1, причем $0 < t_1 < t_2 < \ldots < t_v < \infty$; числа $p_j > 0$, t_j, j = 1,2,...,v удовлетворяют системе уравнений:

$$m_i = \sum_{j=1}^{v} t_j^i p_j; \qquad i = 0,1,\ldots,k; \qquad p_0 = 1. \tag{6}$$

Последний результат позволяет непосредственно находить верхние (нижние) (или гарантированные) оценки вероятности того, что случайная величина (прибыль, доходность и др.) превысит некоторый заданный уровень ε, т.е. величину вероятности P(r ≥ ε). Рассмотрим случай, когда известны два первые момента распределения случайной величины m_1, m_2 (и конечно m_0 = 1). Запишем уравнения для моментов при условии, что гарантированная оценка вероятности P(r ≥ ε) или F(ε) =1 - P(r ≥ ε) достигается на ступенчатом распределении F(t), имеющем v = 2 точки роста. Имеем:

$$\begin{cases} p_1 + p_2 = 1, \\ p_1 t_1 + p_2 t_2 = m_1, \\ p_1 t_1^2 + p_2 t_2^2 = m_2 \end{cases} \tag{7}$$

систему из трех уравнений с четырьмя неизвестными. Очевидно, что

$$F^x(\varepsilon) = \begin{cases} 0 \text{ при } \quad 0 \le \varepsilon \le t_1, \\ p_1 \text{ при } \quad t_1 \le \varepsilon \le t_2, \\ 1 \text{ при } \quad \varepsilon \ge t_2. \end{cases} \tag{8}$$

Последнее соотношение позволяет освободиться от одного неизвестного в системе уравнений (9) и тем самым найти гарантированную оценку $F^x(\varepsilon)$.

Так как F(ε) = F(ε - 0), т.е. функция распределения является непрерывной слева функцией, то наибольшее значение интеграла

$$J(F) = F(\tau) = \int_0^{\tau+0} dF(t) \tag{9}$$

при $F(t) \in F_0$ достигается на ступенчатом распределении $F(t)$, для которого среди точек роста t_1, t_2 имеется точка ε. Из соотношения (10) следует, что наибольший интерес представляет случай $F^x(\varepsilon) = p_1$. Тогда полагаем $t_1 = \varepsilon$ и решаем систему из трех уравнений с тремя неизвестными

$$\begin{cases} p_1 + p_2 = 1, \\ p_1\varepsilon + p_2 t_2 = m_1, \\ p_1\varepsilon^2 + p_2 t_2^2 = m_2 \end{cases} \qquad (10)$$

Откуда после несложных преобразований находим

$$F^x(\varepsilon) = \begin{cases} 0 & \text{при} \quad \varepsilon = 0, \\ \dfrac{m_2 - m_1^2}{m_2 - 2m_1\varepsilon + \varepsilon^2} & \text{при} \quad 0 \le \varepsilon \le m_1, \\ 1 & \text{при} \quad \varepsilon \ge m_1. \end{cases} \qquad (11)$$

$$P(\varepsilon) = \begin{cases} 1 & \text{при} \quad \varepsilon = 0, \\ \dfrac{(m_1 - \varepsilon)^2}{m_2 - 2m_1\varepsilon + \varepsilon^2} & \text{при} \quad 0 \le \varepsilon \le m_1, \\ 0 & \text{при} \quad \varepsilon \ge m_1. \end{cases} \qquad (12)$$

Последний результат был впервые получен с помощью специального метода [2], который не допускает дальнейших обобщений и уже при трех и более моментах не работает.

Пусть $k = 3$, т.е. известны три момента m_1, m_2, m_3. Рассуждая аналогично, как и в случае двух моментов, несложно получить следующий результат:

$$P(\varepsilon) = \begin{cases} 1 - p_1 & \text{при} \quad \varepsilon = 0, \\ 1 - p_1 - p_2 & \text{при} \quad 0 \le \varepsilon \le \min\left(\dfrac{m_2}{m_1}, \dfrac{m_3}{m_2}, \sqrt{\dfrac{m_3}{m_1}}\right), \\ 0 & \text{при} \quad \varepsilon \ge \min\left(\dfrac{m_2}{m_1}, \dfrac{m_3}{m_2}, \sqrt{\dfrac{m_3}{m_1}}\right). \end{cases} \qquad (13)$$

Наибольший интерес представляет случай $P(\varepsilon) = 1 - p_1 - p_2$, опуская промежуточные выкладки, получим окончательное соотношение для гарантированной оценки $P(\varepsilon)$.

$$P(\varepsilon) = \frac{3m_2m_1^2\varepsilon^2 - 3m_1m_2^2\varepsilon - m_1^3\varepsilon^3 + m_2^3}{2m_2^2\varepsilon^2 + m_3^2 + m_1m_3\varepsilon^2 - 3m_2m_3\varepsilon - m_1m_2\varepsilon^3}. \qquad (14)$$

Для произвольного числа фиксированных моментов распределения ($k>3$) гарантированные оценки находятся аналогичным образом, как в случае двух и трех известных моментов распределения, но для практического нахождения этих оценок должны быть использованы численные методы [3].

При использовании утверждения 1 остается открытым вопрос о необходимом количестве моментов для нахождения оценок показателей качества.

Пусть $\rho = (r_1, r_2, ..., r_n) \in R^n$ есть выборка значений показателя качества. Составляющие выборки $r_i > 0$ есть независимые одинаково распределенные величины из некоторого неизвестного распределения $F(t)$. Определим множество F_1 как множество всех возможных функций распределения $F(t)$, из которых может быть получена выборка X, т.е. множество функций распределения F_1 определим в виде:

$$F_1 = \{F(t) : F^{-1}(\xi_i) = r_i\}. \qquad (15)$$

Запись $F^{-1}(\xi_i)=r_i$ следует понимать как решение уравнения $F(r_i) = \xi_i$, в котором ξ_i есть реализация равномерно распределенной случайной величины r на интервале [0,1].

Пусть необходимо найти на множестве F_1 нижнюю и верхнюю оценку (границу) функции распределения $F(t)$ для заданного t=const или найти гарантированные оценки для функции распределения на множестве F_1, т.е. найти такие $F_x(t)$ и $F^x(t)$, что

$$F_x(t) = \min_{F(t)\in F_1} F(t); \qquad F^x(t) = \max_{F(t)\in F_1} F(t); \qquad (16)$$

На основе выборки ρ определим n выборочных моментов распределения $F(t)$ по следующим соотношениям:

$$m_i = \frac{1}{n}\sum_{j=1}^{n} r_j^i; \qquad (17)$$

Определим множество функций распределения F_0, у которых моменты распределения равны выборочным моментам, полученным на основе выборки ρ по соотношениям (19), т.е.

$$F_0 = \{F(t) : \int_0^{\infty} t^i dF(t) = m_i; i = 1,2,...,n\}.$$

Рассмотрим задачу: найти на множестве F_0 нижнюю и верхнюю оценку (границу) для функции распределения F(t) для заданного t, т.е. найти такие $F_x(t)$ и $F^x(t)$, что

$$F_x(t) = \min_{F(t)\in F_0} F(t); \qquad F^x(t) = \max_{F(t)\in F_0} F(t). \qquad (18)$$

Имеет место утверждение.

Утверждение 2. Задачи, определяемые соотношениями (16) и (18), эквивалентны друг другу.

Доказательство. Для доказательства утверждения необходимо доказать, что множества распределений F_0 и F_1 равны между собой. Равенство двух множеств F_0 и F_1 понимается как тождественное равенство, т.е. оно означает, что каждый элемент одного множества принадлежит другому множеству и наоборот.

Пусть $F_1(t)$ есть некоторая функция, такая, что $F_1(t) \in F_1$, т.е. выборка ρ могла быть получена из распределения $F_1(t)$. Каждое значение выборки r_i можно рассматривать как решение уравнения $F(r_i) = \xi_i$, в котором ξ_i есть реализация равномерно распределенной величины x на интервале [0,1]. Докажем, что $F_1(t) \in F_0$.

Так как F_0 есть множество функций распределения, у которых первые n моментов равны выборочным моментам, определенным по выборке ρ с помощью соотношения (17), то утверждение, что $F_1(t) \in F_0$ является очевидным.

Пусть $F_0(t) \in F_0$, $F_0(t)$ есть функция распределения, у которой первые n моментов равны выборочным моментам. Докажем, что $F_0(t) \in F_1$, т.е. может быть функцией распределения, из которой получена выборка ρ. Для этого достаточно доказать, что выборочные моменты $m = (m_1, m_1, \ldots, m_n)$ однозначно определяют выборку ρ. Докажем следующее утверждение.

Утверждение 3. Пусть $X^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$ – одно из решений (возможно комплексных) системы уравнений

$$\begin{cases} x_1 + x_2 + \ldots + x_n = b_1; \\ x_1^2 + x_2^2 + \ldots + x_n^2 = b_2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_1^n + x_2^n + \ldots + x_n^n = b_n, \end{cases} \tag{19}$$

тогда система уравнений (19) не имеет других решений, кроме тех, которые получаются действием на множество X^* группы перестановок.

Доказательство утверждения 3. Введем следующее обозначение

$$s_k = x_1^k + x_2^k + \ldots + x_n^k; \qquad k = 1,2,\ldots,n.$$

Для любого k многочлен s_k является симметричным многочленом, т.е. таким многочленом, который не меняется ни при какой перестановке переменных и поэтому, как следует из основной теоремы о симметричных многочленах [4], может быть представлен и притом единственным образом в виде многочлена от элементарных симметричных функций:

$$\sigma_1 = x_1 + x_2 + \ldots + x_n;$$

$$\sigma_2 = x_1 x_2 + x_1 x_3 + \ldots + x_{n-1} x_n;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\sigma_{n-1} = x_1 x_2 \ldots x_{n-1} + x_1 x_2 \ldots x_{n-2} x_n + \ldots + x_2 x_3 \ldots x_n;$$

$$\sigma_n = x_1 x_2 \ldots x_n;$$

Для степенных сумм s_k существует и притом единственное обратное представление, выражающее элементарные симметричные многочлены через степенные суммы. Это представление задается второй формулой Варинга [4]

$$\sigma_k = \sum \frac{(-1)^{w_1+\ldots+w_k+k}}{1^{w_1}2^{w_2}\ldots k^{w_k} w_1! w_2! \ldots w_k!} s_1^{w_1}\ldots s_k^{w_k},$$

где суммирование распространяется на все наборы неотрицательных целых чисел $w_1, w_2, \ldots, w_n$ со свойством:

$$w_1 + 2w_2 + \ldots + nw_n = k.$$

Таким образом, система уравнений (19) эквивалентна следующей системе

$$\begin{cases} c_1 = x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n; \\ c_2 = x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ c_{n-1} = x_1x_2\ldots x_{n-1} + x_1x_2\ldots x_{n-2}x_n + \ldots + x_2x_3\ldots x_n; \quad (20) \\ c_n = x_1x_2\ldots x_n; \end{cases}$$

где

$$c_k = \sum \frac{(-1)^{w_1+\ldots+w_k+k}}{1^{w_1}2^{w_2}\ldots k^{w_k} w_1! w_2! \ldots w_k!} b_1^{w_1}\ldots b_k^{w_k}; k = 1,2,\ldots,n. \quad (21)$$

Рассмотрим уравнение

$$x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = 0, \quad (22)$$

где

$$a_k = (-1)^k c_k; k = 1,2,\ldots,n.$$

Так как множество X^* удовлетворяет системе уравнений (19), а следовательно, и системе уравнений (22), то, как это следует из обратной теоремы Виета, все элементы множества X^* являются корнями уравнения (22).

Предположим далее, что система уравнений (20) кроме множества корней X^* имеет набор корней $X_* \neq X^*$. Так как по доказанному выше все элементы множества X_* являются корнями уравнения (22), то тем самым мы приходим к выводу, что все элементы множества

$$X_*^* = X^* \cup X_*$$

являются корнями уравнения (19), причем в силу неравенства $X_* \neq X^*$ число N элементов множества X_* больше чем n. Но это противоречит основной теореме алгебры, из которой следует, что уравнение (22) имеет ровно n корней. Указанное противоречие и доказывает утверждение 3.

Утверждение 3 доказывает, что справедливо утверждение 2.

Следовательно, вместо исходной задачи, определяемой соотношением (16), можно решать задачу, определяемую соотношениями (18), т.е. рассматривать задачу нахождения верхней и нижней оценки функции распределения на множестве распределений, у которых n моментов равны n выборочным моментам, найденным по соотношению (19) для выборки X, объемом n измерений.

Таким образом, для получения гарантированной оценки показателя качества, в случае малых выборок следует использовать такое количество моментов распределения (определенных в соответствии с соотношением (17)), которое равно объему выборки. При этом возникает ряд весьма интересных вопросов, связанных с получаемыми оценками, которые требуют своего дальнейшего исследования, среди них: вопрос о численном решении системы уравнений (8); вопрос о свойствах получаемых оценок и др.

Литература

1. Айвазян С. А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ. 1998. – 1022 с.

2. Гермейер Ю.Б., Иргер Д.С., Калабухова Е.П. О гарантированных оценках надежности системы при неполных сведениях о надежности элементов// ЖВМ и МФ. 1966. Т.6. №4. С.733-747.

3. Ломакин М.И. Гарантированные оценки вероятности безотказной работы в классе распределений с фиксированными моментами // Известия АН СССР, Автоматика и телемеханика. 1990. №1. С.154-161.

4. Окунев Л.Я. Высшая алгебра. – М.: Просвещение. 1966. – 335 с.