УДК 519.24

ИНТЕРВАЛЬНАЯ ОЦЕНКА ФУНКЦИИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

Е. Л. Кулешов

Дальневосточный федеральный университет,
690950, г. Владивосток, ул. Суханова, 8
E-mail: kuleshov.el@dvfu.ru

На основе асимптотики Муавра — Лапласа построена интервальная оценка функции распределения вероятностей, представляющая собой интервал со случайными границами, покрывающий истинное значение функции распределения с заданным коэффициентом доверия. Показано, что использование асимптотики вместо биномиального распределения вероятностей приводит к ошибке, величина которой допустима при небольших размерах выборки и монотонно снижается с ростом размера выборки.

Ключевые слова: интервальная оценка, функция распределения вероятностей, коэффициент доверия, критерий согласия, статистическая гипотеза.

Введение. Статистический анализ опытных данных часто сводится к проверке соответствия эмпирической функции распределения вероятностей измеренной случайной величины и предполагаемой теоретической [1–3]. Пусть $F(x) = P(\xi \le x)$, $-\infty < x < \infty$, — функция распределения вероятностей исследуемой случайной величины ξ, где $P(\xi \le x)$ — вероятность события $\xi \le x$. Представим измерения величины ξ выборкой $x_1, \ldots, x_n$ размера n. Если $\nu(x_i \le x)$ — число элементов x_i выборки таких, что каждый элемент $x_i \le x$, то эмпирическая функция распределения вероятностей (точечная оценка функции F) $\hat{F}(x) = \nu(x_i \le x)/n$. Процедура измерения случайной величины ξ и вычисления оценки $\hat{F}$ непосредственно связана со следующей вероятностной схемой Бернулли. Выполняется последовательность из n опытов, в каждом из которых измеряется значение x_i случайной величины ξ. Всякое событие $\xi \le x$ будем рассматривать как успех, вероятность которого $p = P(\xi \le x) = F(x)$. Число успехов $\nu(x_i \le x)$ имеет биномиальное распределение вероятностей: $P(\nu = k) = C_n^k p^k q^{n-k}$, $k = 0, 1, \ldots, n$, где C_n^k — число сочетаний из n по k и $q = 1 - p$ — вероятность неудачи. Пусть $\mathbf{M}$, $\mathbf{D}$ — операторы математического ожидания и дисперсии соответственно. Тогда $\mathbf{M}\nu = np$, $\mathbf{D}\nu = npq$, $\mathbf{M}\hat{F}(x) = F(x)$ и $\mathbf{D}\hat{F}(x) = F(x)[1 - F(x)]/n$. Таким образом, биномиальное распределение вероятностей оценки $\hat{F}$ зависит от неизвестной функции F. Это не позволяет использовать точечную оценку $\hat{F}$ непосредственно для анализа данных. Поэтому представленные в литературе алгоритмы проверки соответствия эмпирического и теоретического распределений вероятностей сводятся к построению на основе оценки $\hat{F}$ более сложных статистик, не зависящих от функции F.

В представленной работе построена интервальная оценка функции распределения вероятностей, которая определяет множество всех возможных пар (коэффициент доверия, доверительный интервал) и задаёт не зависящее от вида функции F вероятностное описание процедуры оценивания функции распределения по экспериментальным данным. На основе интервальной оценки предложено вполне очевидное правило проверки статистических гипотез о соответствии эмпирического распределения вероятностей и предполагаемого теоретического.

Интервальная оценка. Если вероятность успеха в n независимых испытаниях постоянна и равна p, $0 < p < 1$, то при $n \to \infty$ в соответствии с локальной теоремой Муавра — Лапласа распределение вероятностей числа успехов сходится к нормальному распределению. Поэтому можно полагать, что при $n \gg 1$ случайная величина

$$\eta = \frac{\hat{F}(x) - F(x)}{\sqrt{F(x)[1 - F(x)]}} \sqrt{n} \tag{1}$$

распределена по нормальному закону с нулевым математическим ожиданием и единичной дисперсией. Это позволяет построить интервальную оценку функции распределения.

Доверительный интервал функции $F(x)$ с заданным коэффициентом доверия $1 - \alpha$ определим условием $P(|\eta| \le \varepsilon) = 1 - \alpha$, где ε — решение уравнения $\Phi(\varepsilon) = 1 - \alpha/2$ и $\Phi(\varepsilon)$ — функция распределения вероятностей случайной величины η. Границы z_1, z_2 доверительного интервала являются решениями уравнения $\eta^2 = \varepsilon^2$ относительно $F(x)$, которое при подстановке соотношения (1) сводится к выражению

$$F^2 - F\frac{2\hat{F}n + \varepsilon^2}{n + \varepsilon^2} + \frac{n}{n + \varepsilon^2}\hat{F} = 0. \tag{2}$$

Отсюда находим

$$z_{2,1} = \frac{2\hat{F}n + \varepsilon^2}{2(n + \varepsilon^2)} \pm \frac{\sqrt{4n\varepsilon^2(1 - \hat{F})\hat{F} + \varepsilon^4}}{2(n + \varepsilon^2)}. \tag{3}$$

Тогда истинное значение функции $F(x)$ покрывается интервалом $[z_1, z_2]$ с вероятностью $1 - \alpha$. Отметим, что формула, аналогичная соотношению (3), использовалась в [4] для приближённого вычисления границ доверительного интервала вероятности успеха в схеме Бернулли.

Длина доверительного интервала

$$z_2 - z_1 = \frac{\sqrt{4n\varepsilon^2(1 - \hat{F})\hat{F} + \varepsilon^4}}{n + \varepsilon^2}. \tag{4}$$

Величина $(1-\hat{F})\hat{F}$ максимальна при $\hat{F} = 0{,}5$ и минимальна при $\hat{F} = 0$ или $\hat{F} = 1$. Поэтому для заданных n и ε максимальное значение длины доверительного интервала

$$\delta = \sqrt{\varepsilon^2/(n + \varepsilon^2)} \tag{5}$$

и минимальное значение $\min(z_2 - z_1) = \delta^2$. Если оценка $\hat{F}$ принимает значения 0, 0,5, 1, то длина доверительного интервала равна δ^2, δ, δ^2 и соответственно доверительный интервал $[z_1, z_2]$ имеет вид

$$[0, \delta^2]; \quad [0{,}5 - 0{,}5\delta,\ 0{,}5 + 0{,}5\delta]; \quad [1 - \delta^2, 1]. \tag{6}$$

Таким образом, график точечной оценки $\hat{F}$ и два отрезка длиной δ и δ^2 позволяют достаточно детально представить расположение границ доверительного интервала.

На основе интервальной оценки можно построить процедуру проверки статистической гипотезы H_0 о том, что функция $F_1(x)$ является истинной функцией распределения вероятностей. Для выбранного α по формуле (3) вычисляются границы z_1, z_2 доверительного

интервала. Если функция $F_1 \in [z_1, z_2]$, то гипотеза H_0 принимается, в противном случае H_0 отклоняется. Величину $z_2 - z_1$ доверительного интервала, а также параметр δ можно интерпретировать как некоторую ошибку или меру неопределённости относительно $F(x)$ — истинной функции распределения вероятностей. Параметр $\delta \in (\sim(1/\sqrt{n}), 1)$, поэтому его значение может служить дополнительным ориентиром правильности выбора α, поскольку слишком малое значение α приводит к большому ε и в соответствии с формулой (5) к большой ошибке δ. Рассмотрим второй пример. Пусть необходимо из множества функций распределения $F_j(x)$, $j = 1, \ldots, m$, выбрать закон распределения вероятностей, наилучшим образом соответствующий данной выборке. Из формулы (4) для каждого $j = 1, \ldots, m$ определим значение $\varepsilon = \varepsilon_j$, при котором достигается условный минимум величины $z_2 - z_1$ при условии $F_j \in [z_1, z_2]$. Затем по формуле (5) для каждого ε_j находим значение δ_j. Можно полагать, что наилучшее приближение для данной выборки обеспечивает функция $F_J(x)$, для которой $\delta_J = \min\{\delta_1, \ldots, \delta_m\}$.

В работе [5] на основе статистики Колмогорова предложена интервальная оценка функции $F(x)$ в виде интервала постоянной длительности, не зависимой от x. Такой подход имеет очевидные недостатки. Так, если для заданного коэффициента доверия $1 - \alpha$ и $\hat{F} = 0{,}5$ найти доверительный интервал $[z_1, z_2]$ и продолжить его при постоянной величине $z_2 - z_1$ на все значения $\hat{F} \in [0, 1]$, то в окрестностях $\hat{F} = 0$ и $\hat{F} = 1$ величина интервала будет слишком большой. Более того, интервал в окрестности $\hat{F} = 0$ будет содержать отрицательные числа, а в окрестности $\hat{F} = 1$ — числа больше единицы, что вообще не имеет смысла.

Анализ погрешности. Отметим, что асимптотика Муавра — Лапласа имеет небольшую погрешность, если вероятность p успеха в одном опыте не является слишком малой (близкой к нулю), а также слишком большой (близкой к единице). Так, в [6] отмечается, что если $np^{3/2} > 1{,}07$, то ошибка при использовании нормальной функции распределения вместо биномиальной не превосходит 0,05. В силу симметрии биномиального распределения относительно событий «успех»—«неудача» это условие следует дополнить неравенством $nq^{3/2} > 1{,}07$. Отсюда, например, для $n = 50, 100, 200$ получаем соответственно интервалы $[0{,}08, 0{,}92]$, $[0{,}05, 0{,}95]$, $[0{,}03, 0{,}97]$, в пределах которых может находиться вероятность p. Для значений параметра p вне указанных интервалов биномиальное распределение имеет сильную асимметрию и его хорошей аппроксимацией является асимптотика Пуассона: $P(\nu = k) = \lambda^k \mathrm{e}^{-\lambda}/k!$, где $\lambda = np$. Если $p \to 0$, то $P(\nu = 0) \to 1$ и для $k > 0$ вероятность $P(\nu = k) \to 0$, следовательно, $\mathbf{D}\nu \to 0$. Это согласуется с точным выражением для $\mathbf{D}\hat{F}$, поскольку $\mathbf{D}\hat{F} \to 0$, если $F \to 0$.

Для вывода формулы (4) использовалось нормальное распределение вместо биномиального. Поэтому можно полагать, что при $\hat{F} = 0{,}5$ вычисления по формуле (4) приводят к точному значению величины доверительного интервала, а максимальная погрешность возникает на границах для $\hat{F} = 0$ и $\hat{F} = 1$, поскольку при $F \approx 0$ и $F \approx 1$ распределение вероятностей величины $\hat{F}$ имеет значительную асимметрию и плохо аппроксимируется нормальным распределением. Минимальное ненулевое значение оценки $\hat{F} = 1/n$, поэтому практический интерес представляет анализ погрешностей формулы (4) в окрестностях значений $\hat{F} = 1/n$. Пусть в распределении Пуассона $p = 1/n$, тогда $\lambda = 1$ и

$$P(0 \le \nu < 3) = P(0 \le \hat{F} < 3/n) = \mathrm{e}^{-1} \sum_{k=0}^{2} \frac{1}{k!} = 0{,}9197. \tag{7}$$

Таким образом, точечная оценка $\hat{F}$ с вероятностью $1 - \alpha = 0{,}9197$ попадает в интервал $[0, 3/n)$, длина которого $\Delta_1 = 3/n$. Этот результат можно интерпретировать как точный,

имея в виду, что при $p = 1/n$ и большом n асимптотика Пуассона является хорошим приближением биномиального распределения. Для сравнения рассмотрим доверительный интервал $[z_1, z_2]$ функции F с коэффициентом доверия $1-\alpha = 0{,}9197$ — вероятности попадания точечной оценки $\hat{F}$ в интервал $[0, 3/n)$. Полагая $\hat{F} = 1/n$, по формуле (4) находим величину доверительного интервала $\Delta = \sqrt{4\varepsilon^2(1-1/n)+\varepsilon^4}/(n+\varepsilon^2)$, где $\varepsilon = 1{,}75$ — решение уравнения $\Phi(\varepsilon) = 1-\alpha/2$ при $\alpha = 0{,}0803$. Величины Δ и Δ_1 как функции аргумента n существенно различаются только для малых размеров выборки. При $n = 6{,}38$ величина $\Delta - \Delta_1 = 0$ и если $n > 6{,}38$, то $\Delta - \Delta_1 > 0$. Причём с ростом n разность $\Delta - \Delta_1$ достигает максимального значения 0,0527 при $n = 14{,}13$, а затем монотонно уменьшается. Например, для $n = 50, 100, 150, 200$ разность $\Delta - \Delta_1$ равна соответственно 0,0271, 0,0150, 0,0103, 0,0078. Таким образом, формулы (3)–(6) определяют интервальную оценку функции распределения вероятностей с допустимой ошибкой при сравнительно небольших размерах выборки.

Заключение. В данной работе на основе асимптотики Муавра — Лапласа построена интервальная оценка функции распределения вероятностей в виде интервала со случайными границами, в пределах которого находится истинное значение функции распределения с заданным коэффициентом доверия. Анализ погрешностей при использовании асимптотики вместо точного биномиального распределения вероятностей показал, что при этом возникает ошибка, величина которой вполне допустима при небольших объёмах выборки и монотонно снижается с ростом объёма выборки. Интервальная оценка может представлять интерес для решения задач определения вида функции распределения вероятностей исследуемой случайной величины.

СПИСОК ЛИТЕРАТУРЫ

1. **Лемешко Б. Ю., Постовалов С. Н.** Применение непараметрических критериев согласия при проверке сложных гипотез // Автометрия. 2001. № 2. С. 88–102.
2. **Клявин И. А., Тырсин А. Н.** Метод подбора наилучшего закона распределения случайной величины по экспериментальным данным // Автометрия. 2013. **49**, № 1. С. 18–25.
3. **Лапко А. В., Лапко В. А.** Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автометрия. 2010. **46**, № 6. С. 47–53.
4. **Гмурман В. Е.** Теория вероятностей и математическая статистика. М.: Высш. шк., 2003. 479 с.
5. **Кендалл М., Стюарт А.** Статистические выводы и связи. М.: Наука, 1973. 900 с.
6. **Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.** Справочник по теории вероятностей и математической статистике. М.: Наука, 1985. 640 с.

Поступила в редакцию 26 февраля 2014 г.