УДК 519.233.22, 519.233.24

ОЦЕНКА ВЕРОЯТНОСТИ НИ РАЗУ НЕ НАБЛЮДЕННОГО СОБЫТИЯ

© Гуров С.И.

Ф-т ВМиК МГУ им. М.В. Ломоносова,
119992, г. Москва, Ленинские горы, МГУ, 2-й уч. корпус, ВМиК,
e-mail: *sgur@cs.msu.ru*

Abstract. Point and interval estimators of probability of the event never observed in a series of tests under the scheme Bernoulli for which classical statistical methods give in practice often unacceptable zero estimator are offered and proved. Bibl. 7.

При испытаниях одного изделия произошёл один отказ. Какова вероятность безотказной работы изделия?

Занаучный юмор. М.: МФТИ, 2000.

Введение. Постановка проблемы

Рассматривается оценивание неслучайной, но неизвестной вероятности p осуществления некоторого случайного события X в единичном испытании. В $n > 0$ испытаниях по схеме Бернулли случайная величина числа успехов $m \in \{0,\, 1,\, \dots,\, n\}$ имеет биномиальное распределение $Bi_m(n,\, p) = \binom{n}{m} p^m (1-p)^{n-m}, \quad p \in \Theta$, где $\Theta = (0, 1)$ – пространство изменения параметра p (в данном случае – открытый одномерный интервал).

Точечная оценка $\widehat{p}_{ml}$ максимального правдоподобия величины p даётся элементарной формулой

$$\widehat{p}_{ml} \;=\; \arg\max_{p \in \overline{\Theta}} L(p,\, x) \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i \;=\; \frac{m}{n}\,. \tag{1}$$

Здесь

—: $L(p,\, x) \;=\; p^m (1-p)^{n-m}$ – функция правдоподобия величины p для биномиальной статистической модели;

—: $x = (x_1,\, \dots,\, x_n)$ – выборка, полученная в результате проведения n элементарных экспериментов по наблюдению события X, $x_i \in \{0,\, 1\}$, $i = \overline{1, n}$, причём в x имеется m значений 1 и $n - m$ значений 0;

—: $\overline{\Theta} = [0, 1]$ – замыкание множества Θ.

Данная оценка является является несмещенной, эффективной и состоятельной. Несмещенная функция оценки её дисперсии есть

$$\frac{m\,(n-m)}{n^3}\,. \tag{2}$$

При $m = 0$ говорят, что имеет место *0-событие*. В том случае формула (1) даёт нулевую точечную оценку вероятности наблюдения X, а формула (2) – нулевое оценочное значение её дисперсии. Всё это приводит к тому, что на практике оценка $\widehat{p} = 0$ часто неприемлема. В этом и состоит *основная проблема* оценки вероятности некоторого ни разу не наблюдённого события. Автору неизвестны *публикации* по данной проблеме.

Целью работы является предложение и обоснование ненулевой точечной оценки некоторого случайного события при осуществлении 0-события.

1. Доверительное оценивание

Частотный подход. В случае 0-события классические методы частотного подхода к решению задач математической статистики [1, 5] определяют нижнюю границу $p^{-}(n)$ доверительного интервала при коэффициенте доверия η как нулевую, а верхнюю $p^{+}(n)$ – как решение (относительно x) уравнения

$$I_x(1,n) \,=\, \eta\,.$$

Здесь $I_x(\cdot,\cdot)$ – отношение неполной B(бетта)-функции к полной B-функции с соответствующими параметрами. Для практических целей обычно достаточно считать $\eta = 0.95$ или $\eta = 0.99$. Таким образом, имеем

$$I_x(1,n) \,=\, n\int\limits_0^x (1-t)^{n-1}dt \,=\, 1-(1-x)^n \,=\, \eta\,,$$

откуда

$$p^{+}(n) \,=\, 1-\sqrt[n]{1-\eta}\,.$$

Так, при $\eta = 0.95$ получаем $p^{+}(10) = 0,2589$ и $p^{+}(100) = 0,0295$[1]. Для $n > 50$ можно считать $p^{+}(n) \approx 3/n$.

Использование $p^{+}(n)$ в качестве точечной оценки p, как правило, является неоправданным, дающим слишком завышенное значение вероятности: с достоверностью η будем иметь $p \leqslant p^{+}$. Однако от точечной оценки не требуется, чтобы отклонение её значения от истинного было односторонним почти всегда.

Бейесовский подход. При использования бейесовского подхода к решению статистических задач встаёт вопрос о конкретизации априорного распределения.

Будем рассматривать наиболее интересную ситуацию отсутствия результатов аналогичных экспериментов, проводимых ранее, т.е. когда использование того или иного метода восстановления априорного распределения (эмпирический бейесовский подход) невозможно. В этих случаях обычно прибегают к закону недостаточного

[1]как обычно, значения приводятся с точностью до последнего знака

основания Лапласа, который устанавливает, что если ничего не известно о параметре и он изменяется на конечном интервале, то в качестве априорного распределения принимают равномерное. Равномерное априорное распределение представим B-распределением

$$Be_p(a,\, b) \;=\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1} \tag{3}$$

с параметрами $a = b = 1$ ($\Gamma(\cdot)$ – гамма-функция). Плотность вероятности апостериорного распределения будет равняться $Be_p(1,\, n+1) = (n+1)(1-p)^n$, его математическое ожидание $\mu = 1/(n+2)$, а медиана – $med = 1 - 1/\sqrt[n]{2}$.

Бейесовскую точечную оценку определяемой величины обычно полагают равной математическому ожиданию или медиане апостериорного распределения, как доставляющие минимумы среднеквадратических потерь и среднего отклонения соответственно. Таким образом, имеем две оценки

$$\widehat{p}_{B_\mu}(n) \;=\; \frac{1}{n+2} \quad \text{и} \quad \widehat{p}_{B_{med}}(n) \;=\; 1 - \sqrt[n]{0.5}\,.$$

Поскольку $1 - \sqrt[n]{0.5} \to \ln 2/n$ при $n \to \infty$, то $\widehat{p}_{B_{med}}(n) \simeq 1/(1{,}443\, n)$. Отметим, что оценка по медиане имеет бо́льшую практическую ценность, как более робастная [7].

В любом случае ясно, что обе данные оценки являются завышенными, поскольку основаны на предположении о равномерном априорном распределении p на интервале $[0, 1]$, что мало согласуется с фактом 0-события при не слишком малых n.

2. Оценка $\widehat{p}_0$

0-событие имеет место, кода в результате проведения n элементарных экспериментов по наблюдению события X получают выборку $x^0 = (0,\, \ldots,\, 0)$ длины $n \geqslant 1$. Считаем, что любая другая информация о событии X отсутствует и не может быть дополнительно получена.

Далее для оценки вероятности p появления X в единичном эксперименте будет использоваться понятие коэффициента доверия $\eta \in (0, 1)$. Пусть $\widehat{p}$ – выбранная оценка вероятности p события X, а $P(n, \widehat{p})$ – вероятность некоторого события, связанного с наблюдённым 0-событием, и на основании которого делаются те или иные выводы, относительно X. Будем считать значение $P = P(n, \widehat{p})$ превосходящим выбранный коэффициент доверия:

$$P \;\geqslant\; \eta\,. \tag{4}$$

При этом будет иметь место непривычная зависимость $P(n, \widehat{p}) \to 1$ при $\widehat{p} \to 0$, что связано с нулевой оценкой p по (1). Поэтому здесь коэффициент доверия (не будем менять терминологию) выражает не степень достоверности некоторого события, а степень «уступки», на которую мы можем пойти для получения оценки, уклоняющийся от теоретически истинного, но неприемлемого для нас значения. В силу этого, интерес будет представлять оценка, максимально возможная при данных предположениях (наиболее удалённая от 0).

Построим две оценки вероятности 0-события, свободные от указанных выше недостатков и основанные на разных идеях.

Оценка $\widehat{p}_\eta$. При истинном значении оцениваемой вероятности p вероятность P наблюдённого 0-события есть $P = (1-p)^n$. По (4) полагаем

$$P = (1-p)^n \geqslant \eta\,,$$

откуда

$$p \leqslant \widehat{p}_\eta = 1 - \sqrt[n]{\eta} \simeq \frac{\ln(1/\eta)}{n}\,.$$

Оценка $\widehat{p}_r$. Мы будем говорить, что некоторое случайное событие X, наблюдаемое в единичном эксперименте по схеме Бернулли с вероятностью $p \in [0,1]$, определяет случайный процесс $\mathfrak{X}_p$ с дискретным временем, который и порождает выборку x^0 как реализацию этого процесса.

Идея получения оценки $\widehat{p}_r(n)$ состоит в замене рассмотрения реализации x^0 процесса $\mathfrak{X}_p$ некоторой другой его реализацией x^1, которая содержит хотя бы одно значение 1.

Построим требуемую реализацию x^1. Рассмотрим процесс $\mathfrak{X}_q$ определяемый вероятностью q наблюдения события X в единичном эксперименте по схеме Бернулли и x^1 – реализация указанного процесса. Пусть объём выборки x^1 есть $N \geqslant 1$, из которых $M \geqslant 1$ значений нулевые. Далее воспользуемся оценкой (1). Определим допустимые значения M и N из достоверности совпадения параметров $p = q$ биномиальных распределений не менее η и естественном требовании минимальности N.

Для решения поставленной задачи воспользуемся точным критерием Фишера для сравнения вероятностей, лежащих в основе двух биномиальных распределений при малых объёмах выборок [5, п. 4.6.7]. Метод основан на анализе т.н. таблиц 2×2. В нашем случае имеем таблицу

0	n	n
M	N-M	N
M	N-M+n	N+n

Применение данного критерия вызвано тем, что использования общего критерия анализа 2×2 таблиц возможно лишь при достаточно больших значениях элементов таблицы, что в нашем случае заведомо не имеет места, поскольку одно из таких значений нулевое.

Вероятность P того, что таблица порождена одним значением вероятности, будет равна

$$P = \frac{n!\,N!\,M!\,(N-M+n)!}{(N+n)!} \cdot \frac{1}{n!\,M!\,(N-M)!} = \frac{N!\,(N-M+n)!}{(N-M)!\,(N+n)!} = \frac{\binom{N}{M}}{\binom{N+n}{M}}\,.$$

Известна (см., например, [2]) асимптотика

$$\frac{\binom{n-s}{k}}{\binom{n}{k}} \sim \exp\left\{-\frac{s\,k}{n} - \frac{s^2k + sk^2}{2\,n^2}\right\},$$

справедливая при $s+k=o(n^{3/4})$ и $n\to\infty$. В нашем случае это даёт

$$P\;=\;\frac{\binom{(N+n)-n}{M}}{\binom{N+n}{M}}\;\sim\;\exp\Big\{-\frac{n\,M}{N+n}\Big(1+\frac{M+n}{N+n}\Big)\Big\}$$

с сохранением условия представления (как легко показать, для $P\to\max$ должно выполняться $M^2=o(N)$, откуда и $n+M=o((N+n)^{3/2})$ при $N\to\infty$, $n=const$). Тогда по (4) имеем

$$-\frac{n\,M}{N+n}\Big(1+\frac{M+n}{N+n}\Big)\;\lesssim\;\ln\eta\,,$$

а полагая по (1), что $\widehat{p}_r=\dfrac{M}{N}$ и считая $N\gg 1$, получим

$$n\,\widehat{p}_r\,(1+\widehat{p}_r)\;\gtrsim\;\ln\frac{1}{\eta}\,. \tag{5}$$

Отсюда, пренебрегая величиной $\widehat{p}_r{}^2$, получим $\widehat{p}_r\simeq\dfrac{\ln(1/\eta)}{n}=\widehat{p}_\eta$.

Таким образом обе построенные оценки практически совпадают. Данную оценку обозначим $\widehat{p}_0$:

$$\widehat{p}_0(n)\;=\;1-\sqrt[n]{\eta}\;\simeq\;\frac{\ln(1/\eta)}{n}\;\simeq\;\frac{1-\eta^2}{2\,\eta\,n}\;\simeq\;\frac{1-\eta}{\eta\,n}\,. \tag{6}$$

Её и предлагается принимать как точечную оценку вероятности 0-события. Приведённые асимптотики (перечисленные в порядке понижения точности с завышением оценки) справедливы для практических значений η и не слишком малых n.

Несколько более грубые рассуждения, основанные на фиксации определённого значения N, приводят, как следствие $P\to\max$, к

$$M\;=\;1\,. \tag{7}$$

Тогда $P\;=\;N/(N+n)$. По (4) имеем

$$N\;=\;\Big]\frac{\eta\,n}{1-\eta}\Big[\tag{8}$$

и по (1) сразу получаем

$$p\;\leqslant\;\widehat{p}\;=\;\frac{M}{N}\;=\;\frac{1-\eta}{\eta\,n}\,,$$

что совпадает с (6).

Ясно, что для реальных значений η и $n>3$

$$\widehat{p}_0\;<\;\widehat{p}_{B_{med}}\;<\;\widehat{p}_{B_\mu}\;<\;p^+\,.$$

3. Интервальное согласованное оценивание

Полученная точечная оценка $\widehat{p}_0$ даёт интервальную оценку p на основе принципа согласованности [3, 4]. Данный принцип, основанный на идее Э. Лемана [6], позволяет в рамках бейесовского подхода конкретизировать априорное распределение оцениваемого параметра. Метод направлен именно на малые вероятности событий.

По принципу согласованности априорное распределение выбираться, в частности, из условия совпадения бейесовской и частотной точечных оценок определяемого параметра. При этом получаемое априорное распределение (укажем что оно есть $f_{a\_priori}(p) = Be_p(1, b)$, где значение b, определяемое по принципу согласованности, см. ниже) в большей степени, чем равномерное распределение, согласуется с наблюдённым 0-событием. Далее, по принципу согласованности, апостериорное распределение есть $f_{a\_post}(p) = Be_p(1, b+n)$ и верхняя граница p_c^+ доверительного интервала $(0, p_c^+)$ для оцениваемой вероятности p, имеющей точечную оценку $\widehat{p}$, есть решение уравнения

$$I_x(1,\, n+b-1) \;=\; \eta\,.$$

Здесь параметр b определяется из условия $\widehat{p} = 1/N = 1/(b+n+1)$, и, таким образом, $b = N - n - 1$. Тогда уравнение для определения $x = \widehat{p}_c^+$ принимает вид

$$I_x(1,\, 1/\widehat{p}-2) \;=\; \eta \quad \text{или} \quad I_x(1,\, N-2) \;=\; \eta \tag{9}$$

и в последнем случае значение N берётся из (8).

Например, при $\eta = 0.95$ и $n = 10$ имеем $N = 190$, $\widehat{p}_r = 0.0053$. Уравнение (9) конкретизируется как $I_x(1,\, 188) = 0.95$, откуда по Таблице 5.2 из [1] получим $\widehat{p}_c^+ \approx 0,016$. Для сравнения: классические методы для данных параметров M и N дают доверительный интервал $(0,\, 0.024)$.

Заключение

Проблема оценки вероятности 0-события не решена окончательно. Предложенная оценка интуитивно кажется слишком заниженной при малых значениях n, когда факт 0-события не противоречит предположению о достаточно больших значениях вероятности p. *Перспективным дальнейшим исследованием* является обоснование точечной оценки вероятности 0-события для малых выборок.

Список литературы

1. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1983. – 464 с.
2. *Гаврилов Г.П., Сапоженко А.А.* Задачи и упражения по дискретной математике. – М.: ФИЗМАТЛИТ, 2004. – 416 с.
3. *Гуров С.И.* Принцип согласованности и бейесовское интервальное оценивание // Таврический вестник информатики и математики, 2003, № 2. – С. 14-27.
4. *Гуров С.И.* Интервальное оценивание на основе принципа согласованности // Вестник Тверского гос. университета. Серия «Прикладная математика», №14 (74), вып. 9, 2008. – С. 77-93.
5. *Закс Л.* Статистическое оценивание. – М.: Статистика, 1976. – 560 с.
6. *Леман Э.* Теория точечного оценивания. – М.: Наука, 1991. –448 с.
7. *Смоляк С.А., Титаренко Б.П.* Устойчивые методы оценивания: (Статистическая обработка неоднородных совкупностей). – М.: Статистика, 1980. – 208 с.

Статья поступила в редакцию 01.09.2009