УДК 519.233.22, 519.233.24

ОЦЕНКА ВЕРОЯТНОСТИ 0-СОБЫТИЯ[1]

Гуров С.И.
ВМиК МГУ им. М.В. Ломоносова, г. Москва

Поступила в редакцию 20.05.2009, после переработки 10.06.2009.

Предлагаются и обосновываются точечная и интервальная оценки вероятности события, ни разу не наблюдавшегося в серии испытаний по схеме Бернулли (0-событие), для которого классическая статистика методы дают нулевую точечную оценку, часто неприемлемую на практике. В работе предлагаются и обосновываются ненулевая точечная и интервальная оценки вероятности 0-события. Для случая 0-события даётся классификация выборок по объёму.

Point and interval probability estimator of the event never observed in a series of Bernulli trials (0-event) for which the classical statistics methods yield the zero point estimation often unacceptable in practice are offered and proved.

Ключевые слова: математическая статистика, точечное оценивание, 0-событие, интервальное оценивание, принцип согласованности, малая выборка.
Keywords: mathematical statistics, point estimator 0-event, interval estimator, consistency principle, small sample.

Введение. Постановка проблемы

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

Точечная оценка $\widehat{p}_{ml}$ максимального правдоподобия величины p даётся «классической формулой» (последнее равенство) для вычисления вероятностей, предложенной ещё в XVII в:

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

Здесь

[1]Работа выполнена при финансовой поддержке РФФИ (коды проектов 07-01-00211-а, 08-01-00405-а) и компании Intel Corporation.

$L(p,\, x) \;=\; L(p \mid m,\, n) \;=\; p^m(1-p)^{n-m}$ — функция правдоподобия для биномиальной статистической модели, где $x = (x_1,\, \ldots,\, x_n)$ — выборка, полученная в результате проведения n элементарных независимых 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 (как обычно, 1 означает наблюдение, а 0 — непоявление события в данном эксперименте);

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

Данная оценка является является несмещенной, эффективной и состоятельной. Несмещенная функция оценки её дисперсии (см. [20, Пример 17.9]) есть

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

При $m = 0$ говорят, что имеет место *0-событие* (см., например, [18, п. 4.5.4]). Точнее, под 0-событием мы будем понимать само случайное событие X, ни разу не наблюдавшееся в серии экспериментов по схеме Бернулли (а не факт получения нулевой выборки, как в [13]). 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 данной работе, являющийся развитием [13], предлагается и обосновывается ненулевая точечная оценка 0-события.

1. Доверительное оценивание

1.1 Частотный подход

В случае 0-события классические методы частотного подхода к решению задач математической статистики [4, с. 107, Таблица 5.2], [18, п. 4.5.4, (4.26)] 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 Эйлера к полной B-функции с соответствующими параметрами:

$$I_x(a, b) \;=\; \frac{1}{B(a,b)} \int\limits_0^x t^{a-1}(1-t)^{b-1} dt\,,$$

$$B(a, b) \;=\; \int\limits_0^1 t^{a-1}(1-t)^{b-1} dt \;=\; \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

($\Gamma(\cdot)$ — гамма-функция). Для практических целей обычно полагают $\eta = 0.95$; другие значения (0.99 или 0.90) используются значительно реже, в зависимости от конкретной практической ситуации. Таким образом, имеем

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

откуда

$$p^+(n) \,=\, 1 - \sqrt[n]{1-\eta}\,. \tag{3}$$

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

На практике использование $p^+(n)$ в качестве точечной оценки p будет оправданным, если наступление события X влечёт серьёзные последствия, требующих соответствующей «подстраховки» (например, при оценке различных рисков). В противном случае такая, дающая завышенное значение вероятности, оценка приводит к тому, что с близкой к 1 достоверностью будем иметь $p \leqslant p^+$. Однако от точечной оценки не требуется, чтобы отклонение её значения от истинного было односторонним почти всегда.

1.2 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 на их основе (эмпирический бейесовский подход) невозможно. В этих случаях обычно прибегают к закону недостаточного основания Лапласа, который устанавливает, что если ничего не известно о параметре и он изменяется на конечном интервале, то в качестве априорного распределения принимают равномерное.

Априорное распределение будем, как принято, выбирать из семейства сопряженных априорных распределений [15] относительно биномиальной статистической модели, которое составляют плотности B-распределений (или распределений Бернулли)

$$Be_p(a,\, b) \,=\, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1} \tag{4}$$

с параметрами $a,\, b > 0$. Равномерное распределение $U(0,1)$ на интервале $(0,1)$ есть B-распределение с параметрами $a = b = 1$. Поскольку функция правдоподобия 0-события есть $L(p \mid 0,\, n) \,=\, (1-p)^n$, то плотность вероятности апостериорного распределения будет $f(p)_{a\_post} \,=\, Be_p(1,\, n+1) \,=\, (n+1)(1-p)^n \propto L(p \mid 0,\, n)\cdot U(0,1)$. Математическое ожидание полученного апостериорного распределения, как известно, есть

$$\mu \,=\, (n+1)\int\limits_0^1 p\,(1-p)^n\, dp \,=\, \frac{I_1(2, n+1)}{n+2} \,=\, \frac{1}{n+2}\,, \tag{5}$$

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

$$\widehat{p}_{B^U_\mu}(n) \,=\, \frac{1}{n+2} \quad \text{и} \quad \widehat{p}_{B^U_{med}}(n) \,=\, 1 - \sqrt[n]{0.5}\,. \tag{6}$$

Оценка $\widehat{p}_{B^U_\mu}(n)$ отражает т.н. закон следования Лапласа [27]. 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 более робастна [26]. Ценным качеством бейесовских оценок является их независимость от дополнительного понятия доверительной вероятности.

В любом случае ясно, что для не слишком малых n обе приведённые оценки являются завышенными, поскольку основаны на предположении о равномерном априорном распределении p на интервале $(0,1)$, мало, при данном условии, согласующимся с фактом 0-события.

2. Оценка $\widehat{p}_0$

0-событие имеет место, кода в результате проведения n элементарных экспериментов по наблюдению события X получают *0-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:

$$P \geqslant \eta\,. \tag{7}$$

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 подходах.

Оценка $\widehat{p}_\eta$. При истинном значении оцениваемой вероятности p вероятность P осуществившегося 0-события есть $P = (1-p)^n$. По (7) 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 условия достоверности равенства $p = q$ не менее η.

Для решения поставленной задачи воспользуемся точным критерием Фишера сравнения вероятностей, лежащих в основе двух биномиальных распределений [18, 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

(8)

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 $P = P(N, M; n)$ того, что таблица порождена одним значением вероятности, будет равна

$$P = \frac{n!\,N!\,M!\,(N-M+n)!}{(N+n)!} \cdot \frac{1}{n!\,M!\,(N-M)!} = \frac{N!\,(N-M+n)!}{(N-M)!\,(N+n)!} = \frac{\binom{N}{M}}{\binom{N+n}{M}}\,. \tag{9}$$

Известна (см., например, [7]) асимптотика

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

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 (7) 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

$$n\,\widehat{p}_r\,(1 + \widehat{p}_r) \gtrsim \ln\frac{1}{\eta}\,. \tag{10}$$

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

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

$$\widehat{p}_0(n) \,=\, 1 - \sqrt[n]{\eta} \,\simeq\, \frac{\ln(1/\eta)}{n} \,\simeq\, \frac{1-\eta^2}{2\,\eta\,n} \,\simeq\, \frac{1-\eta}{\eta\,n}\,; \tag{11}$$

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

$$M \,=\, 1\,. \tag{12}$$

Тогда $P \,=\, N/(N+n)$, по (7) имеем

$$N \,=\, \left]\frac{\eta\,n}{1-\eta}\right[\tag{13}$$

и по (1) сразу получаем $p \leqslant \widehat{p} = M/N \,=\, (1-\eta)/(\eta\,n)$, что совпадает с (11).

Очевидно, для реальных значений η и $n > 3$

$$\widehat{p}_0(n) \,<\, \widehat{p}_{B^U_{med}}(n) \,<\, \widehat{p}_{B^U_{\mu}}(n) \,<\, p^+(n)\,.$$

3. Интервальное согласованное оценивание

Полученная точечная оценка $\widehat{p}_0$ позволяют дать интервальную оценку p на основе принципа согласованности [11, 12]. Данный принцип, основанный идее Э. Лемана [22, Гл. 4, п. 2, Пример 2.7], позволяет в рамках бейесовского подхода конкретизировать априорное распределение оцениваемого параметра. Метод ориентирован 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 $f_{a\_priori}(p) = Be_p(1,b)$ (где b — некоторый параметр) в большей степени, чем равномерное распределение, согласуется с наблюдённым 0-событием. Далее, 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 есть решение уравнения

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

По принципу согласованности, 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

$$I_x(1,\, 1/\widehat{p} - 2) \,=\, \eta \quad \text{или} \quad I_x(1,\, N-2) \,=\, \eta \tag{14}$$

и в последнем случае значение N берётся из (13).

Например, при $\eta = 0.95$ и $n = 10$ имеем $N = 190$, $\widehat{p}_0 = 0.0052$. Уравнение (14) конкретизируется как $I_x(1,\, 188) = 0.95$, откуда по Таблице 5.2 из [4] получим $\widehat{p}_c^+ \approx 0,016$. Для сравнения: классические методы для данных параметров M и N дают доверительный интервал (0, 0.024).

4. Случай малой выборки

Предложенная оценка $\widehat{p}_0$ интуитивно кажется слишком заниженной для малых значениях n. Построим оценку $\widehat{p}(n)$ для этого случая.

При совсем малых n факт 0-события не противоречит предположению о достаточно больших значениях вероятности p. Поэтому оправданным представляется следующий подход. Для некоторой N-элементной выборки с M единичными значениями найдём по (9) вероятность $P(N, M; n)$ того, что таблица (8) порождена одним значением вероятности, осредним оценку M/N в соответствии с введённым вероятностным распределением на выборках, и данное среднее значение

$$\widehat{p}_N(n) \;=\; \frac{\sum\limits_{M=0}^{N} \frac{M}{N} \cdot P(N, M; n)}{\sum\limits_{M=0}^{N} P(N, M; n)} \tag{15}$$

будем принимать за искомую оценку для данного $N \geqslant 1$.

Пользуясь, например, методом математической индукции, элементарно показывается, что

$$\sum_{M=0}^{N} P(N, M; n) \;=\; \sum_{M=0}^{N} \frac{\binom{N}{M}}{\binom{N+n}{M}} \;=\; \sum_{M=0}^{N} \frac{N!\,(N-M+n)!}{(N-M)!\,(N+n!)} \;=\; \frac{N+n+1}{n+1}\,,$$

и знаменатель (15) определён. Для числителя аналогично показывается, что

$$\sum_{M=0}^{N} \frac{M}{N} \cdot P(N, M; n) \;=\; \sum_{M=1}^{N} \frac{M}{N} \cdot \frac{N!\,(N-M+n)!}{(N-M)!\,(N+n!)} \;=\; \frac{N+n+1}{(n+1)(n+2)}\,.$$

Отсюда

$$\widehat{p}_N(n) \;=\; \frac{1}{n+2} = \widehat{p}_{B_\mu}(n)$$

(и приятной неожиданностью оказывается независимость $\widehat{p}_N(n)$ от N, освобождая нас от необходимости определять N или брать ещё одно осреднение по его значениям).

Полученный результат заставляет сделать вывод, что при малых значениях n обоснованной точечной оценкой вероятности 0-события является бейесовская оценка по математическому ожиданию при равномерном априорном распределении.

Интересно заметить, что формулу (5) можно проинтерпретировать как вычисление среднего значения вероятности p при распределении $Be_p(1,\, n+1)$, выражающем, в рамках фидуциального фишеровского подхода [32, 2, 3], степень уверенности в равенстве текущего значения p действительному значению вероятности

0-события. Кроме того, элементарно показывается, что при $n \ll N$ справедлива асимптотика

$$P(N, M; n) \;=\; \frac{\binom{N}{M}}{\binom{N+n}{M}} \;\sim\; \Bigl(1 - \frac{M}{N}\Bigr)^{n}.$$

Таким образом, (15) оказывается дискретным аналогом (5), что и объясняет совпадение оценок $\widehat{p}_N(n)$ и $\widehat{p}_{B^U_\mu}(n)$.

При “средних”, не слишком малых и не слишком больших n представляется в качестве априорного распределения использовать ступенчатую функцию, учитывающую не все возможные значения p, а лишь те, что с достоверностью η не противоречат предположению о равенстве текущего значения p действительному значению вероятности 0-события:

$$f_{a\_priori}(p) \;=\; \begin{cases} 1/\widehat{p}_\eta, & \text{если } 0 \leqslant p \leqslant \widehat{p}_\eta\,, \\ 0, & \text{иначе}\,. \end{cases}$$

В итоге получаем (ср. с (5)) оценку $\widehat{p}_{B_\eta}(n)$:

$$\widehat{p}_{B_\eta}(n) \;=\; \frac{1}{(n+2)\cdot \widehat{p}_\eta}\, I_{\widehat{p}_\eta}(2, n+1)\,, \qquad \text{где } \widehat{p}_\eta \;=\; 1 - \sqrt[n]{\eta}\,. \tag{16}$$

Например, при $\eta = 0.95$ имеем $\widehat{p}_{B_\eta}(10) = 0.0272$ и $\widehat{p}_{B_\eta}(20) = 0.0259$.

Отметим, что значения неполной B-функции для представляющих интерес значений параметров в нашем случае по таблицам (например, [4]) не определяются. При этом возможно использование формулы

$$I_x(2, n+1) \;=\; \sum_{k=2}^{n+2} C^k_{n+2}\, x^k (1-x)^{n+2-k}$$

(где значения слагаемых быстро убывают с ростом k).

5. Когда какую оценку использовать?

Сразу укажем, что мы не рассматриваем случаи, когда ясны принципы (точнее, желательность более вероятного отклонения) выбора точечной оценки в данной предметной области исследования: здесь всё зависит от того, насколько желательным или нежелательным является появление данного редкого события.

Если указанные принципы отсутствуют, то для ответа на поставленный в заголовке вопрос необходимо определиться, что понимать под «малой выборкой».

Разные авторы по-разному определяют это понятие: выборку считают малой, если её объём не превосходит 200 [19], или 50 [30], или 30 [9], или 10–20 [21], или 10–15 [25], или «меньше расчетного числа, определенного при помощи специальной номограммы достаточно больших чисел» [23], или если наблюдается «факт отсутствия устойчивости информативных свойств и статистических характеристик» [29]. Часто вообще не определяют это понятие. В БСЭ имеется статья «Малые выборки» [24], но в специализированной энциклопедии [6] аналогичной статьи нет.

Наша точка зрения была высказана в [10]: выборка считается малой, если при её обработке методами, основанными на группировке наблюдений и аппроксимационными методами, нельзя достичь заданных точности и достоверности. Для случая 0-события данное положение требуется конкретизировать, а именно, указать, при каких значениях n использовать оценку $\widehat{p}_{B^U_\mu}(n)$, и при каких — $\widehat{p}_0(n)$. Понятно, что абсолютно объективных критериев такого выбора существовать не может. Мы, однако, предложим указанное разбиение, основанное на статистической достоверности результатов.

Нижняя граница. Прежде всего, кажется ясным, что при совсем малых значениях n никаких статистических выводов делать вообще нельзя. Заметим, что при $n < 4$ имеем $\widehat{p}_{B_{med}}(n) < \widehat{p}_{B^U_\mu}(n)$ и обратное отношение при бо́льших n. Указанное значение представляется естественной границей для отделения понятия «малая выборка» от случая недостаточности данных для любых статистических выводов. Таким образом считаем, что при $1 \leqslant n \leqslant 3$ можно только констатировать факт 0-события при данном числе испытаний.

Аналогичный вывод сделан в работе [14]: *Один из основных вопросов математической статистики: какова должна быть минимально необходимая информация для получения требуемой достоверности результата. ... Если подразумевать под условиями отсутствие каких-либо ограничений по точности конечного результата статистического анализа, то ответ на поставленный вопрос дал Р. Фишер [31, 28]. Минимальное число образцов не может быть меньше 4. В противном случае, неизбежно возникает систематическая ошибка (смещение). Наличие смещения — первый признак отсутствия достаточности статистики [22]. Ряд авторов подтверждал вывод Фишера.* Добавим от себя — ср. с процитированным выше пассажем из [29].

Также при проверке гипотезы о значении отношения ξ наблюдаемых абсолютных частот a и b на основе χ^2-критерия со статистической надёжностью 95% требуется (см., например, [18, (4.33)])

$$\widehat{\chi}^2 = \frac{(\xi a - b)^2}{\xi(a+b)} < \chi^2 = 3.841\,.$$

При определении равенства вероятностей, порождающих выборки как реализации случайных процессов, полагаем $\xi = 1$, что приводит к соотношению $|a-b| < 3.841$. Поскольку применение данного критерия предполагает $0 < a \leqslant b$, вместо 0-события рассматриваем противоположное ему *полное событие*, для которого $b = n$. Таким образом, для того, чтобы с указанной надёжностью считать выборку с a значениями 1 другой реализацией того же случайного процесса, что и породивший нулевую выборку $x^0 = (0, \ldots, 0)$ той же длины, необходимо, чтобы значение $n - a$ не превосходило 3[2]. Таким образом, различие может быть статистически определено лишь при длине выборки $4 \leqslant n$.

Верхняя граница. Верхнюю границу для малой выборки в случае 0-события естественно установить равной N по (13) при $n = 1$. Это значение практически совпадает с условием $\widehat{p}_{B^U_\mu}(n) < 1 - \eta$ для того же значения достоверности η. Меньшие значения n не дают возможности статистически достоверно определить совпадение вероятностей единичных событий, связанных с данными выборками.

[2]Интересно, что граничное значение 3 (т.н. «бонгартовская тройка») часто возникает в комбинаторных исследованиях на неслучайность событий [5, 17, 16].

При $\eta = 0.95$ для искомой границы получаем значение $n = 19$. Заметим, что оно влечёт $\widehat{p}_{B^U_\mu}(n) \approx 5\%$. Отметим, что на практике значение 5% часто принимают за границу понятия «редкого события» (см., например, [1]).

В результате предлагается следующая классификация 0-выборок по объёму с указанием точечной оценки $\widehat{p}$ вероятности 0-события.

n	Тип 0-выборки, $\widehat{p}$
$1, 2, 3$	никаких оценок дать нельзя
от 4 до 19	«малая» 0-выборка, $\widehat{p} = \widehat{p}_{B^U_\mu}(n)$
более 20	«большая» 0-выборка, $\widehat{p} = \widehat{p}_0(n)$

Резкий скачок значения предложенной оценки при переходе от «малой» выборки к «большой» вызван экстремальностью самого исследуемого понятия: появляется возможность с достаточной достоверностью статистически фиксировать совпадение или несовпадение вероятностей единичных событий, определяющих выборки по схеме Бернулли. В малых выборках осуществление 0-события представляется вполне возможным даже при p, не обязательно близких к 0, в то время как в больших выборках оно с необходимостью означает либо крайне малую величину p, либо вообще невозможность события X.

Автор выражает глубокую признательность Ю.И. Журавлёву за неизменную поддержку и В.Е. Бенингу за ценные консультации.

Список литературы

[1] Бейли Н. Математика в биологии и медицине. — Электронный ресурс http://www.biometrica.tomsk.ru/beili_2_2.htm

[2] Бернштейн С.Н. О «доверительных» вероятностях Фишера /С.Н. Бернштейн. Собрание сочинений. Том IV. Теория вероятностей и математическая статистика (1911-1946). — М.: Наука, 1964. — С. 386–393.

[3] Большев Л.Н. Комментарий к работе С.Н. Бернштейна «О “доверительных” вероятностях Фишера» / С.Н. Бернштейн. Собрание сочинений. Том IV. Теория вероятностей и математическая статистика (1911-1946). — М.: Наука, 1964. — С. 566–569.

[4] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. — М.: Наука, 1983.

[5] Бонгард М.М. Проблема узнавания. — М.: Наука, 1967.

[6] Вероятность и математическая статистика. Энциклопедия. — М.: Научное изд-во Большая Российская энциклопедия, 1999.

[7] Гаврилов Г.П., Сапоженко А.А. Задачи и упражнения по дискретной математике. — М.: ФИЗМАТЛИТ, 2004.

[8] Гаскаров Д.В., Шаповалов В.И. Малая выборка. — М.: Статистика, 1978.

[9] Гмурман В.Е. Теория вероятностей и математическая статистика. — М.: Высшая школа, 1977.

[10] Гуров С.И. Оценка надёжности классифицирующих алгоритмов. — М.: Издательский отдел ф-та ВМиК МГУ, 2002 г.

[11] Гуров С.И. Принцип согласованности и бейесовское интервальное оценивание // Таврический вестник информатики и математики, 2003, Вып. 2. — С. 14-27.

[12] Гуров С.И. Интервальное оценивание на основе принципа согласованности // Вестник Тверского государственного университета. Серия «Прикладная математика», № 14 (74), вып. 9, 2008. — С. 77-93.

[13] Гуров С.И. Оценка вероятности ни разу не наблюдённого события // Таврический вестник информатики и математики, 2009, Вып. 2 (в печати).

[14] Гусев А.В., Лидский Э.А., Мироненко О.В. Малые выборки при оценке работоспособности и надежности электронных компонентов. Часть I // Chip news — Инженерная микроэлектроника, 2002, № 1, — С. 52–56. (См. также электронный ресурс http://www.chipinfo.ru/literature/chipnews/about.html.)

[15] Де Гроот М. Оптимальные статистические решения. — М.: Мир, 1974.

[16] Донской В.И., Башта А.И. Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992.

[17] Закревский А.Д. Логика распознвания. — М.: Едиториал УРСС, 2003.

[18] Закс Л. Статистическое оценивание. — М.: Статистика, 1976.

[19] Кендал М., Стюарт А. Теория распределений. — М.: Наука, 1966.

[20] Кендал М., Стюарт А. Статистические выводы и связи. — М.: Наука, 1973.

[21] Кремер Н.Ш. Теория вероятностей и математическая статистика. — М.: ЮНИТИ-ДАНА, 2000.

[22] Леман Э. Теория точечного оценивания. — М.: Наука, 1991.

[23] Методы статистического анализа и обработка малого числа наблюдений при контроле качества и надежности приборов и машин. — Л.: Изд. ЛДНТП, 1974.

[24] Прохоров Ю. В. Малая выборка / БСЭ. — М.: Сов. энциклопедия, 1969–1978.

[25] Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. — М.: Наука, 1965.

[26] Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: (Статистическая обработка неоднородных совкупностей). — М.: Статистика, 1980.

[27] Феллер В. Введение в теорию вероятностей и её прилжения, т. 1–2 — М.: Мир, 1984.

[28] Фишер Р. Статистические методы для исследователей. — М.: Гостехиздат, 1958.

[29] Фурсов В.А. Идентификация моделей систем формирования изображений по малому числу наблюдений. — Самара: Самар. гос. аэрокосм. ун-т., 1998.

[30] Шор Я.Б. Статистические выводы анализа и контроля надежности и качества. — М.: Сов. радио, 1962.

[31] Fisher R.A. On the mathematical foundations of theoretical statistics // Phil. Trans. Roy. Soc., Ser. A, 1921, v. 222.

[32] Fisher R.A. The fiducial argument in statistical inference // Annals of Eugenics, Vol. 5, 1935. — С. 391–398.