УДК 519.233+519.226+51-76

АНАЛИЗ ДИСКРЕТНЫХ ДАННЫХ. ОСНОВНЫЕ ПОНЯТИЯ

В. Г. Панов

Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

Обзорная статья посвящена методам анализа дискретных данных. Основное внимание уделяется точным методам, основанным на представлении дискретных вероятностных распределений в виде степенных распределений (power series distributions). Рассматриваемые примеры относятся к задачам медицины и биологии.

Ключевые слова: дискретные распределения; степенные распределения; точные методы статистического анализа; статистические гипотезы; доверительные области; мощность критерия; вероятность покрытия.

1. Введение

Исследования в медицине [1–5], эпидемиологии [6–10], биологии [11–16], экономике [17–21] и других науках постоянно сталкиваются с необходимостью анализа дискретных данных, для которых, разумеется, разработаны разнообразные методы, однако в большинстве своем они предполагают использование приемов, связанных с большими выборками. Если же выборка существенно ограничена или аппроксимация асимптотическими распределениями недостаточно хорошая, то использование таких методов сомнительно. В этом случае желательно иметь средства точного анализа данных, не обращающиеся к предельным распределениям того или иного типа.

Исторически именно точные методы были первыми, в которых математика применялась для анализа случайных данных, это классическая задача оценки вероятности события при наличии равновероятных исходов и схема повторных независимых испытаний Бернулли [22, 23]. В XX веке актуальность тематики точного анализа дискретных данных была поддержана необходимостью развития непараметрических методов анализа непрерывных данных [24–26]. Однако основное развитие методов теории вероятностей было связано с непрерывными распределениями и предельными теоремами.

Важным событием в возрождении интереса к точным методам и фактически становлением этой тематики как отдельного направления статистических исследований стала почти одновременная публикация работ по точному анализу таблиц сопряженности 2×2 в работах Фишера [27, 28], Ирвина [29] и Йейтса [30]. Однако развитие этих идей сдерживалось значительной вычислительной сложностью приложений полученных теоретических результатов. В дальнейшем были разработаны методы, ускоряющие такие вычисления (на основе быстрого преобразования Фурье), а также были разработаны новые алгоритмы таких вычислений. Увеличение вычислительных возможностей и появление новых алгоритмов сделали методы точного анализа более доступными, что было реализовано в нескольких компьютерных программах (например, StatXact [31], а также отдельные модули точных статистических вычислений в программах SPSS, SAS, R и др.).

Несмотря на значительную потребность в методах точного анализа дискретных данных, изложение темы точных статистических вычислений в учебниках и монографиях (если таковое вообще есть) до сих пор остается на уровне примеров и «рецептов», за исключением точного теста Фишера (см., например, [32, 33]).

Ниже основное внимание будет уделено именно точным методам анализа дискретных случайных величин, хотя и некоторые приемы асимптотических вычислений также будут приведены.

Технически аппарат, используемый для точного анализа дискретных случайных величин, восходит к работе Р. А. Фишера [28], в которой вероятность попадания в хвост распределения представлялась как отношение двух многочленов, причем числитель был частью знаменателя. Это отношение Фишер использовал для вычисления того, что сейчас называется верхней границей доверительного интервала. Двусторонний вариант такого отношения был предложен в работе [34]. В дальнейшем такое представление использовалось и в других работах, см., например, [35–37]. В работе [38] этот подход был систематизирован и развит с приложениями к различным задачам статистического анализа В частности, в этой работе вводится понятие доказательной функции (evidence function, или функции *p*-значения), которая объединяет три ключевых средства анализа данных: *p*-значение, точечную оценку и доверительный интервал. С математической точки зрения это понятие и сопутствующие конструкции довольно элементарны и не требуют знаний, выходящих за пределы вузовского курса алгебры. Однако оно предоставляет эффективную теоретическую и вычислительную среду для точного анализа дискретных данных.

2. Основные целочисленные случайные величины

Использование методов теории вероятностей в анализе данных основано на предположении о случайности рассматриваемых исходов. В обсуждаемых ниже вопросах предполагается, что случайность исхода описывается некоторой дискретной случайной величиной. Например, числа в таблицах сопряженности представляют собой дискретные случайные величины, принимающие целочисленные значения. Основные распределения такого типа рассмотрены ниже.

2.1. Биномиальное распределение

Допустим, что имеется однородная совокупность, для каждого объекта которой некоторая случайная величина $\mathcal{Y}$ принимает значения 0 или 1 с одной и той же вероятностью $\pi = \mathrm{P}(\mathcal{Y} = 1)$ для каждого объекта (испытания Бернулли; π называется вероятностью успеха). Пусть из этой совокупности извлекается выборка объем n, и $\mathcal{Y}_i$ означает значение случайной величины $\mathcal{Y}$ для i-го объекта, причем будем предполагать, что значения $\mathcal{Y}_i$ для разных объектов независимы. Тогда случайная величина $\mathcal{B} = \sum_i \mathcal{Y}_i$ есть (случайное) число объектов, для которых $\mathcal{Y} = 1$ (биномиальная случайная величина). Плотность распределения этой случайной величины задается равенством

$$\mathrm{P}(\mathcal{B} = m; \pi) = C_n^m \pi^m (1-\pi)^{n-m},\ m = 0, 1, 2, \ldots, n, \quad C_n^m = \frac{n!}{m!(n-m)!} \tag{1}$$

Нетрудно проверить, что сумма двух независимых биномиальных распределений $\mathcal{A}$ и $\mathcal{B}$ с одинаковой вероятностью успеха π, но разным числом испытаний n и m соответственно, также имеет биномиальное распределение с вероятностью успеха π и числом испытаний $n + m$.

Пример 1. Допустим, что для какого-то лекарства исследуется наличие побочных эффектов. Пусть π – неизвестная вероятность того, что у пациентов данной группы будет обнаружено побочное действие лекарства. Будем предполагать, что распределение числа пациентов, у которых будет обнаружен побочный эффект, подчиняется биномиальному распределению $\mathcal{B}(n, \pi)$, где n – число испытуемых.

2.2. Полиномиальное распределение

Если в схеме повторных независимых испытаний исходов будет не два, а больше, например, m, то для описания исхода такого испытания для каждого объекта необходимо вводить m случайных величин $\mathcal{Y}_{ij}$ или одну m-мерную случайную величину $\mathcal{Y}_i = (\mathcal{Y}_{i1}, \mathcal{Y}_{i2}, ..., \mathcal{Y}_{im})$, определяемую условиями $\sum_j \mathcal{Y}_{ij} = 1$ и $\mathcal{Y}_{ij} = 1$, если для i-го объекта случился j-й исход. Очевидно, выполняется равенство $\sum_{ij} \mathcal{Y}_{ij} = n$ – полное число проведенных испытаний (или, что тоже самое, число испытуемых объектов). При этом ясно, что независимых случайных величин из $(\mathcal{Y}_{i1}, \mathcal{Y}_{i2}, ..., \mathcal{Y}_{im})$ будет $m - 1$, так как $\sum_j \mathcal{Y}_{ij} = 1$.

Как и для биномиального распределения, будем считать, что для каждого объекта проводимых испытаний вероятность данного исхода из m возможных одна и та же $\pi_j = \mathrm{P}(\mathcal{Y}_{ij} = 1)$, $j = 1, 2, \ldots, m$. Обозначим $n_j = \sum_i \mathcal{Y}_{ij}$ – количество объектов, имеющих исход на уровне j, $\mathcal{M} = \sum_i \mathcal{Y}_i$ – многомерная случайная величина, j-я компонента которой есть биномиальная случайная величина с вероятностью успеха π_j, $\sum_j \pi_j = 1$ (полиномиальное, или мультиномиальное распределение). Тогда плотность полиномиального распределения имеет вид

$$\mathrm{P}(\mathcal{M} = (n_1, n_2, \ldots, n_m); \pi_1, \pi_2, ..., \pi_m) = \frac{n!}{n_1! n_2! ... n_m!} \pi_1^{n_1} \pi_2^{n_2} ... \pi_m^{n_m}, \quad \sum_i n_i = n \tag{2}$$

Пример 2. В условиях примера 1 можно рассматривать более подробно, какие именно побочные явления могут появиться от применения исследуемого лекарства. Например, пусть с вероятностями π_1, π_2, π_3 в данной группе испытуемых может проявиться только один из симптомов: головная боль, жар или мышечный спазм, а с вероятностью π_4 не будет никаких побочных явлений, $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$. Тогда распределение числа появлений того или иного симптома в группе из n испытуемых описывается полиномиальным распределением (2).

2.3. Отрицательное биномиальное распределение

Пусть $\mathcal{Y}_i$, $i = 1, 2, 3, \ldots$ – последовательность независимых испытаний Бернулли с одной и той же вероятностью успеха π, $r \geq 1$ – фиксированное целое число. Случайная величина $\mathcal{NB}$ определяется как число успехов, которые были до наступления r-й неудачи. Иначе говоря, если $k + r$ – номер испытания, в котором наступила r-я неудача, то $\mathcal{NB} = k$. Плотность отрицательного биномиального распределения задается равенством

$$\mathrm{P}(\mathcal{NB} = k; \pi) = C_{k+r-1}^{r-1} \pi^k (1-\pi)^r, k = 0, 1, 2, ... \tag{3}$$

Заметим, что отрицательное биномиальное распределение определяется различно и приводит к различным выражениям для плотности вероятности в зависимости от того, означает π вероятность успеха или неудачи и продолжаются испытания до r-й неудачи или r-го успеха, а также от того, что подсчитывает случайная величина – число успехов/неудач или число испытаний. Например, если в той же постановке случайная величина $\mathcal{X}$ подсчитывает число неудач в последовательности испытаний Бернулли с вероятностью успеха π, проводимых до r-го успеха, то, очевидно, что $\mathcal{X}$ имеет отрицательное биномиальное распределение и

$$\mathrm{P}(\mathcal{X}=k;\pi)=C_{k+r-1}^{r-1}\pi^r(1-\pi)^k,\ k=0,\ 1,\ 2,\ldots$$

Но если $\mathcal{X}$ подсчитывает число испытаний Бернулли с вероятностью успеха π до наступления r-го успеха (неудачи), то также получим отрицательное биномиальное распределение (так как $n = k + r$)

$$\mathrm{P}(\mathcal{X}=n;\pi)=C_{n-1}^{r-1}\pi^r(1-\pi)^{n-r},\ n=r,\ r+1,\ r+2,\ldots$$

$$(\mathrm{P}(\mathcal{X}=n;\pi)=C_{n-1}^{r-1}\pi^{n-r}(1-\pi)^{r},\ n=r,\ r+1,\ r+2,\ldots)$$

Если же $\mathcal{X}$ подсчитывает число успехов в заданном числе n испытаний Бернулли, то получаем биномиальное распределение (1).

Пример 3. Пусть в условиях примера 1 испытания проводятся последовательно на однородной группе испытуемых до обнаружения побочных эффектов лекарства ровно у k испытуемых. Число требуемых для этого испытаний имеет отрицательное биномиальное распределение с вероятностью «успеха» π (здесь это вероятность появления побочных эффектов).

2.4. Гипергеометрическое распределение

Пусть имеется N объектов, из которых ровно n имеют некоторую интересующую нас характеристику (некоторое заболевание, дефект и т. п.). Из этой совокупности извлекается (без возвращения) выборка из m элементов. Случайная величина $\mathcal{HG}$ определяется как число объектов, имеющих заданную характеристику в случайной выборке объема m. Плотность распределения этой случайной величины равна

$$\mathrm{P}(\mathcal{HG}=k)=\frac{C_n^k\cdot C_{N-n}^{m-k}}{C_N^m},\ k=0,\ 1,\ 2,\ldots,\ m \tag{4}$$

Очевидно, гипергеометрическое распределение можно получить как условное распределение из двух одинаковых независимых биномиальных распределений. А именно, если $\mathcal{A}$ и $\mathcal{B}$ независимые биномиальные распределения с одной и той же вероятностью успеха и одним и тем же числом испытаний n, то распределение одного из них при фиксированной сумме этих случайных величин имеет гипергеометрическое распределение

$$\mathrm{P}(\mathcal{A}=k\,|\,\mathcal{A}+\mathcal{B}=m)=\frac{C_n^k\cdot C_n^{m-k}}{C_{2n}^m}$$

Пример 4. В группе из N больных некоторым определенным заболеванием имеется n больных с легким течением и $N - n$ со средним или тяжелым. Из этой группы случайным образом выбрано m больных. Тогда число больных в этой выборке с легким течением заболевания имеет гипергеометрическое распределение.

2.5. Распределение Пуассона

Случайная величина $\mathcal{P}$ имеет распределение Пуассона, если ее плотность задается равенством

$$P(\mathcal{P} = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \ldots, \tag{5}$$

где параметр $\lambda > 0$.

Это распределение используется для представления вероятности наступления редких событий, например, числа ДТП на некотором участке дороги в течение дня, числа новых случаев некоторого заболевания с низкой распространенностью и т. п. Однако фактически его можно рассматривать просто как одно из целочисленных дискретных распределений. Использование этого распределения для описания именно редких событий связано с тем, что можно строго вывести выражение для плотности распределения числа наступивших событий, происходящих в течение некоторого времени или в некоторой пространственной области, предполагая, что

(1) вероятность появления события в малом интервале времени (или области пространства) пропорциональна размеру этого интервала;

(2) вероятность двух и более событий в малом интервале времени (области пространства) пренебрежимо мала;

(3) события в непересекающихся интервалах времени (областях пространства) независимы.

При этих предположениях искомая плотность будет именно плотностью распределения Пуассона. В такой трактовке параметр λ определяет интенсивность наступления событий в течение рассматриваемого промежутка времени (среднее число появлений подсчитываемых событий в течение заданного промежутка времени).

Одно из полезных свойств распределения Пуассона состоит в его аддитивности, т. е. сумма попарно независимых распределений Пуассона с параметрами λ_i также распределена по закону Пуассона с параметром $\sum_i \lambda_i$.

Пример 5. Пусть в течение суток на станцию скорой помощи поступает в среднем 12 вызовов в связи с ожогами детей. Тогда средняя интенсивность таких обращений равна 0,5 в час. Тогда вероятность поступления k таких обращений в течение часа вычисляется по формуле (5).

Интересно также, что биномиальное распределение можно получить как условное распределение одного распределения Пуассона при условии постоянства его суммы с другим независимым распределением Пуассона. А именно, если $\mathcal{P}_1$ и $\mathcal{P}_2$ – распределения Пуассона с параметрами λ_1 и λ_2, то

$$P(\mathcal{P}_1 = k \mid \mathcal{P}_1 + \mathcal{P}_2 = m) = \frac{m!}{k!(m-k)!} \pi^k (1-\pi)^{m-k}, \quad \pi = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Важно также, что распределение Пуассона является предельным для отрицательного биномиального при $r \to \infty$. Довольно часто в биологических задачах именно отрицательное биномиальное распределение оказывается более подходящим, чем пуассоновское, см., например, [39–41].

3. Степенно́е представление целочисленной случайной величины

Любую дискретную случайную величину можно считать принимающей значения 0, 1, 2,… (до некоторого *n* или до бесконечности), делая при необходимости перекодировку ее значений. Большинство таких целочисленных случайных величин можно представить в следующем степенном виде.

Пусть ϕ – положительный параметр, $f(\phi)$ – сходящийся степенной ряд, в частности, возможно, полином, $f(\phi)=\sum_{k=0}^{\infty} c(k)\phi^k$, причем $c(k) \ge 0$, k = 0, 1, 2,… Обозначим $\Omega = \{k \mid c(k) > 0\} \ne \varnothing$. Тогда

$$f(\phi)=\sum_{k\in\Omega} c(k)\phi^k$$

Функция $f(\phi)$ определяет дискретную (целочисленную) случайную величину κ, определенную на множестве Ω (называемом *носителем* случайной величины κ, или выборочным пространством для этой случайной величины), следующим равенством (κ зависит от параметра ϕ)

$$\mathrm{P}(\kappa = k; \phi) = \frac{c(k)\phi^k}{f(\phi)},\ k\in\Omega \qquad (6)$$

Очевидно, любую целочисленную случайную величину можно представить в виде (6), считая, что $\phi = 1$, если она не зависит от какого-либо положительного параметра. Заметим также, что функция $f(\phi)$ является просто ненормированной производящей функцией случайной величины κ, которую можно назвать *производящим степенным рядом*, или *производящим полиномом* (если множество Ω конечно) для случайной величины κ. Такое задание случайной величины называется степенным (в англ. литературе *power series distribution* – см. [42, 43]).

Иногда удобно перейти к экспоненциальной форме функции $f(\phi)$ с параметром $\beta = \ln(\phi)$

$$h(\beta)=f(e^\beta)=\sum_{k\in\Omega} c(k)e^{\beta k};$$

$$\mathrm{P}(\kappa = k; \beta) = \frac{c(k)e^{\beta k}}{h(\beta)}$$

С помощью функции $f(\phi)$ (или его экспоненциального варианта $h(\beta)$) можно вычислить числовые характеристики κ следующим образом.

Теорема 1. Если среднее μ и дисперсия σ^2 случайной величины κ конечны, то

$$\mu = \phi\frac{d}{d\phi}\left[\ln f(\phi)\right] = \frac{f_*(\phi)}{f(\phi)} = \frac{h'(\beta)}{h(\beta)};$$

$$\sigma^2 = \mu + \phi^2\frac{d^2}{d\phi^2}\left[\ln f(\phi)\right] = \frac{f_{**}(\phi)}{f(\phi)} - \left(\frac{f_*(\phi)}{f(\phi)}\right)^2 = \frac{h''(\beta)}{h(\beta)} - \left(\frac{h'(\beta)}{h(\beta)}\right)^2,$$

где $\quad f_*(\phi)=\sum_{k\in\Omega} kc(k)\phi^k, \quad f_{**}(\phi)=\sum_{k\in\Omega} k^2c(k)\phi^k.$

Доказательство получается прямым вычислением.

Важное свойство порождающих степенных рядов (как и общих производящих функций) состоит в следующем.

Теорема 2. Если $\mathcal{K}_1$, $\mathcal{K}_2$,…, $\mathcal{K}_N$ – независимые случайные величины, определенные степенными рядами $f_1(\phi)$, $f_2(\phi)$,…, $f_N(\phi)$, заданными, соответственно, на носителях Ω_1, Ω_2, …, Ω_N, то случайная величина $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2 + \ldots + \mathcal{K}_N$ определяется степенным рядом

$$f(\phi) = \prod_{k=1}^{N} f_k(\phi)$$

Доказательство проводится по индукции (см., например, [38, с. 9]).

Пример 6. Приведем степенные представления рассмотренных ранее случайных величин.

(1) *Биномиальное распределение* $\mathcal{B}$ зависит от параметра π (вероятность успеха), относительно которого формула (1) принимает вид

$$\mathrm{P}(\mathcal{B} = m; \pi) = C_n^m \pi^m (1-\pi)^{n-m} = C_n^m (\frac{\pi}{1-\pi})^m (1-\pi)^n = C_n^m (\frac{\pi}{1-\pi})^m (1+\frac{\pi}{1-\pi})^{-n}$$

Следовательно, для шанса успеха $\phi = \dfrac{\pi}{1-\pi}$ это равенство можно представить в виде

$$\mathrm{P}(\mathcal{B} = m; \phi) = \frac{C_n^m \phi^m}{(1+\phi)^n}, \quad f(\phi) = (1+\phi)^n$$

(2) *Распределение Пуассона*. Здесь очевидно, что это случайная величина, определяемая степенным рядом, а именно рядом для e^{λ}

$$\mathrm{P}(\mathcal{P} = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{\lambda^k / k!}{e^{\lambda}}, k = 0, 1, 2, \ldots,$$

(3) *Отрицательное биномиальное распределение*. Из равенства $\mathrm{P}(\mathcal{NB} = k; \pi) = C_{k+r-1}^{r-1} \pi^k (1-\pi)^r, k = 0, 1, 2, \ldots$ сразу получаем

$$\mathrm{P}(\mathcal{NB} = k; \phi) = \frac{C_{k+r-1}^{r-1} \phi^k}{f(\phi)}, k = 0, 1, 2, \ldots,$$

где $\quad \phi = \pi, f(\phi) = (1-\phi)^{-r} = \sum_{u=0}^{\infty} C_{u+r-1}^{r-1} \phi^u.$

(4) *Гипергеометрическое распределение*. Для получения степенного представления этой случайной величины удобно перейти к ее обобщению (нецентральному гипергеометрическому распределению) на основе показанной выше связи гипергеометрического и биномиального распределений. А именно, пусть $\mathcal{A}(n, \pi_1)$ и $\mathcal{B}(m, \pi_0)$ – взаимно независимые биномиальные распределения. Рассмотрим распределение одного из них при заданной их сумме $\mathrm{P}(\mathcal{A} = k \mid \mathcal{A} + \mathcal{B} = l)$.

$$\mathrm{P}(\mathcal{A} = k \mid \mathcal{A} + \mathcal{B} = l) = \frac{\mathrm{P}(\mathcal{A} = k)\mathrm{P}(\mathcal{B} = l-k)}{\mathrm{P}(\mathcal{A} + \mathcal{B} = l)} = \frac{\mathrm{P}(\mathcal{A} = k)\mathrm{P}(\mathcal{B} = l-k)}{\sum_{u \in \Omega} \mathrm{P}(\mathcal{A} = u)\mathrm{P}(\mathcal{B} = l-u)} =$$

$$= \frac{\dfrac{C_n^k \phi_1^k}{(1+\phi_1)^n} \cdot \dfrac{C_m^{l-k} \phi_0^{l-k}}{(1+\phi_0)^m}}{\sum_{u \in \Omega} \dfrac{C_n^u \phi_1^u}{(1+\phi_1)^n} \cdot \dfrac{C_m^{l-u} \phi_0^{l-u}}{(1+\phi_0)^m}} = \frac{C_n^k C_m^{l-k} (\dfrac{\phi_1}{\phi_0})^k \phi_0^l}{\sum_{u \in \Omega} C_n^u C_m^{s-u} (\dfrac{\phi_1}{\phi_0})^u \phi_0^s} = \frac{c(k)\phi^k}{\sum_{u \in \Omega} c(u)\phi^u}, \quad f(\phi) = \sum_{u \in \Omega} c(u)\phi^u,$$

где $\quad \Omega = \{l_1, l_1+1, \ldots, l_2\}, l_1 = \max(0, l-m), l_2 = \min(n, l), \phi = \phi_1 / \phi_0 = \dfrac{\pi_1 / (1-\pi_1)}{\pi_0 / (1-\pi_0)}, c(k) = C_n^k C_m^{l-k}.$

Таким образом, нецентральное гипергеометрическое распределение является степенным, а обычное гипергеометрическое распределение получается из этого представления при $\phi = 1$. Заметим также, что параметр ϕ, от которого зависит нецентральное гипергеометрическое распределение, представляет собой отношение шансов для параметров распределений $\mathcal{A}(n, \pi_1)$ и $\mathcal{B}(m, \pi_0)$.

(5) *Полиномиальное распределение*. Это вероятностное распределение также является степенным, однако для его представления в этом виде необходимо рассматривать многочлен от многих переменных.

Пусть $\boldsymbol{\phi} = (\phi_1, \phi_2,\ldots, \phi_N)$ – вектор параметров, $\boldsymbol{\mathcal{K}} = (\mathcal{K}_1, \mathcal{K}_2,\ldots, \mathcal{K}_N)$ – вектор из N дискретных целочисленных случайных величин, $\mathcal{K}_i = \mathcal{K}_i(\phi_i)$, принимающий значения из не более чем счетного N-мерного множества Ω. Обозначим (векторное обозначение экспоненты многомерных векторов)

$$\phi^{\boldsymbol{k}} = \prod_{i=1}^{N} \phi_i^{k_i}, \quad \boldsymbol{k} = (k_1, k_2, \ldots, k_N)$$

Важно отметить, что $\boldsymbol{\phi}^{\boldsymbol{k}}$ – скаляр, а не вектор. Функция

$$f(\phi) = \sum_{\boldsymbol{u} \in \Omega} c(\boldsymbol{u}) \phi^{u}$$

представляет собой степенной ряд (в частности, полином) от N переменных. Если $\boldsymbol{k}$ – N-мерный вектор из Ω, то N-мерная случайная величина $\boldsymbol{\mathcal{K}}$ называется *степенной*, если выполняется равенство

$$\mathrm{P}(\mathcal{K} = \boldsymbol{k}; \phi) = \frac{c(\boldsymbol{k}) \phi^{k}}{f(\phi)}$$

Соответствующее экспоненциальное представление имеет вид (штрих означает транспонирование)

$$\mathrm{P}(\mathcal{K} = \boldsymbol{k}; \phi) = \frac{c(\boldsymbol{k}) \exp(\boldsymbol{\beta} \boldsymbol{k}')}{h(\boldsymbol{\beta})};$$
$$h(\boldsymbol{\beta}) = \sum_{\boldsymbol{u} \in \Omega} c(\boldsymbol{u}) \exp(\boldsymbol{\beta} \boldsymbol{u}')$$

В частности, для полиномиального распределения (2) получаем его представление в виде степенной случайной величины

$$\mathrm{P}(\mathcal{M} = (n_1, n_2, \ldots, n_m); \pi_1, \pi_2, \ldots, \pi_m) = \frac{n!}{n_1! n_2! \ldots n_m!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_m^{n_m}.$$

В качестве иллюстрации полезности полиномиального представления приведем выражения для вероятности хвостов распределения целочисленной случайной величины. Пусть $\mathcal{K}$ – целочисленная случайная величина, зависящая от параметра ϕ и определенная степенным рядом $f(\phi)$ по формуле (6). Тогда, очевидно, имеют место равенства

$$\begin{aligned} \mathrm{P}(\mathcal{K} = k; \phi) &= \frac{f(\phi; k)}{f(\phi)} \\ \mathrm{P}(\mathcal{K} \le k; \phi) &= \frac{f(\phi; \le k)}{f(\phi)} \\ \mathrm{P}(\mathcal{K} \ge k; \phi) &= \frac{f(\phi; \ge k)}{f(\phi)}, \end{aligned} \tag{7}$$

где
$$f(\phi;k)=c(k)\phi^k$$
$$f(\phi;\le k)=\sum_{m\in\Omega,\, m\le k} c(m)\phi^m$$
$$f(\phi;\ge k)=\sum_{m\in\Omega,\, m\ge k} c(m)\phi^m$$
$$f(\phi)=\sum_{m\in\Omega} c(m)\phi^m$$

Таким образом, при известном законе распределения случайной величины κ мы может точно вычислить вероятность соответствующих левых частей равенств (7), которую можно использовать для нахождения p-значения.

4. Статистические методы анализа

Методы статистического анализа дискретных данных, в частности таблиц сопряженности 2×2, использующие вероятности попадания в хвост распределения (иначе говоря, вычисление p-значения), можно разделить на точные и асимптотические методы. Точные методы анализа используют конкретное распределение, которое порождает имеющиеся данные, а асимптотические обращаются к общим критериям, которые являются предельными распределениями для случайных величин, построенных на основе точных распределений. При этом асимптотические критерии не учитывают выборочные схемы получения данных и потому рассматриваются как универсальные. Однако для малых выборок они могут давать довольно заметные ошибки в величине p-значения или размере доверительного интервала или фактической вероятности попадания в него. Кроме того, существуют методы, занимающие промежуточное положение между точными и асимптотическими, которые основаны на вычислении скорректированных тем или иным способом выражениях для p-значений.

Статистический анализ на основе вычисления вероятности попадания в хвост распределения можно проиллюстрировать на следующем примере. Пусть имеется выборка (x_1, x_2,... x_n) из нормального распределения с неизвестным средним μ и неизвестной дисперсией σ^2. Тогда для проверки гипотезы H_0: $\mu = \mu_0$ с помощью t-критерия Стьюдента надо ввести случайную величину ($\bar{x}$ – выборочное среднее, s^2 – исправленная выборочная дисперсия)

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \tag{8}$$

которая имеет распределение Стьюдента с n – 1 степенью свободы. Вывод о том, согласуется ли гипотеза H_0 с имеющейся выборкой, определяется по тому, насколько велико или мало значение статистики t. Считая, что рассматривается двусторонний вариант (т. е. альтернативная гипотеза H_1: $\mu \ne \mu_0$), вычисляется значение (8) и его абсолютная величина сравнивается с «критическим значением» $t_{кр}$ распределения Стьюдента с n – 1 степенями свободы, которое удовлетворяет условию

$$\mathrm{P}(|t| > t_{кр}) = \alpha,$$

где α – заданный уровень значимости (скажем, 0,05). Если вычисленное значение (8) меньше, чем $t_{кр}$, то гипотеза H_0 не отвергается, а если больше – отвергается на заданном уровне значимости α.

Другой подход состоит в том, что по данной выборке вычисляется *p*-значение, на основании величины которого делается заключение о принятии/непринятии гипотезы H_0. А именно, вычисляется вероятность (здесь *t* – распределение Стьюдента с *n* – 1 степенями свободы)

$$p = P\left(|t| > \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right|\right). \tag{9}$$

Если эта вероятность мала (например, меньше 0,05), то это является свидетельством против гипотезы H_0, если же $p > 0{,}05$, то гипотезу H_0 принимают.

Таким образом, гипотеза H_0 принимается или не принимается в зависимости от того, выполняется или не выполняется неравенство $P(|t| > |t_{\text{набл}}|) > \alpha$, $t_{\text{набл}} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

Введем индикаторную переменную δ скобкой Айверсона $\delta = [P(|t| > |t_{\text{набл}}|) > \alpha]$ (Iverson's bracket: если P – некоторое высказывание, то скобка Айверсона от этого высказывания определяется условиями: [P] = 1, если P – истинно, и [P] = 0, если P ложно). Тогда переменная δ определяет нерандомизированный критерий для проверки гипотезы H_0: $\delta = [p > \alpha]$, где *p*-значение *p* вычисляется по формуле (9). Иначе говоря, гипотеза H_0 принимается тогда и только тогда, когда $\delta = 1$. Равносильно можно говорить, что принимается гипотеза H_a, где $\delta = [p \le \alpha] = a$.

Если для вычисления *p*-значения используется некоторая случайная величина, которая (а) строится по выборке; (б) вероятностное распределение которой будет точно известно при выполнении гипотезы H_0, то ее можно использовать и в качестве обычного статистического критерия для проверки гипотезы H_0, и в качестве средства для вычисления *p*-значения и задания нерандомизированного критерия δ с помощью скобки Айверсона. В этом случае *p*-значение является вероятностью попадания точно в один (при односторонней альтернативе) или в один из двух хвостов (при двусторонней альтернативе) этой случайной величины. Однако при проверке статистических гипотез для дискретных случайных величин (в частности, для таблиц 2×2) часто оказывается целесообразным несколько ослабить это правило вычисления *p*-значения, что делает необходимым использование *p*-значения как самостоятельного средства проверки статистических гипотез.

Степенное представление целочисленной случайной величины позволяет удобно оперировать с ключевыми конструкциями математического аппарата проверки статистических гипотез – уровнем значимости (размером критерия) и мощностью. Кроме того, с проверяемой статистической гипотезой (односторонней или двусторонней) о значении параметра ϕ очевидно связана задача нахождения области, в которой этот параметр находится с заданной вероятностью (доверительная область), которую также можно найти с помощью функции *p*-значения. Вначале рассмотрим случай односторонней гипотезы.

4.1. Односторонняя гипотеза. Уровень значимости и мощность критерия

Пусть κ – целочисленная случайная величина, зависящая от параметра ϕ, для которой выдвинута гипотеза H_0: $\phi \le \phi_0$ vs H_1: $\phi > \phi_0$. Возможность вычисления *p*-значения опирается на следующее свойство случайной величины κ.

Теорема 3. Если κ – целочисленная случайная величина, зависящая от параметра ϕ и определяемая степенным рядом с носителем Ω, и $k \in \Omega$ отлично от inf Ω и sup Ω, то функция $P(\kappa \ge k; \phi)$ монотонно возрастает по ϕ, а $P(\kappa \le k; \phi)$ – монотонно убывает по ϕ.

Доказательство получается элементарным вычислением производной от каждой вероятности по ϕ с использованием формул (7) (см. подробности в [36, с. 33, 34]).

Таким образом, при фиксированном значении $k \in \Omega$ правосторонняя область $\kappa \ge k$ становится тем больше, чем больше значение параметра ϕ. Соответственно, левостороння область $\kappa \le k$ становится тем больше, чем меньше значение ϕ. Поэтому разумно рассматривать вероятности хвостов $P(\kappa \ge k; \phi)$ и $P(\kappa \le k; \phi)$ как меру справедливости соответствующих односторонних гипотез. А именно, если в результате наблюдения (измерения) над случайной величиной κ, зависящей от параметра ϕ, получено некоторое значение k, и для некоторого значения $\phi = \phi_0$ вероятность попадания значений κ в область $\kappa \ge k$ мала (например, меньше 0,05), можно сделать заключение, что на самом деле значение параметра ϕ должно быть больше, чем ϕ_0 (чтобы в соответствии с Теоремой 3 вероятность попадания в область $\kappa \ge k$ стала больше).

Определение 1. Для заданного значения $\kappa = k$ функция от параметра ϕ

$$p(k, \phi) = P(\kappa \ge k; \phi) \tag{10}$$

называется *функцией p-значения* для односторонних гипотез H_0: $\phi \le \phi_0$ vs H_1: $\phi > \phi_0$. Соответственно, функция от параметра ϕ

$$p(k, \phi) = P(\kappa \le k; \phi) \tag{11}$$

называется *функцией p-значения* для односторонних гипотез H_0: $\phi \ge \phi_0$ vs H_1: $\phi < \phi_0$.

Замечание 1. Выражения (10), (11) известны также под названиями доказательная функция, доверительная кривая, функция доверительного интервала, конфлюэнтная функция.

Пример 7. Пусть в описанных в примере 1 условиях испытание проводилось на 10 пациентах, и у двоих обнаружены побочные явления. Тогда для оценки того, насколько это свидетельствует в пользу наличия побочных эффектов исследуемого лекарства, разумно вычислить вероятность того, что такое или еще большее число испытуемых будет иметь побочные явления от приема этого лекарства. Иначе говоря, необходимо вычислить вероятность $P(\kappa \ge 2; \phi)$, которая, однако, остается зависящей от неизвестного параметра ϕ. Рассмотрим зависимость этой вероятности от ϕ. Из равенств (7) следует

$$P(\kappa \ge 2; \phi) = 1 - P(\kappa < 2; \phi) = 1 - (P(\kappa = 0; \phi) + P(\kappa = 1; \phi)) = 1 - \frac{1 + 10\phi}{(1+\phi)^{10}},$$

т. е. функция p-значения здесь равна $p(2, \phi) = 1 - \frac{1 + 10\phi}{(1+\phi)^{10}}$. Так как шанс ϕ является монотонно возрастающей функцией от вероятности π, то вместо гипотезы H_0: $\pi \le \pi_0$ против конкурирующей гипотезы H_1: $\pi > \pi_0$ можно рассматривать соответствующую пару гипотез относительно параметра ϕ: H_0: $\phi \le \phi_0$ vs H_1: $\phi > \phi_0$. Очевидно, в данном примере нас интересует гипотеза малого значения вероятности π. Поэтому, выбрав некоторое пороговое значение ϕ_0, мы должны оценить, каково может быть максимальное значение $P(\kappa \ge 2; \phi)$ при справедливости гипотезы H_0, т. е. при выполнении неравенства $\phi \le \phi_0$.

Таким образом, если мы хотим проверить гипотезу H_0: $\phi \le \phi_0$ при условии $\mathcal{K} = k$, то одностороннее точное *p*-значение для проверки этой гипотезы можно определить как (индекс *c* от *conventional p-value*)

$$p_c(k;\phi_0) = \sup_{\phi \le \phi_0} \mathrm{P}(\mathcal{K} \ge k;\phi) = \frac{f(\phi_0;\ge k)}{f(\phi_0)}, \tag{12}$$

т. е. $p_c(k;\phi_0) = \sup\limits_{\phi \le \phi_0} p(k;\phi)$. Аналогично для проверки односторонней гипотезы H_0: $\phi \ge \phi_0$ при условии $\mathcal{K} = k$ одностороннее точное *p*-значение можно принять равным

$$p_c(k;\phi_0) = \sup_{\phi \le \phi_0} p(k;\phi) = \sup_{\phi \le \phi_0} \mathrm{P}(\mathcal{K} \le k;\phi) = \frac{f(\phi_0;\le k)}{f(\phi_0)}. \tag{13}$$

Определяемые равенствами (12), (13) *p*-значения обладают теми свойствами, которые ожидаются от этих величин: малое *p*-значение означает маловероятную справедливость соответствующей гипотезы. Большое *p*-значение (обычно больше принятого порогового значения, например, 0,05) свидетельствует о согласовании выборочных данных и выдвинутой гипотезы на заданном уровне порогового значения. Заметим также, что из теоремы 3 следует, что для гипотезы H_0: $\phi \le \phi_0$ *p*-значение (12) является возрастающей функцией параметра ϕ_0, а для гипотезы H_0: $\phi \ge \phi_0$ *p*-значение (13) является убывающей функцией этого параметра. Эти свойства выполняются также и для вводимых ниже других способов вычисления *p*-значения.

Если мы проверяем гипотезу, что H_0: $\pi \le 0{,}1$, т. е. H_0: $\phi \le 0{,}111... = 1/9$, то соответствующее точное *p*-значение (12) будет равно

$$\mathrm{P}(\mathcal{K} \ge 2;\ 1/9) = 0{,}264.$$

Следовательно, на основании этого *p*-значения можно сделать вывод, что гипотеза H_0 согласуется с имеющимися данными.

Следующее определение вводит другой способ вычисления *p*-значения [38, 44, 45].

Определение 2. Если $\mathcal{K}$ – дискретная случайная величина, зависящая от параметра ϕ, то *односторонним средним p-значением* (*one-sided mid-p-value*) для проверки гипотезы H_0: $\phi \le \phi_0$ против альтернативы H_1: $\phi > \phi_0$ называется функция от ϕ_0

$$p_m(k;\phi_0) = p_c(k;\phi_0) - 0{,}5\frac{f(\phi_0;k)}{f(\phi_0)} = \frac{f(\phi_0;\ge k) - 0{,}5 f(\phi_0;k)}{f(\phi_0)}, \tag{14}$$

где $p_c(t;\phi_0)$ – точное *p*-значение (12).

Эта величина также обладает теми же свойствами, которые должны быть у *p*-значения, в частности, она монотонна по ϕ (см. теорему 3).

Аналогичная формула для левосторонней области очевидна:

$$p_m(k;\phi) = \mathrm{P}(\mathcal{K} \le k;\phi) - 0{,}5(\mathcal{K} = k;\phi); \quad p_m(k;\phi_0) = \frac{f(\phi_0;\le k) - 0{,}5 f(\phi_0;k)}{f(\phi_0)}. \tag{15}$$

Таким образом, вероятность значения $\mathcal{K} = k$ распределяется поровну между правой и левой областями (хвостами распределения). Очевидно, что одностероннее среднее p-значение (14), (15) всегда меньше, чем точное p-значение (12), (13) соответственно.

Пример 7 (продолжение). Одностороннее среднее p-значение (14) для проверки гипотезы H_0: $\phi \le \phi_0$ равно 0,264 – 0,194/2 = 0,167.

Еще один способ введения p-значения основан на часто применяемой нормальной аппроксимации дискретных распределений. При большом числе проведенных испытаний односторонняя вероятность $P(\kappa \le t;\ \phi)$, где t – произвольное число, обычно достаточно хорошо аппроксимируется функцией распределения нормальной (гауссовой) случайной величины с подходящими параметрами. Например, для биномиального распределения κ это следует из интегральной теоремы Муавра – Лапласа [22, 23]

$$P(\kappa \le t;\ \pi) = \frac{1}{2} + \Phi\left(\frac{t - n\pi}{\sqrt{n\pi(1-\pi)}}\right),$$

где $\Phi(x)$ – функция Лапласа, определяемая равенством $\Phi(x) = \dfrac{1}{\sqrt{2\pi}}\int\limits_0^x e^{-t^2/2}dt.$

Следовательно, асимптотическое одностороннее p-значение $p_a(t;\ \pi)$ для биномиального распределения можно определить равенствами

$$\begin{aligned} p_a(t;\ \pi) &= \tfrac{1}{2} + \Phi(z); \\ p_a(t;\ \pi) &= \tfrac{1}{2} - \Phi(z); \end{aligned} \qquad (16)$$

для левосторонней области и правосторонней области соответственно, где $z = \dfrac{t - n\pi}{\sqrt{n\pi(1-\pi)}}$. Аналогично будут определяться асимптотические односторонние p-значения для других дискретных распределений.

Пример 7 (продолжение). Здесь $n = 10$, $\pi = 0{,}1$, $t = 2$. Значит, асимптотическое правостороннее p-значение равно $p_a(2;\ 0{,}1) = \frac{1}{2} - \Phi(1{,}05) = 0{,}5 - 0{,}3531 = 0{,}1469$.

Замечание 1. Выше были введены три варианта односторонних p-значений: обычный, основанный на вероятности хвоста распределения, среднее p-значение и асимптотическое p-значение. Причем последние два, строго говоря, не являются вероятностями хвостов распределения данных. Однако все они обладают теми желательными свойствами, которые позволяют делать выводы о справедливости соответствующей односторонней гипотезы. Для непрерывной случайной величины p-значение, как правило, является вероятностью некоторого события в изучаемом выборочном пространстве. Для дискретной случайной величины это может быть не так. В целом при использовании того или иного варианта p-значения для дискретной случайной величины необходимо иметь в виду следующие особенности:

– p-значение есть мера правдоподобия данной гипотезы, принимающая значение от 0 до 1. Оно является случайной величиной на выборочном пространстве;
– p-значение выводится на основе некоторых вероятностных рассуждений, но не обязательно само является вероятностью какого-либо события;
– в частности, p-значение не всегда равно вероятности того, что нулевая гипотеза ложна и не тождественно уровню значимости или величине ошибки первого рода.

Напомним некоторые известные понятия, представленные в удобном для дальнейшего применения виде. Пусть выдвинута некоторая статистическая гипотеза H_0 о значениях параметра ϕ, от которого зависит случайная величина $\mathcal{K}$ с носителем Ω, и задан уровень значимости α_0. Нерандомизированный критерий δ определяется равенством $\delta(k;\, \alpha_0) = [p(k) \le \alpha_0]$, где $p(k)$ – p-значение, определяемое тем или иным образом.

Определение 3. *Критической областью* $R(\alpha_0) = R_0$ статистического критерия δ, соответствующей уровню значимости α_0, называется множество значений $t \in \Omega$, при которых гипотеза H_0 отвергается:

$$R(\alpha_0) = \{k \in \Omega \,\big|\, p(k;\, \alpha_0) \le \alpha_0\} = \{k \in \Omega \,\big|\, \delta(k;\, \alpha_0) = 1\}.$$

Определение 4. *Максимальным достигаемым* (фактическим) уровнем значимости критерия δ, или *максимальным фактическим размером* критерия δ, или максимальным фактическим уровнем ошибки I-го рода, называется

$$\alpha = \sup_{H_0} \sum_{k\in\Omega} \left[p(k) \le \alpha_0\right] \mathrm{P}(\mathcal{K} = k;\, \phi), \tag{17}$$

где точная верхняя грань вычисляется по области значений параметра ϕ, описываемой гипотезой H_0. Например, для гипотезы H_0: $\phi \le \phi_0$ имеем

$$\alpha = \sup_{\phi\le\phi_0} \sum_{k\in\Omega} \left[p(k) \le \alpha_0\right] \mathrm{P}(\mathcal{K} = k;\, \phi).$$

Иначе величину (17) можно описать как максимальную (по ϕ, удовлетворяющем H_0) вероятность отвергнуть H_0, когда она верна.

Важной характеристикой статистических тестов является их устойчивость к заданному уровню значимости, т. е. к заданному размеру критерия.

Определение 5. Статистический тест δ называется консервативным, если $\alpha \le \alpha_0$, и *либеральным* в противоположном случае.

Иначе говоря, консервативный критерий сохраняет заданный размер (не увеличивает его), а либеральный – нарушает (увеличивает).

Пример 7 (продолжение). Рассмотрим фактический уровень значимости для каждого рассмотренного выше способа вычисления p-значения. Примем в качестве номинального уровня значимости обычное значение $\alpha_0 = 0{,}05$. Для рассмотренных выше способов вычисления p-значения получаем, что уровень значимости вычисляется по формуле ($u = c,\, m,\, a$ соответственно, для обычного p-значения (12), среднего p-значения (14) и асимптотического p-значения (16))

$$\alpha_u(\phi) = \sum_{k\in\Omega} \left[p_u(k;\, \phi) \le \alpha_0\right] \mathrm{P}(\mathcal{K} = k;\, \phi)$$

и в зависимости от конкретного способа вычисления имеет график, показанный на рис. 1.

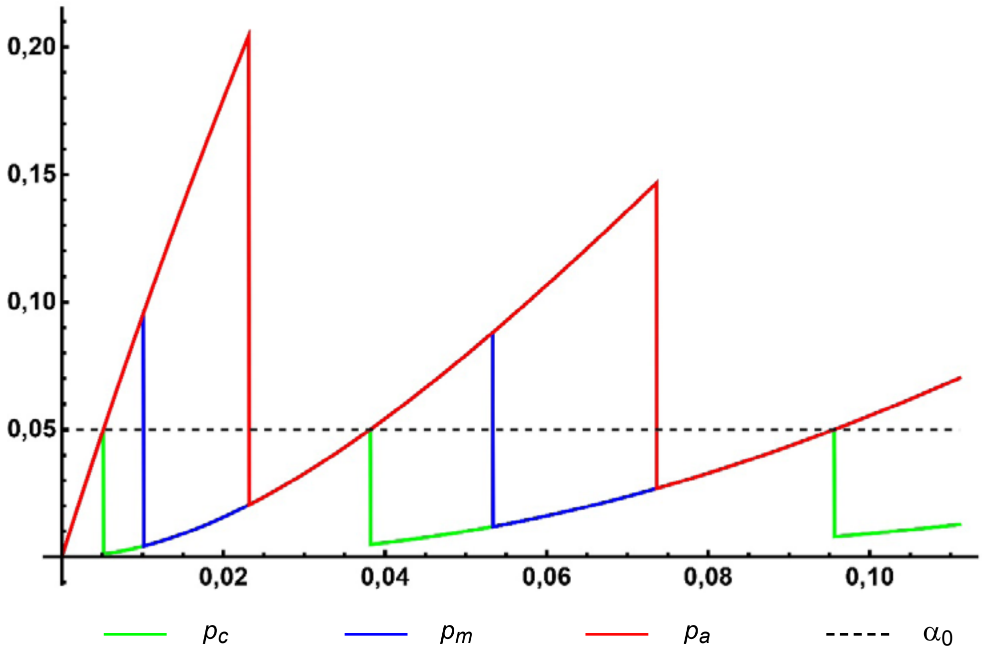


Рис. 1. Уровень значимости критерия, определяемого *p*-значением как функция от параметра ϕ. Зеленая линия – *p*-значение вычисляется по формуле (12); синяя линия – среднее *p*-значение (14); красная линия – асимптотическое *p*-значение (16). Части зеленой и синей линий, находящиеся под красной линией, совпадают с ней. Пунктирная линия показывает номинальный выбранный уровень значимости α_0 = 0,05

При этом, как нетрудно проверить, выполнятся неравенство $\sup\limits_{\phi\le\phi_0}\alpha_c(\phi) = 0{,}05,\ \phi_0 = 1/9$. Следовательно, тест δ, основанный на вычислении *p*-значения по формуле (12), является консервативным. Напротив, остальные два *p*-значения определяют либеральные статистические критерии: $\sup\limits_{\phi\le\phi_0}\alpha_m(\phi) = 0{,}096,\ \phi_0 = 1/9$ и это значение достигается при ϕ = 0,01; $\sup\limits_{\phi\le\phi_0}\alpha_a(\phi) = 0{,}2,\ \phi_0 = 1/9$ достигается при ϕ = 0,02.

Таким образом, вычисление *p*-значения по формуле (12) приводит к консервативному статистическому критерию, а по формулам (14) и (16) – к либеральному.

Другая важная характеристика статистического критерия – мощность, т. е. вероятность отклонить нулевую гипотезу, когда она неверна. Имея в виду, что обычно проверяемая гипотеза формулируется как утверждение об отсутствии эффекта (в том или ином смысле – см. [46, 47]), неформально можно сказать, что мощность критерия показывает его способность обнаружить *имеющийся* эффект. В этом контексте уровень значимости показывает возможность критерия (ложно) обнаружить отсутствующий эффект, т. е. сделать вывод из имеющихся данных о наличии эффекта, которого на самом деле нет. Уровень значимости (α) и мощность критерия (*pwr*) можно определить равенствами (эти величины определяются для конкретной нулевой гипотезы и ее альтернативы, а также статистического критерия)

$$\alpha = \mathrm{P}(\overline{\mathrm{H}}_0 \mid \mathrm{H}_0),\ pwr = 1 - \mathrm{P}(\mathrm{H}_0 \mid \mathrm{H}_1)$$

или

$$pwr = \mathrm{P}(\text{reject } \mathrm{H}_0 \mid \mathrm{H}_1 \text{ is true}). \tag{18}$$

Очевидно, что для простых основной и конкурирующей гипотез вычисление мощности не представляет труда, но для сложных гипотез это сделать затруднительно, так как такие гипотезы не определяют однозначно соответствующие вероятностные распределения. Тем не менее в большинстве случаев можно построить функцию двух переменных, которая дает представление о поведении мощности критерия в зависимости от конкретных значений основной и конкурирующей гипотез.

Пусть $\mathcal{K}$ – целочисленная случайная величина с носителем Ω и H_0: $\phi \le \phi_0$, H_1: $\phi > \phi_0$. Рассмотрим для каждого $\phi_1 \le \phi_0$ критическую область $R_0(\phi_1) = \{k \in \Omega \mid p(k;\, \phi_1) \le \alpha_0\}$, где α_0 – некоторый назначенный уровень значимости (например, 0,05), $p(k;\, \phi)$ – p-значение, вычисляемое тем или иным образом. Тогда функция мощности (18) принимает вид

$$pwr(\phi_1, \phi_2) = \sum_{k \in R_0(\phi_1)} \mathrm{P}(\mathcal{K} = k;\, \phi_2), \quad \phi_2 > \phi_0 \tag{19}$$

Для первоначальной оценки мощности критерия вместо этой функции можно рассматривать функцию $pwr(\phi_2)$ одной переменной ϕ_2, которая является нижней гранью полной функции (19) по всем значениям ϕ_1, допускаемым гипотезой H_0, с фиксированным значением ϕ_2. Для такой функции выполняется неравенство $pwr(\phi_1, \phi_2) \ge \inf_{\phi_1 | H_0} pwr(\phi_1, \phi_2) = pwr(\phi_2)$, где $\phi_1 \middle| H_0$ означает, что ϕ_1 удовлетворяет гипотезе H_0. Вместо нахождения этой функции можно воспользоваться тем, что часто существует такое значение $\phi_1^0 \middle| H_0$, для которого критическое множество $R_0(\phi_1^0)$ будет минимальным среди всех $\phi_1 | H_0$. Тогда, очевидно, для любого $\phi_1 | H_0$ выполняется неравенство $\sum_{k \in R_0(\phi_1)} \mathrm{P}(\mathcal{K} = k \mid \phi_2) \ge \sum_{k \in R_0(\phi_1^0)} \mathrm{P}(\mathcal{K} = k \mid \phi_2)$. Например, для рассматриваемого случая гипотез H_0: $\phi \le \phi_0$ vs H_1: $\phi > \phi_0$ таким минимальным множеством будет $R_0(\phi_0)$ и, следовательно, функция

$$pwr(\phi_2) = \sum_{k \in R_0(\phi_0)} \mathrm{P}(\mathcal{K} = k \mid \phi_2) \tag{20}$$

является оценкой снизу для полной функции мощности (19).

Пример 7 (продолжение). Для обычного p-значения, вычисляемого по формуле (12) и $\pi = 0{,}1$, т. е. $\phi_0 = 1/9$, критическая область равна

$$R_0(\phi_0) = \{4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\},$$

а при меньших значениях ϕ к нему добавляются значения 3, 2, 1. Поэтому вычисленная по формуле (20) оценка снизу функции (19) мощности критерия имеет вид

$$pwr(\phi) = \sum_{k=4}^{10} P(\mathcal{K} = k;\, \phi), \tag{21}$$

где значения $\mathrm{P}(\mathcal{K} = k;\, \phi)$ вычисляются по формуле (7). На рис. 2 показан график этой функции на отрезке [1/9, 2].

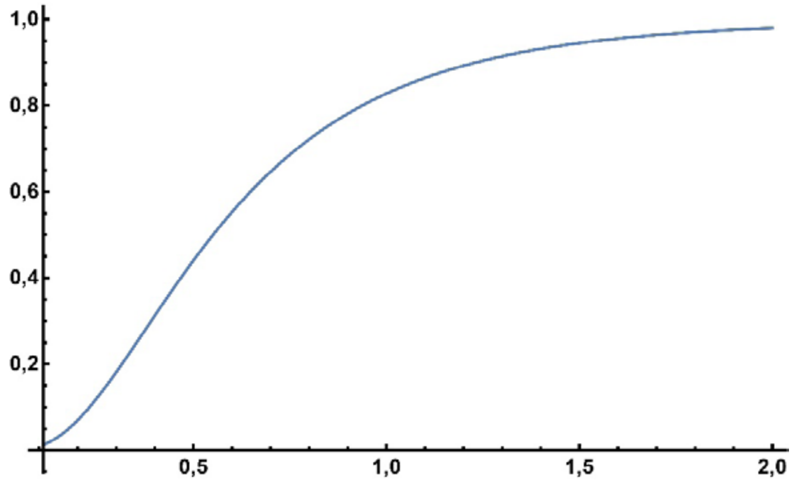


Рис. 2. График функции мощности (21) статистического критерия для примера 1 с консервативным тестом

Разумеется, для конкретных гипотез H_0: $\phi = \phi_1$ vs H_1: $\phi = \phi_2$, $\phi_1 < \phi_0 < \phi_2$ функция мощности может значительно отличаться от ее нижней оценки (20). Например, на рис. 3 показаны графики функции мощности $pwr(\phi_1, \phi)$, $1/9 \le \phi = \phi_2 \le 1$, полученные по формуле (21) при значениях $\phi_1 = 0{,}01$ (красная кривая), $\phi_1 = 0{,}05$ (синяя), $\phi_1 = 0{,}1$ (зеленая кривая). Пунктиром показано значение мощности 0,8.

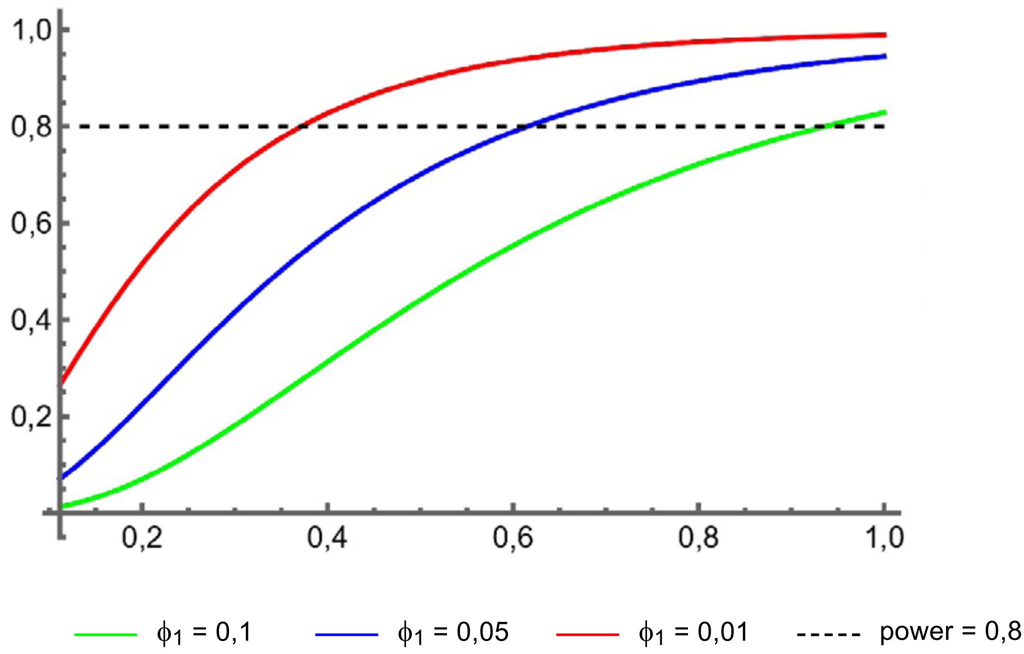
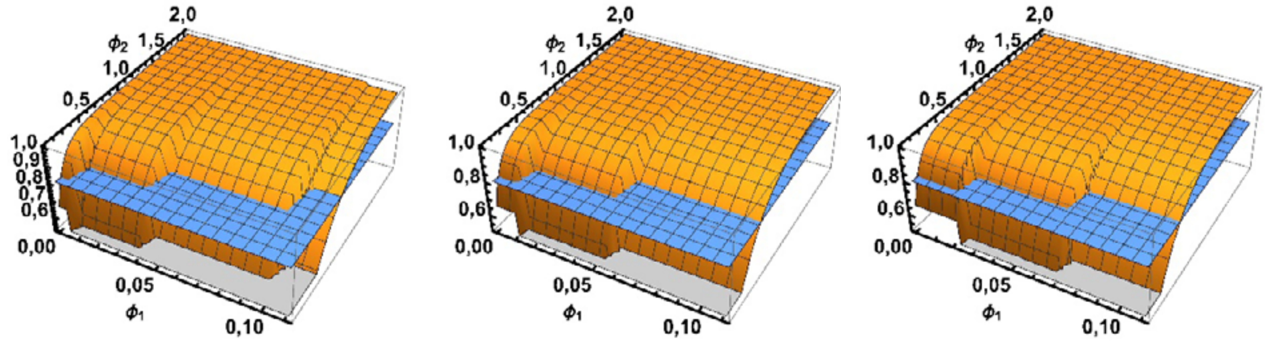


Рис. 3. Сечения ϕ_1 = const функции мощности (19) для примера 1: $\phi_1 = 0{,}01$ (красная кривая); $\phi_1 = 0{,}05$ (синяя кривая); $\phi_1 = 0{,}1$ (зеленая кривая)

Как видно из рис. 3, чем больше различаются значения ϕ_1 и $\phi_0 = 1/9$, тем сильнее различаются графики мощности (19) и (20). При $\phi_1 = 0{,}1$ функция мощности (19) практически совпадает с функцией (21).

Представление о различии функций мощности (19) для различных статистических тестов на основе вычисления *p*-значения по формулам (12), (14), (16) можно увидеть на рис. 4 (для примера 7).



(а) (б) (в)

Рис. 4. Поверхность мощности для примера 7 с (а) консервативным тестом; (б) либеральным тестом на основе среднего *p*-значения; (в) либеральным тестом на основе асимптотического *p*-значения. Голубым цветом показан уровень мощности 0,8

Из рис. 1 и 4 видно, что статистический критерий на основе среднего *p*-значения занимает промежуточное положение между точным и асимптотическим критериями не только по величине достигаемого уровня значимости, но и по мощности. В этом смысле его можно считать приемлемым компромиссом в тех случаях, когда точный критерий на основе (12) слишком консервативен, а на основе (16) – слишком либерален.

В целом можно отметить, что вычисление фактического размера критерия и его мощности становится все более распространенным в медико-биологической практике. Разработка новых алгоритмов и теоретических методов такого вычисления делает эти задачи значительно более доступными даже для пользователей, не имеющих глубоких знаний в математической статистике. При этом, несмотря на наличие общей теории и доступность программных средств вычисления характеристик статистических критериев, корректное применение этих алгоритмов, методов и программ предполагает точное понимание всех необходимых шагов таких вычислений и учет специфики конкретной ситуации.

4.2. Односторонние доверительные области

Важной частью анализа целочисленных случайных величин и, в частности, таблиц сопряженности 2×2, является нахождение доверительных интервалов для неизвестных параметров этих распределений. Проверки гипотезы H_0: $\phi = \phi_0$ о точном значении такого параметра или гипотез H_0: $\phi \le \phi_0$, H_0: $\phi \ge \phi_0$ о его односторонней границе часто недостаточно для оценки истинного значения этого параметра. Например, пусть сравнивается эффективность нового метода лечения некоторого заболевания и ω – шанс выздоровления после применения этого метода лечения. Проверяемая гипотеза может предполагать отсутствие эффективности этого метода, т. е. H_0: $\phi \le 1{,}0$ против гипотезы H_1: $\phi > 1{,}0$, где ϕ – шанс выздоровления. Предполагая биномиальную модель описания выздоровления, мы получим, что при $n = 100$ испытуемых и $k = 55$ выздоровевших *p*-значение (12) равно 0,184, а при $n = 1\,000$ испытуемых и $k = 550$ выздоровевших это же *p*-значение равно 0,00009. При этом в обоих случаях частота выздоровления одна и та же 0,55. Таким образом, в зависимости от объема исследований мы можем сделать противоположные выводы: для $n = 100$ испытуемых и $k = 55$ выздоровевших можно считать, что лечение неэффективно, а при $n = 1\,000$ испытуемых и $k = 550$ выздоровевших, что оно значимо эффективно.

Понятно, что эти противоположные выводы возникают вследствие значительного различия числа испытуемых в двух опытах. Однако эти вычисления *p*-значения не дают возможности оценить, *насколько* эффективно лечение. Как таковое *p*-значение показывает, насколько вероятно наблюдаемое значение отклика в предположении выполнения гипотезы H_0, но никак не описывает насколько велик и *предметно значим* сам эффект, приводящий к такому значению отклика. Иначе говоря, кроме величины *p*-значения, необходимо иметь и оценку величины интересующего исследователя эффекта, например, вероятности выздоровления из примера выше. Одним из таких методов является построение доверительных интервалов для интересующей нас величины.

Для односторонних гипотез соответствующие доверительные интервалы будут ограничены только с одной стороны, т. е. будут лучами на числовой прямой. Рассмотренный ниже метод построения таких интервалов представляет собой обращение критерия проверки гипотезы H_0 [46, 47] и удобен для рассмотрения именно дискретных случайных величин.

Пусть рассматривается гипотеза H_0: $\phi \le \phi_0$ (H_0: $\phi \ge \phi_0$), где ϕ – параметр, от которого зависит целочисленная случайная величина κ. В результате наблюдения получено значение $\kappa = k$. Используя эту информацию, требуется найти такую область изменения параметра ϕ, в которой с заданной вероятностью $(1 - \alpha_0)$, $0 < \alpha_0 < 1$, находится значение параметра ϕ_0.

Определение 6. *Правосторонней* (*левосторонней*) *доверительной областью* уровня $(1 - \alpha_0)$, $0 \le \alpha_0 \le 1$, для параметра ϕ_0 называется множество

$$\{\phi_0 \mid H_0\colon \phi \le \phi_0 \text{ не отвергается на уровне значимости } \alpha_0\}, \tag{22}$$

соответственно

$$\{\phi_0 \mid H_0\colon \phi \ge \phi_0 \text{ не отвергается на уровне значимости } \alpha_0\}. \tag{23}$$

Поскольку во всех рассматриваемых определениях *p*-значения они будут монотонны по параметру ϕ (для разных гипотез монотонность будет разной – см. теорему 3), то множество (22) ограничено снизу, а множество (23) – ограничено сверху. Минимальное (максимальное) значение множества (22) (соответственно, (23)) называется *левой*, или *нижней* (*правой*, или *верхней*) доверительной границей параметра ϕ_0.

Так как принятие или непринятие гипотезы H_0 производится на основе величины *p*-значения, то такие доверительные области будут различными в зависимости от способа вычисления *p*-значения. Например, для обычного выражения (12) для *p*-значения гипотеза H_0: $\phi \le \phi_0$ будет приниматься при выполнении неравенства

$$\frac{f(\phi_0; \ge k)}{f(\phi_0)} \ge \alpha_0.$$

При этом так как левая часть неравенства является монотонно возрастающей непрерывной функцией от ϕ_0, изменяющейся от 0 до 1, то существует значение ϕ_0, для которого выполняется равенство $f(\phi_0; \ge k) = \alpha_0 f(\phi_0)$, которое и является левой границей доверительной области (22). Таким образом, левая (правая) граница доверительной области (22) (соответственно, (23)) уровня $(1 - \alpha_0)$, $0 < \alpha_0 < 1$, является решением уравнения

$$\sum_{m\geq k} c(m)\phi^m = \alpha_0 \sum_{m\in\Omega} c(m)\phi^m \tag{24}$$

соответственно,

$$\sum_{m\leq k} c(m)\phi^m = \alpha_0 \sum_{m\in\Omega} c(m)\phi^m. \tag{25}$$

Замечание 3. Множества (22), (23) можно представить в виде $\{\phi_0 \mid p(k;\ \phi_0) \geq \alpha_0\}$, где $p(k;\ \phi_0)$ – соответствующее p-значение. Следовательно, доверительная область есть множество $p^{-1}(k;\ [\alpha_0;\ 1])$ (прообраз отрезка по второй переменной). Если рассматривать обычное определение p-значения (12), (13), то по теореме 3 функция $p(k;\ \phi_0)$ монотонна по ϕ_0, что гарантирует, что доверительная область связна, т. е. представляет собой интервал (бесконечный с одной стороны).

Геометрически граница доверительной области получается пересечением графика функции $p(k;\ \phi)$ и горизонтальной линии α_0.

Пример 7 (продолжение). Для гипотезы H_0: $\phi \leq \phi_0$, $\alpha_0 = 0{,}05$ получаем 95 %-ю доверительную область для параметра ϕ_0, решая уравнение (24) или находя точку пересечения графика функции $p(2;\ \phi)$, определяемой равенством (12), с линией $\alpha_0 = 0{,}05$. Аналогично для гипотезы H_0: $\phi|\geq|\phi_0$, 95 %-я доверительная область для параметра ϕ_0 находится решением уравнения (25) или поиском точки пересечения графика функции $p(2;\ \phi)$, определяемой равенством (13), с линией $\alpha_0 = 0{,}05$ (см. рис. 5).

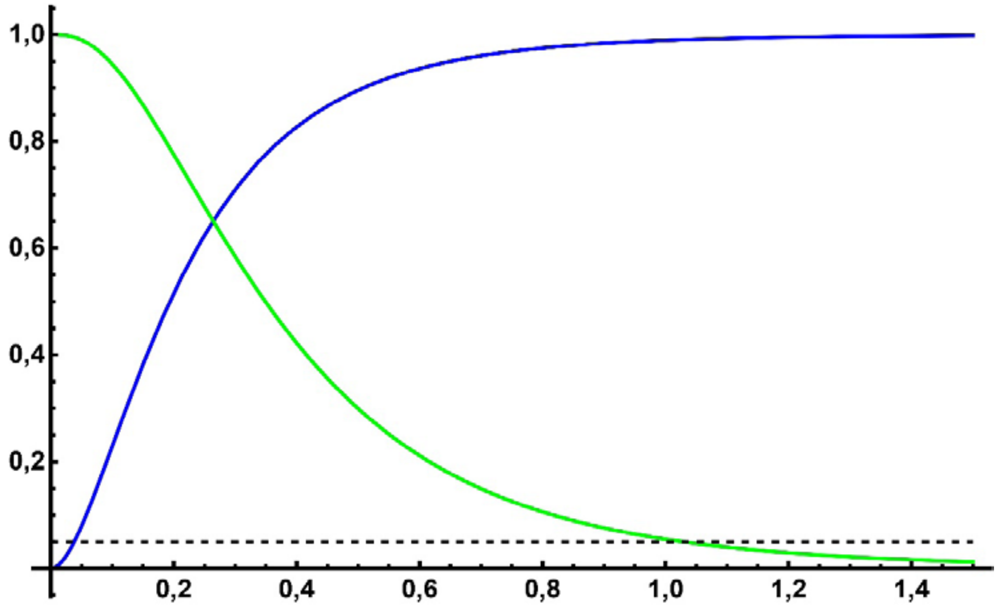


Рис. 5. График $p(2;\ \phi)$ для примера 7 и его пересечение с прямой $\alpha_0 = 0{,}05$. Синяя линия – p-значение для гипотезы H_0: $\phi \leq \phi_0$; зеленая – для гипотезы H_0: $\phi \geq \phi_0$

Искомое значение для левой границы доверительной области равно 0,038. Следовательно, доверительная область есть луч $[0{,}038;\ \infty)$. Для правой границы 1,028, доверительная область есть отрезок $[0;\ 1{,}028]$. Заметим также, что в отрезке $[0{,}038;\ 1{,}028]$ параметр ϕ_0 находится с вероятностью $0{,}95^2 = 0{,}9025$.

Аналогично находятся границы доверительных областей для других способов вычисления p-значения. А именно, для среднего p-значения (14), (15) границы доверительной области уровня $(1 - \alpha_0)$, $0 < \alpha_0 < 1$, находятся как решения уравнений

$$f(\phi; \geq k) - 0{,}5f(\phi; k) = \alpha_0 f(\phi),$$
$$f(\phi; \leq k) - 0{,}5f(\phi; k) = \alpha_0 f(\phi). \qquad (26)$$

Для асимптотического вычисления p-значения (16) односторонние границы доверительных областей для параметра π находятся как решения уравнения (меньший корень дает левую границу, больший – правую)

$$(t - n\pi)^2 = z_{1-\alpha_0} n\pi(1 - \pi),$$

где $z_{1-\alpha_0}$ – квантиль уровня $1 - \alpha_0$ стандартного нормального распределения. Явное решение этого уравнения имеет вид

$$\pi = 0{,}5(b \pm \sqrt{b^2 - 4ac}), \quad b = \frac{2t + z^2_{1-\alpha_0}}{n + z^2_{1-\alpha_0}}, \quad c = \frac{t^2}{n(n + z^2_{1-\alpha_0})} \qquad (27)$$

Пример 7 (продолжение). Для одностороннего среднего p-значения левая и правая границы доверительных областей, соответственно, равны 0,053 и 0,876. Для асимптотического p-значения эти границы таковы (в отличие от предыдущих случаев здесь это границы для биномиальной вероятности π): 0,086 и 0,399. Для шанса ϕ эти значения равны 0,094 и 0,664 соответственно.

При сопоставлении односторонних областей для различных способов вычисления p-значения необходимо учитывать, что среднее и асимптотическое p-значения либеральны и, следовательно, реальная вероятность того, что оцениваемый параметр находится в определяемой ими области будет меньше заданного уровня доверительной вероятности. Иначе говоря, они дают более оптимистичную доверительную область, чем консервативное p-значение (12), (13).

Замечание 4. Как мы видели, для определения фактического p-значения для данного наблюдаемого значения $\kappa = k$ и для определения доверительной области параметра распределения κ используется одна и та же величина – вычисляемое тем или иным образом p-значение. График зависимости p-значения от параметра ϕ позволяет найти как величину p-значения для проверяемой гипотезы, так и доверительную область для этого параметра. Например, пусть рассматривается гипотеза H_0: $\phi \leq \phi_0$ с уровнем значимости $\alpha_0 = 0{,}05$ и наблюдаемым значением $\kappa = k$ случайной величины κ, зависящей от параметра ϕ. Построим график $p(k; \phi)$ как функции от ϕ. Тогда p-значение для проверки гипотезы H_0 находится как ордината точки пересечения этого графика с прямой $\phi = \phi_0$, а левая граница односторонней доверительной области – как абсцисса точки пересечения этого графика с прямой $\alpha = \alpha_0$ (см. рис. 6).

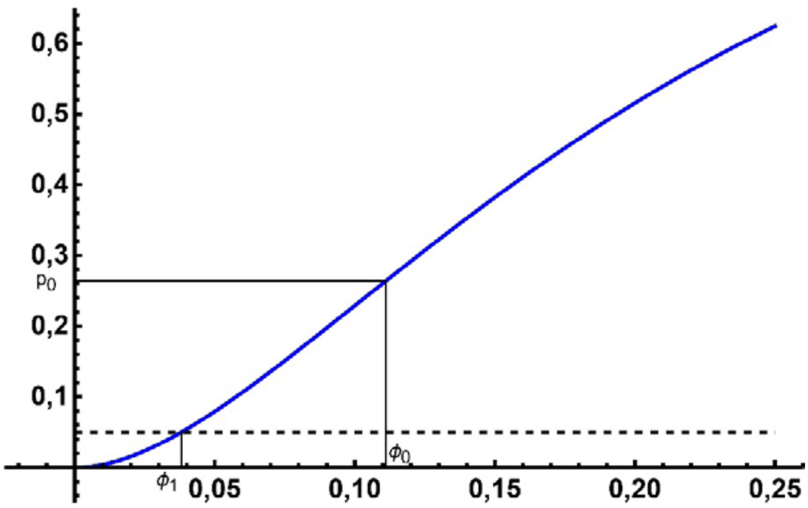


Рис. 6. График p-значения для проверки гипотезы H_0: $\phi \le \phi_0$ при заданном значении $\kappa = k$ случайной величины κ (синяя линия), а также нахождение соответствующего уровня значимости p_0 и правосторонней доверительной области с левой границей ϕ_1. Пунктирная линия отмечает уровень значимости α_0

4.3. Двусторонняя гипотеза

При двустороннем оценивании рассматривается гипотеза H_0: $\phi = \phi_0$ с различными альтернативами, обычно H_1: $\phi \ne \phi_0$. В практическом анализе данных такие гипотезы более естественны, так как не предполагают какую-либо априорную информацию о направленности эффекта. Например, при сравнении двух методов лечения исходная нулевая гипотеза состоит в том, что оба метода имеют одинаковую эффективность. Альтернатива может быть как отрицанием этого предположения, так и выражением большей эффективности какого-то из этих методов.

На основании результатов для одностороннего оценивания можно ожидать, что для проверки справедливости гипотезы H_0 необходимо найти такую функцию $p(k, \phi)$ вычисления p-значения, для которой будут выполняться следующие условия:

– $p(k, \phi)$ непрерывная функция от ϕ, принимающая значения от 0 до 1;

– $p(k, \phi)$ возрастает от 0 при изменении ϕ от 0 до некоторого значения ϕ_1 и убывает до 0, начиная с некоторого значения ϕ_2 (возможно, совпадающего с ϕ_1).

Тогда как малые, так и большие значения ϕ будут иметь малые значения $p(k, \phi)$, что можно считать свидетельством против гипотезы H_0: $\phi = \phi_0$.

В качестве такой функции можно выбрать функцию удвоения вероятности минимального хвоста для данного значения $\kappa = k$ (twice the smaller tail, или double tail method, см., например, [35, 38]), полагая ее равной 1, если удвоенная минимальная вероятность хвоста будет больше 1. Таким образом, положим

$$p(k;\phi) = \min\{2\min\{\mathrm{P}(\kappa \le k;\phi), \mathrm{P}(\kappa \ge k;\phi)\}, 1\}. \tag{28}$$

На рис. 7 показан график этой функции для примера 7.

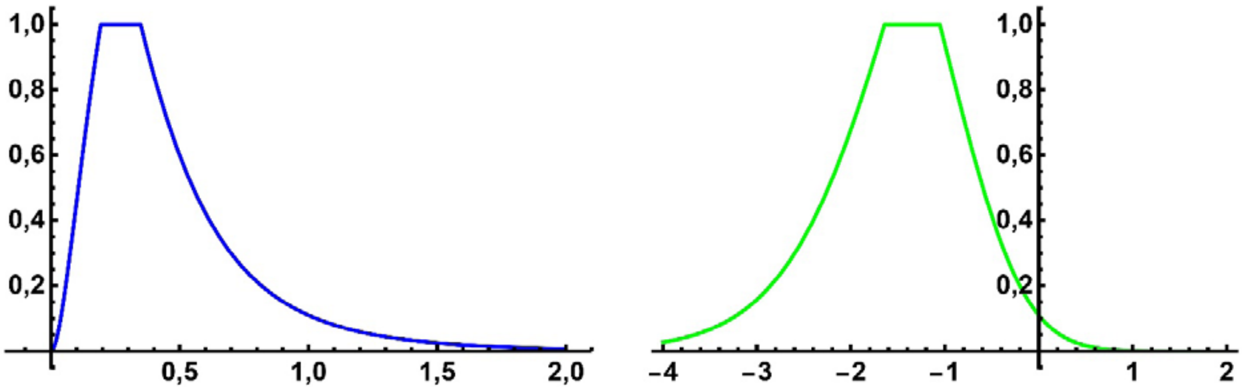


Рис. 7. График функции $p(2, \phi)$ для примера 7 относительно ϕ (слева) и относительно переменной $\beta = \ln \phi$ (справа)

С помощью функции (28) можно следующим образом решить обычные задачи оценивания в их двустороннем варианте [38]:

– Для гипотезы H_0: $\phi = \phi_0$ при $\kappa = k$ двустороннее p-значение определяется как $p(k, \phi_0)$.

– Двусторонний доверительный интервал уровня $(1 - \alpha_0)$, $0 \le \alpha_0 \le 1$, для параметра ϕ_0 есть множество таких значений ϕ, для которых H_0: $\phi = \phi_0$ не отвергается на уровне α_0, т. е. это множество $\{\phi \mid p(k; \phi) \ge \alpha_0\}$, или $p^{-1}(k; [\alpha_0; 1])$, как и для одностороннего случая. Границы этого интервала находятся как решение уравнения $p(k; \phi) = \alpha_0$.

– Значение $\hat{\phi}(k)$ которое содержится во всех двусторонних доверительных интервала при $0 \le \alpha_0 \le 1$, будем называть *точечной оценкой* параметра ϕ. Как следует из рис. 7, в этом подходе такое число может быть неединственным.

Пример 7 (продолжение). Вычислим двустороннее p-значение для гипотезы H_0: $\phi = 1/9$ при $\kappa = 2$:

$$p(10; 1/9) = \min\{2\min\{\mathrm{P}(\kappa \le 2; 1/9), \mathrm{P}(\kappa \ge 2; 1/9)\}, 1\} = 0{,}528.$$

Двусторонний доверительный интервал для ϕ с доверительной вероятностью 0,95 для p-значения, определенного равенством (28), есть отрезок [0,026; 1,253]. Вследствие наличия постоянного участка для функции (28) точечная оценка, описанная выше, определена неоднозначно – пересечение всех доверительных интервалов есть отрезок [0,194; 0,349].

Из свойств одностороннего p-значения и формулы (28) следует, что выполняются следующие полезные свойства:

– Если $1 > \alpha_{01} > \alpha_{02} > 0$, то двусторонний тест, значимый на уровне α_{02}, будет также значим и на уровне α_{01}.

– Вложенность доверительных областей: если $1 > \alpha_{01} > \alpha_{02} > 0$, то двусторонняя доверительная область уровня $(1 - \alpha_{01})$ содержится в двусторонней доверительной области уровня $(1 - \alpha_{02})$.

– Связность доверительной области: для любого $0 < \alpha_0 < 1$ двусторонняя доверительная область уровня $(1 - \alpha_0)$ представляет собой интервал.

Имея в виду введенные выше выражения для одностороннего среднего (14)–(15) и асимптотического (16) p-значений, можно определить аналогичные двусторонние варианты. Например, двусторонний вариант среднего p-значения определяется равенством

$$p_m(k;\ \phi_0) = 2\min\{\mathrm{P}(\varkappa < k;\ \phi_0),\ \mathrm{P}(\varkappa > k;\ \phi_0)\} + \mathrm{P}(\varkappa = k;\ \phi_0), \tag{29}$$

так как в силу равенства $\mathrm{P}(\varkappa < k;\ \phi) + \mathrm{P}(\varkappa > k;\ \phi) + \mathrm{P}(\varkappa = k;\ \phi) = 1$ правая часть равенства (29) не будет превышать 1.

Асимптотическое двустороннее *p*-значение можно определить равенством

$$p_a(k;\ \pi) = 2\min\left\{\frac{1}{2} - \Phi(z),\ \frac{1}{2} + \Phi(z)\right\},\ z = \frac{k - n\pi}{\sqrt{n\pi(1-\pi)}}. \tag{30}$$

Пример 7 (продолжение). Построим графики *p*-значений и найдем 95 %-е доверительные интервалы для формул (29), (30) для проверки гипотезы H_0: $\phi = 1/9$ при $\varkappa = 2$. Получаем (рис. 8):

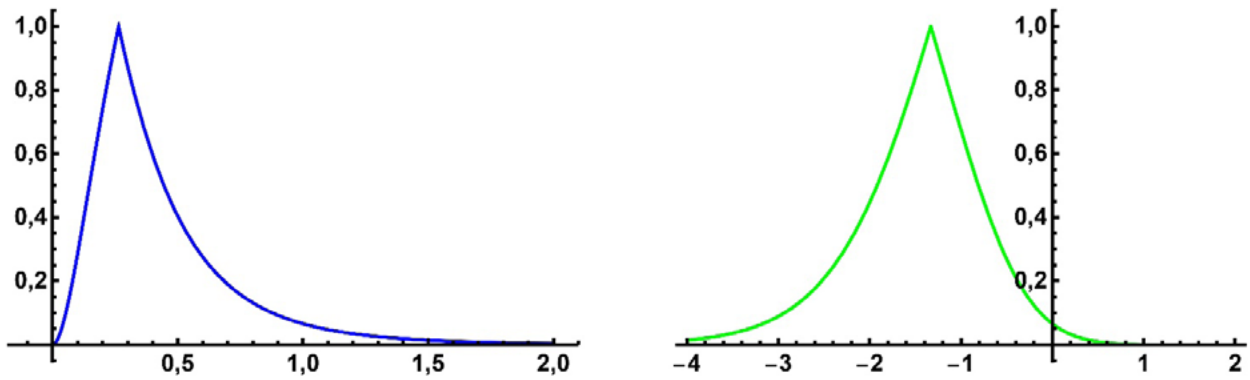


Рис. 8. График функции $p_m(2,\ \phi)$ для примера 7 в исходном масштабе (слева) и в логарифмическом относительно переменной $\beta = \ln\phi$ (справа)

95 %-й доверительный интервал для значения ϕ относительно *p*-значения (29) равен [0,036; 1,081]. Точечная оценка этого параметра равна 0,264 и достигается в точке максимума функции $p_m(2,\ \phi)$ (рис. 9).

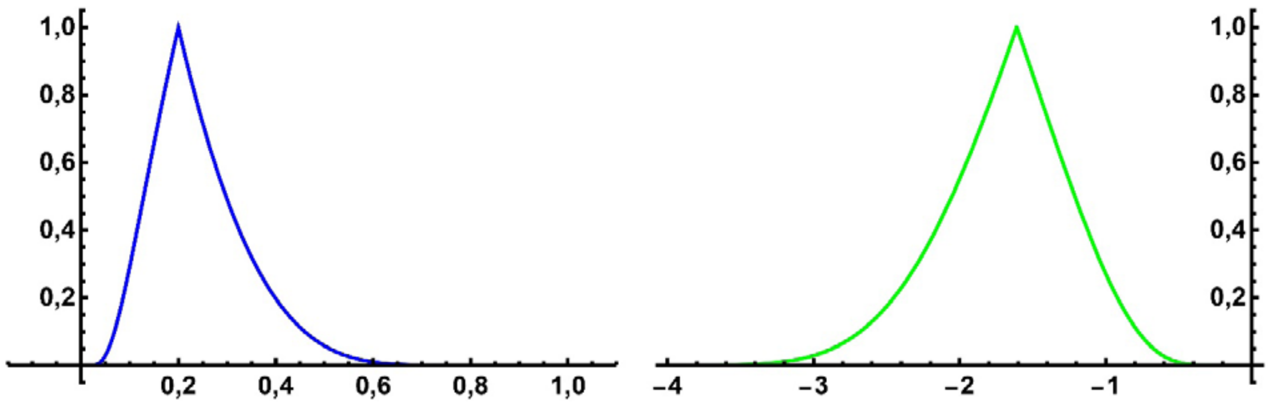


Рис. 9. График функции $p_a(2,\ \pi)$ для примера 7 в исходном масштабе (слева) и в логарифмическом относительно переменной $\beta = \ln\pi$ (справа)

95 %-й доверительный интервал для значения π относительно *p*-значения (30) равен [0,057; 0,510]. Точечная оценка этого параметра равна 0,2 и достигается в точке максимума функции $p_a(2,\ \phi)$ (рис. 9).

Можно отметить, что для двустороннего среднего *p*-значения (29) точка $\hat{\phi}$ максимума функции $p_m(k,\ \phi)$ удовлетворяет равенству $f(\hat{\phi};\ > k) = f(\hat{\phi};\ < k)$ и называется *несмещенной медианной оценкой параметра* ϕ. Она, очевидно, содержится в каждом доверительном интервале, т. е. является, таким образом, точечной оценкой параметра ϕ. Для рассматриваемого примера 7 несмещенная медианная оценка параметра π равна 0,264.

4.4. Вероятность покрытия и средняя длина доверительного интервала

Как мы видели, ключевые характеристики статистических выводов – *p*-значение и доверительный интервал можно получить из одной функции, а именно, функции *p*-значения, вычисление которой может быть определено различными способами, однако в большинстве случаев используется вероятность попадания в хвосты некоторого распределения. Одна из важных характеристик, которую можно использовать при теоретическом анализе того или иного метода вычисления функции *p*-значения, называется *вероятность покрытия*. Она позволяет определить, насколько хорошо данная функция *p*-значения задает доверительные интервалы для интересующего исследователя параметра. Смысл понятия *вероятность покрытия доверительным интервалом* состоит в том, что это есть доля тех доверительных интервалов (для заданного параметра ϕ) среди всех доверительных интервалов, которые для фиксированного значения параметра ϕ содержат этот параметр.

Более формально определение можно дать следующим образом. Пусть некоторый параметр ϕ принимает значения из множества Φ (например, для биномиальной вероятности $\Phi = (0;\ 1)$) и его истинное значение равно ϕ_0, т. е. $\phi_0 \in \Phi$. После анализа экспериментальных данных делается предположение о том, что на самом деле параметр содержится в более узком множестве *CS* (confidence set) $\phi_0 \in CS \subset \Phi$. Это предположение может быть верным или неверным, и общепринятым подходом является задание некоторого уровня вероятности, с которым это утверждение будет истинно (*доверительная вероятность*). Как правило, доверительное множество *CS* выбирается в соответствии с некоторым правилом (процедурой, формулой) ξ, которое зависит от экспериментальных данных и, следовательно, имеет случайный характер, который управляется неизвестной функцией распределения F, порождающей эти данные. В свою очередь, функция F зависит от параметра ϕ.

Определение 7. Вероятность покрытия доверительными множествами *CS,* выбираемыми в соответствии с правилом ξ, истинного значения ϕ_0 параметра ϕ, называется вероятность $C(CS, \phi_0)$, определяемой равенством

$$C(\phi_0) = P_{\phi_0}(\phi_0 \in CS(\xi)), \tag{31}$$

где множества $CS(\xi)$ берутся по всем допустим значениям правила ξ, а P_{ϕ_0} означает, что вероятность вычисляется с помощью функции распределения F при значении параметра $\phi = \phi_0$. Эта вероятность называется также фактическим уровнем (вероятностью) покрытия параметра ϕ_0.

Пример 8. Если $CI(\phi_0, \alpha_0)$ – доверительный интервал уровня $1 - \alpha_0$ для параметра ϕ_0 случайной величины $\mathcal{T}$ с носителем Ω, то вероятность покрытия (31) можно представить в виде

$$\mathrm{C}(\phi_0) = \sum_{u \in \Omega} \left[\phi_0 \in CI(\phi_0, \alpha_0)\right] \mathrm{P}_{\phi_0}(\mathcal{T} = \phi_0), \tag{32}$$

где [ ], как и раньше, скобка Айверсона.

Например, если рассматривается задача оценки биномиальной вероятности π в *n* испытаниях, то формула (32) принимает вид

$$C(\pi_0)=\sum_{k=0}^{n}\left[\pi_0\in CI(\pi_0,\alpha_0,k)\right]C_n^k\pi_0^k(1-\pi_0)^{n-k},$$

где $CI(\pi_0, \alpha_0, k)$ – доверительный интервал для биномиальной вероятности при условии, что наблюдается k успехов в n испытаниях.

Таким образом, величина вероятности покрытия существенно зависит от того, каким образом будет построен доверительный интервал и, следовательно, от того, какой выбрана функция p-значения.

Пример 7 (продолжение). Ниже показаны графики вероятности покрытия доверительными интервалами, построенными на основе обычной, средней и асимптотической двусторонней функций p-значения (соответственно (28), (29), (30)). Особенность дискретных распределений в том, что такие графики имеют пилообразный характер. При этом существенно, что при заданном уровне значимости вероятность покрытия может быть не меньше заданного доверительного уровня при любых значениях оцениваемого параметра (в заданном множестве) или иногда меньше этого уровня. В первом случае говорят о консервативном методе задания доверительного интервала, во втором о либеральном. Как видно из рис. 10, ситуация по отношению к вероятности покрытия такая же, как и ранее для статистического теста на основе соответствующей функции p-значения: доверительные интервалы, построенные по обычному p-значению (12), – консервативны, по среднему и асимптотическому p-значению – либеральны, причем доверительный интервал по среднему p-значению занимает промежуточное положение между иногда чрезмерно консервативным доверительным интервалом по обычному p-значению и чрезмерно либеральным по асимптотическому p-значению.

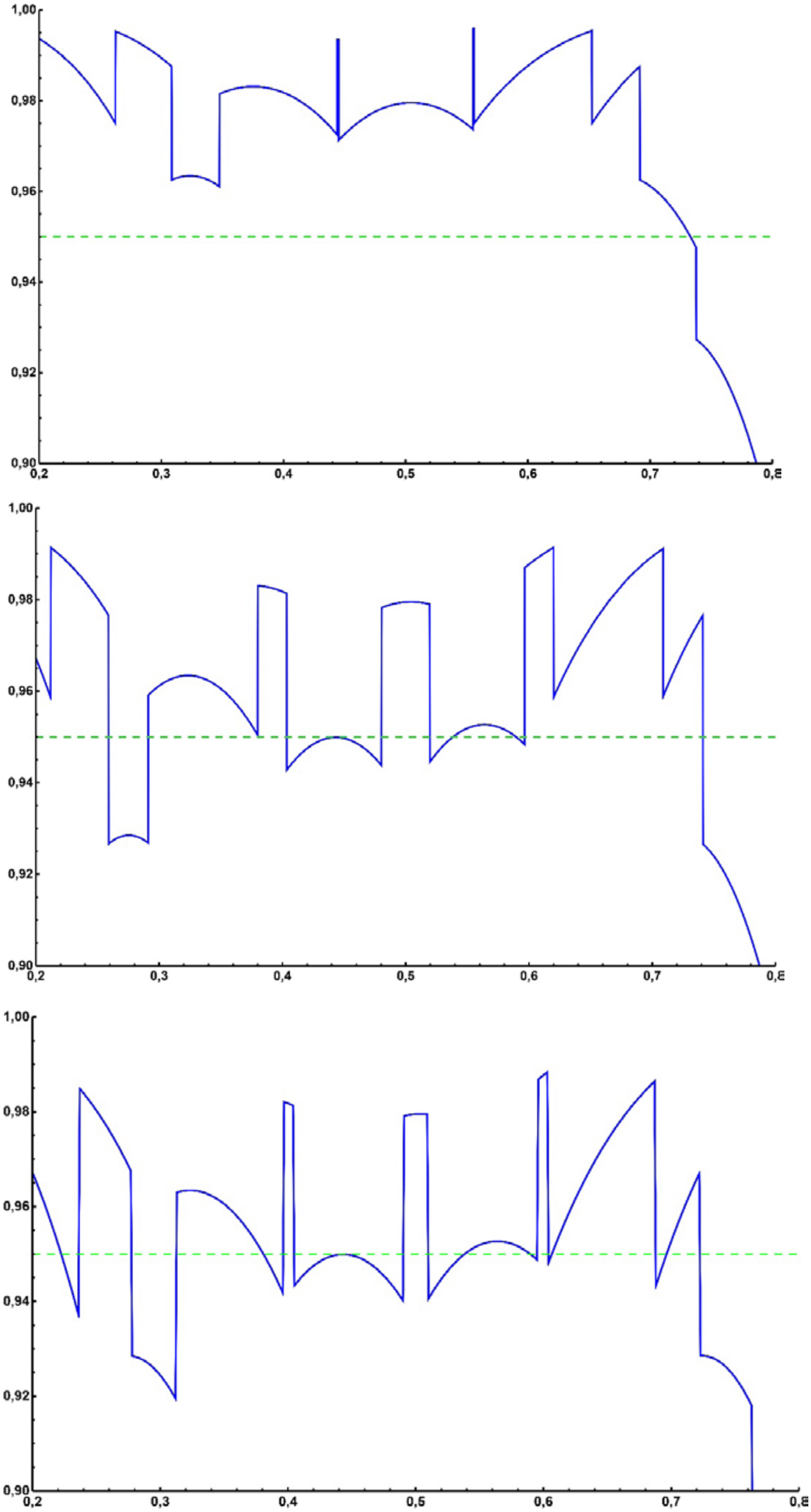


Рис. 10. Графики вероятности покрытия доверительным интервалом неизвестного значения параметра биномиального распределения для примера 7. Ось абсцисс – возможные значения вероятности оцениваемого параметра, ось ординат – соответствующая вероятность покрытия. Сверху вниз для доверительных интервалов на основе обычной функции p-значения (28), функции среднего p-значения (29), функции асимптотического p-значения (30)

С понятием «вероятность покрытия» тесно связана величина средней длины доверительного интервала, который строится по заданному правилу.

Определение 8. *Средней длиной* доверительного интервала для фиксированного значения оцениваемого параметра $\phi = \phi_0$, определяемого данным правилом, называется среднее значение длин этих интервалов, содержащих значение ϕ_0.

Пример 8 (продолжение). Из определения 8 следует, что в обозначениях примера 8 средняя длина $\mathrm{len}(\phi_0)$ доверительных интервалов, содержащих параметр ϕ_0, вычисляется по формуле

$$\mathrm{len}(\phi_0) = \sum_{u \in \Omega} \left[\phi_0 \in CI(\phi_0, \alpha_0) \right] \left\| CI(\phi_0, \alpha_0) \right\| \mathrm{P}_{\phi_0}(\mathcal{T} = \phi_0), \tag{33}$$

где $\left\| CI(\phi_0,\ \alpha_0) \right\|$ – длина интервала $CI(\phi_0,\ \alpha_0)$. В частности, для примера 7 оценки биномиальной вероятности эта формула принимает вид

$$\mathrm{len}(\pi_0) = \sum_{k=0}^{n} \left[\pi_0 \in CI(\pi_0, \alpha_0, k) \right] \left\| CI(\pi_0, \alpha_0, k) \right\| C_n^k \pi_0^k (1 - \pi_0)^{n-k}$$

Понятно, что вероятность покрытия доверительными интервалами и средняя длина доверительного интервала могут служить ориентирами в выборе подходящего способа вычисления доверительного интервала, однако на практике предпочтение отдается вероятности покрытия. На рис. 11 показаны графики средней длины доверительных интервалов для трех способов определения функции *p*-значения (28), (29) и (30) при уровне значимости $\alpha_0 = 0{,}05$.

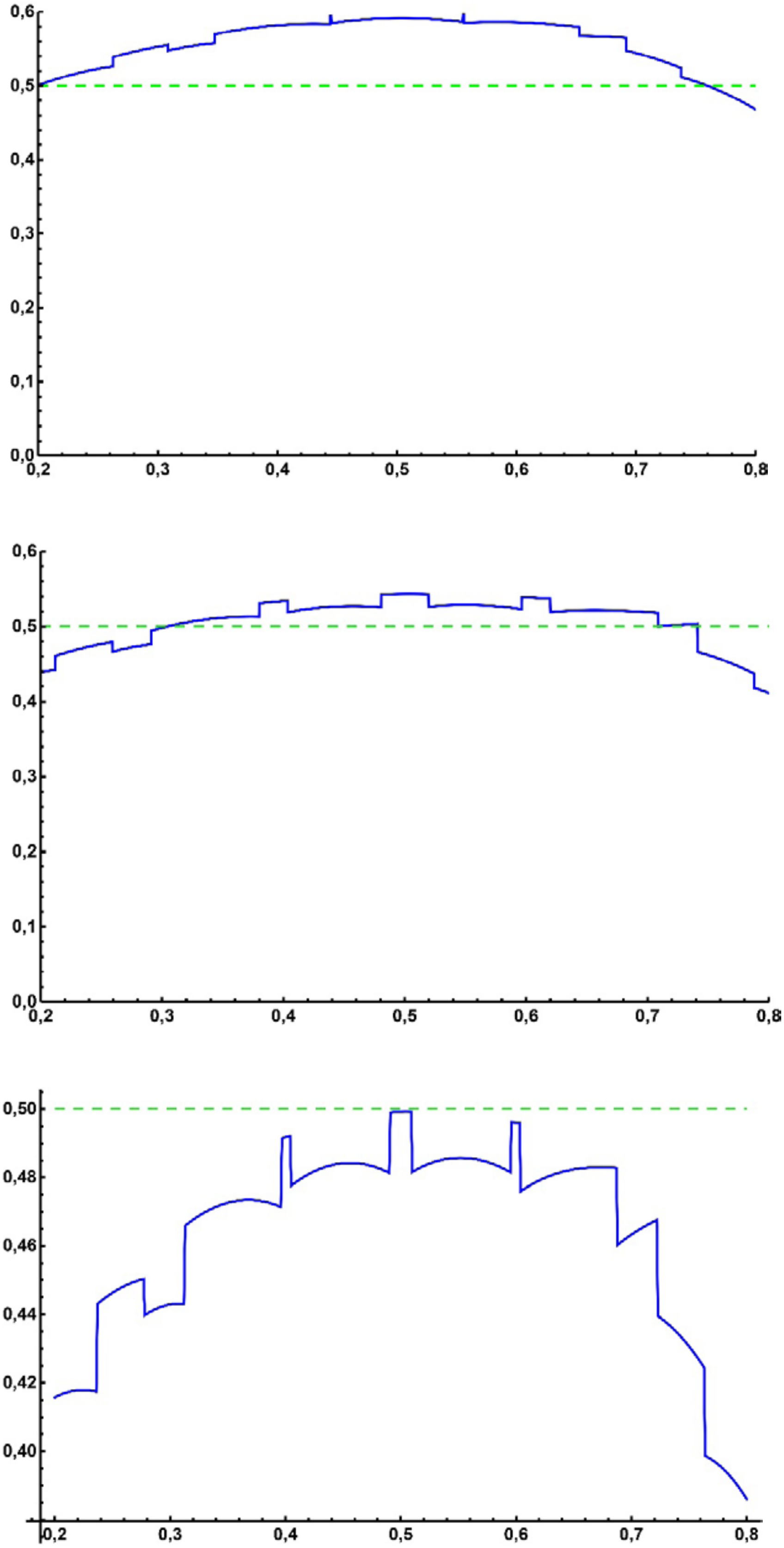


Рис. 11. Графики средней длины доверительных интервалов неизвестного значения параметра биномиального распределения для примера 7. Ось абсцисс – возможные значения вероятности оцениваемого параметра, ось ординат – соответствующая средняя длина. Сверху вниз для доверительных интервалов на основе: обычной функции *p*-значения (28), функции среднего *p*-значения (29), функции асимптотического *p*-значения (30)

Как видно из рис. 11, консервативная функция *p*-значения (28) порождает более широкие интервалы, чем две другие функции, причем наименьшие интервалы порождает асимптотическая функция (30), которая, однако, имеет худшую вероятность покрытия (см. рис. 10). Поэтому задача определения «наилучшего» способа определения доверительного интервала на основе понятий вероятности покрытия и средней длины доверительного интервала требует внимательного анализа имеющихся данных, так как эти критерии взаимно дополнительны. Увеличение одной из этих характеристик приводит к увеличению и другой, в то время как требуется увеличение вероятности покрытия при уменьшении длины интервала. Разумной видится следующая стратегия: среди всех рассматриваемых способов вычисления доверительного интервала следует выбрать те, для которых вероятность покрытия будет наибольшей для того диапазона значений, в котором рассматриваемый параметр должен находится, а среди этих методов выбрать тот, для которого средняя длина интервала будет наименьшей.

Пример 7 (продолжение). При описанных в примере 7 условиях ($n = 10$, $k = 2$) можно предположить, что оцениваемая биномиальная вероятность обнаружения побочных явлений находится в пределах (0,1; 0,3). Найдем минимальную вероятность покрытия при $\pi \in (0{,}1;\ 0{,}3)$ и максимальную среднюю длину доверительных интервалов при этом условии для функций *p*-значений (28)–(30) (рис. 12–14).

Таким образом, гарантированную вероятность покрытия дает только консервативный метод (28), который, однако, имеет наиболее широкие доверительные интервалы. Тем не менее, если известно, что значение оцениваемой вероятности не может превышать 0,25, то такую же гарантированную вероятность обеспечивает и функция (29), имеющая при этом заметно меньшую среднюю длину доверительных интервалов. Наиболее точные доверительные интервалы (наименьшей средней длины) обеспечивает асимптотическая функция (30), которая, однако, имеет худшую вероятность покрытия. Для ее корректного применения с гарантированной вероятностью покрытия необходимо ограничиваться теми интервалами возможных значений параметра, на которых вероятность покрытия больше заданного номинального уровня (например, (0,18; 0,24)).

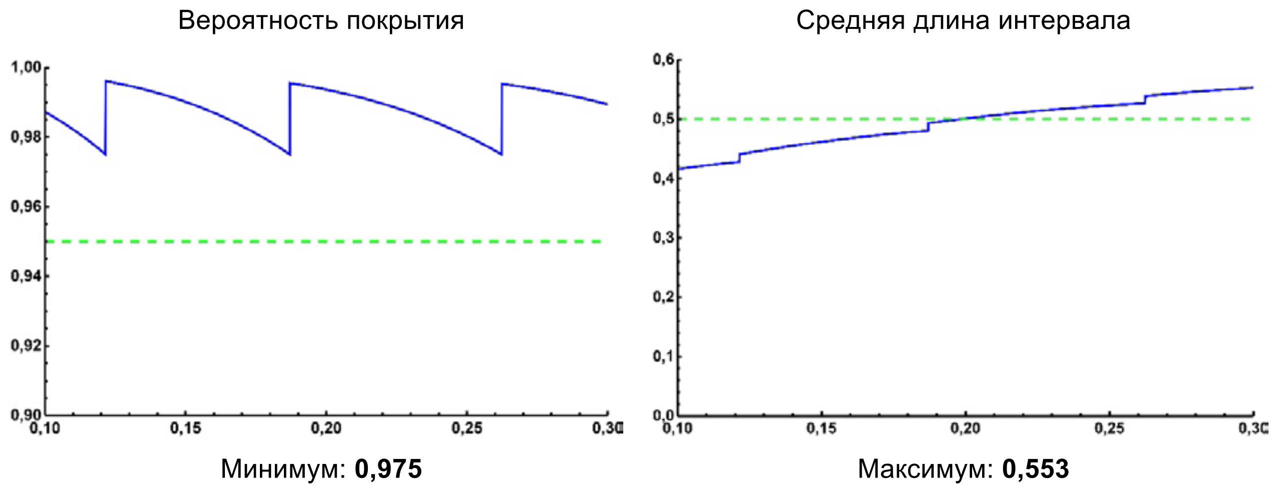


Рис. 12. Вероятность покрытия и средняя длина доверительных интервалов на интервале (0,1;0,3) для примера 7 для двусторонней функции *p*-значения (28). По оси абсцисс отложены возможные значения вероятности для примера 7, по оси ординат – вероятность покрытия (слева) и средняя длина доверительного интервала (справа)

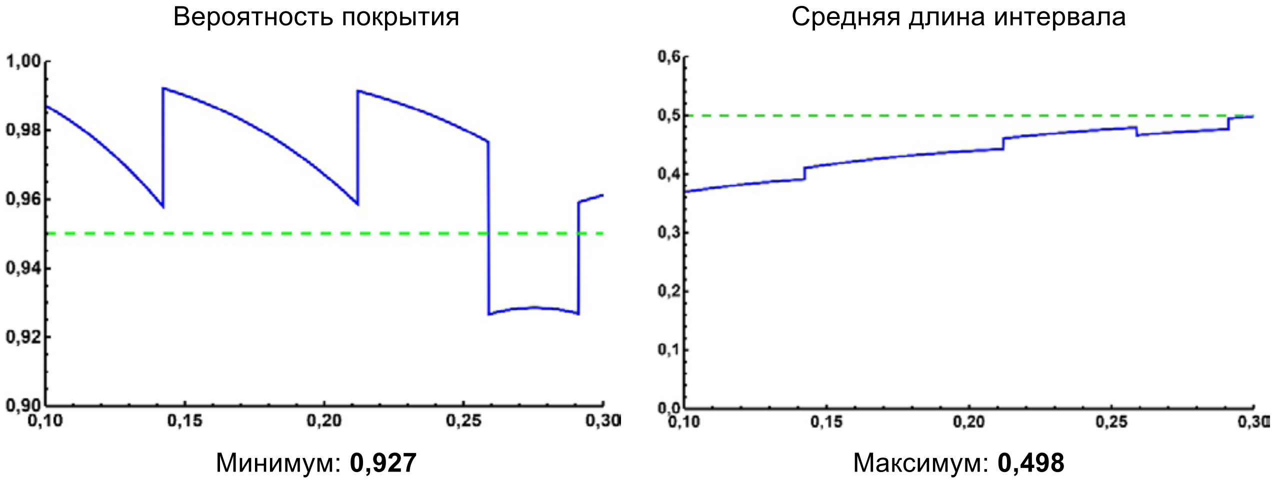


Рис. 13. Вероятность покрытия и средняя длина доверительных интервалов на интервале (0,1; 0,3) для примера 7 для функции *p*-значения (29)). По оси абсцисс отложены возможные значения вероятности для примера 7, по оси ординат – вероятность покрытия (слева) и средняя длина доверительного интервала (справа)

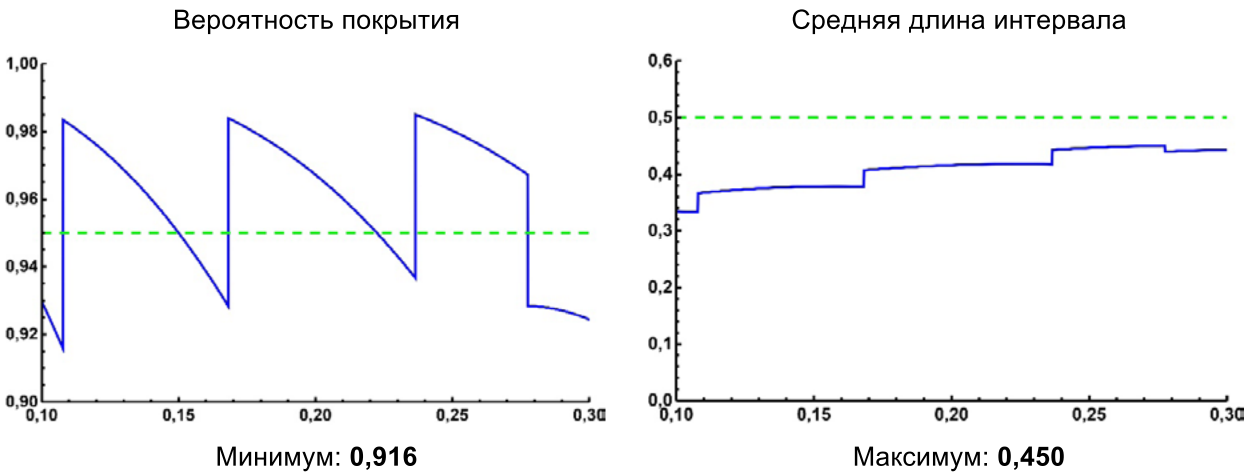


Рис. 14. Вероятность покрытия и средняя длина доверительных интервалов на интервале (0,1; 0,3) для примера 7 для функции р-значения (30). По оси абсцисс отложены возможные значения вероятности для примера 7, по оси ординат – вероятность покрытия (слева) и средняя длина доверительного интервала (справа)

5. Заключение

1. Анализ дискретных данных представляет специфическую область теоретической и прикладной статистики, имеющую широкие приложения в различных науках. Использование для анализа таких данных методов непрерывных случайных величин дает в лучшем случае приближение, точность которого остается дискуссионной, а чаще всего необсуждаемой. Однако для малых выборок эта точность может оказаться существенно недостаточной для получения статистически значимых результатов. В то же время теоретические исследования способов точного анализа дискретных данных ведутся уже довольно давно, и полученные важные результаты этой работы реализованы в нескольких статистических пакетах общего назначения и некоторых специализированных. Однако грамотное приложение этих методов и применение компьютерных программ требует знания основных понятий точного анализа

данных и понимания особенностей применения соответствующих методов. Несмотря на простоту аналитического аппарата изложенных методов, идеи точного анализа используют фундаментальные понятия теории статистического оценивания и вывода. Их знание необходимо не только при анализе дискретных данных, но и для понимания статистики вообще. Приведенные в работе определения и примеры помогут как в первоначальном знакомстве с этими понятиями, так и в последующем освоении методов точного анализа данных.

6. Список литературы

1. *Флетчер*, *Р.* Клиническая эпидемиология : Основы доказат. медицины / Р. Флетчер, С. Флетчер, Э. Вагнер; пер. с англ. под общ. ред. С. Е. Бащинского и С. Ю. Варшавского; [Пер. с англ. А. Д. Деев и др.]. – 3-е изд. – М. : Медиа Сфера, 2004. – 347 с. – ISBN 5-89084-011-8. – EDN QLFECB.
2. *Гринхальх*, *Т.* Основы доказательной медицины / Т. Гринхальх ; Триша Гринхальх ; пер. с англ.; под ред. И. Н. Денисова, К. И. Сайткулова. – 3-е изд. – М. : ГЭОТАР-Медиа, 2009. – 282 с. – ISBN 978-5-9704-1347-0. – EDN QLULXN.
3. *Эттингер*, *А. П.* Что такое доказательная медицина? / А. П. Эттингер, М. Е. Жарова // Доказательная гастроэнтерология. – 2021. – Т. 10, № 1. – С. 38–48. – DOI 10.17116/dokgastro20211001138. – EDN WDYYQS.
4. Общая эпидемиология с основами доказательной медицины / под ред. В. И. Покровского, Н. И. Брико. – 2-е изд., испр. и доп. – М. : ГЭОТАР-Медиа, 2012. – 496 с.
5. *Гланц*, *С.* Медико-биологическая статистика / С. Гланц ; под ред. Н. Е. Бузикашвили и Д. В. Самойлова. – М. : Практика, 1999. – 459 с. – ISBN 5-89816-009-4.
6. *Брико*, *Н. И.* Эпидемиология : учеб. в 2 т. / Н. И. Брико, Л. П. Зуева, В. И. Покровский [и др.]. – М. : Издательство «Медицинское информационное агентство», 2013. – 832 с. – ISBN 978-5-9986-0110-1. – EDN RTTTRD.
7. *Власов*, *В. В.* Эпидемиология : учеб. пособие для студентов вузов, обучающихся по специальности 040300 – Медико-профилактическое дело / В. В. Власов. – 2-е изд., испр. – М. : ГЭОТАР-Медиа, 2006. – 462 с. – ISBN 5-9704-0265-6. – EDN QLNFJP.
8. *Rothman*, *K. J.* Modern epidemiology / K. J. Rothman, S. Greenland, T. L. Lash. – 3rd ed., 2011. – 758 p. – EDN YDXWPS.
9. *Schoenbach*, *V. J.* Understanding the Fundamentals of Epidemiology / V. J. Schoenbach, W. D. Rosamond. – Department of Epidemiology School of Public Health University of North Carolina at Chapel Hill, 2000. – 584 c.
10. *Ahrens*, *W.* Handbook of Epidemiology / W. Ahrens, I. Pigeot. – Springer, 2005. – 1594 p. – ISBN 3-540-00566-8.
11. *Платонов*, *А. Е.* Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А. Е. Платонов. – М., 2000. – 52 с. – ISBN 5-7901-0022-8. – EDN PBDIJN.
12. *Малета*, *Ю. С.* Математические методы статистического анализа в биологии и медицине / Ю. С. Малета, В. В. Тарасов. – М. : Изд-во МГУ, 1982. – 179 с.
13. *Quinn*, *G. P.* Experimental Design and Data Analysis for Biologists / G. P. Quinn, M. J. Keough. – Cambridge University Press, 2002. – 557 p. – ISBN 0-521-00976-6.
14. *Любищев*, *А. А.* Дисперсионный анализ в биологии / А. А. Любищев – М. : МГУ, 1986. – 93 с.

15. Mathematics for Ecology and Environmental Sciences / под ред. Y. Takeuchi, Y. Iwasa, K. Sato. – Springer, 2007. – 186 p. – ISBN 10 3-540-34427-6.
16. *Legendre*, *P.* Numerical Ecology / P. Legendre, L. Legendre. – Elsevier, 1998. – 870 p. – ISBN 0-444-89250-8.
17. *Магнус*, *Я. Р.* Эконометрика. Начальный курс : учеб. / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М. : Издательский дом «Дело» РАНХиГС, 2021. – 504 с. – ISBN 978-5-85006-296-5. – EDN MDSVQL.
18. *Мхитарян*, *В. С.* Эконометрика : учеб. пособие / В. С. Мхитарян, М. Ю. Архипова, В. П. Сиротин. – М. : Евразийский открытый институт, 2012. – 224 с. – ISBN 978-5-374-00053-5. – EDN RBAOLL.
19. *Джонстон*, *Дж.* Эконометрические методы / Дж. Джонстон. – М. : Статистика, 1980. – 440 с.
20. Handbook of Computational Econometrics / под ред. D. A. Belsley, E. J. Kontoghiorghes. – Wiley, 2009. – 516 p. – ISBN 978-0-470-74385-0.
21. *Грин*, *У. Г.* Эконометрический анализ / У. Г. Грин; пер. с англ. под науч. ред. С. С. Синельникова и М. Ю. Турунцевой. – М. : Изд. дом «Дело», 2016. Кн. 1. – 760 с. Кн. 2 – 1476 с. – ISBN 978-5-7749-1158-5.
22. *Феллер*, *В.* Введение в теорию вероятностей и ее приложения = An Introduction to probability theory and its applications / В. Феллер; пер. со второго англ. изд. и предисл. Ю. В. Прохорова. – 2-е изд. – М. : URSS, 2009. – 751 с. – ISBN 978-5-397-01036-8. – EDN QJVXJV.
23. *Боровков*, *А. А.* Теория вероятностей : учеб. пособие для студентов высших учебных заведений, обучающихся по направлению подготовки 010100 «Математика» / А. А. Боровков. – 5-е изд., существенно перераб. и доп. – М. : URSS, 2009. – 652 с. – ISBN 978-5-397-00582-1. – EDN QJVLHP.
24. *Гаек*, *Я.* Теория ранговых критериев / Я. Гаек, З. Шидак. – М. : Наука, 1971. – 376 с.
25. *Холлендер*, *М.* Непараметрические методы статистики / М. Холлендер, Д. Вульф. – М. : Финансы и статистика, 1983. – 518 с.
26. *Corder*, *G. W.* Nonparametric statistics for non-statisticians. A Step-by-Step Approach / G. W. Corder, D. I. Foreman. – Wiley, 2009. – 264 p. – ISBN 978-0-470-45461-9.
27. *Fisher*, *R. A.* Statistical Methods for Research Workers. – 5th ed. / R. A. Fisher. – Oliver and Boyd, Edinburgh, 1934. – 336 p.
28. *Fisher*, *R. A.* The logic of inductive inference (with discussion) / R. A. Fisher // J. Royal Stat. Soc. – 1935. – Vol. 98. – Part I: 35–82.
29. *Irwin*, *J. O.* Tests of significance for differences between percentages based on small numbers / J. O. Irwin // Metron. – 1935. – Vol. 12: 83–94.
30. *Yates*, *F.* Contingency tables involving small numbers and the chi-square test / F. Yates // Supplement to the J. Royal Stat Soc, Series B. – 1934. – 1: 217–235.
31. Cytel Software Corporation. StatXact for Windows, version 6, Cytel Software Corporation, Cambridge, MA.
32. *Agresti*, *A.* Categorical data analysis / A. Agresti. – Wiley, 2002. – 732 p. – ISBN 0-471-36093-7.
33. *Sakamoto*, *Y.* Categorical data analysis by AIC / Y. Sakamoto // Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach: Vol. 2. Multivariate Statistical Modeling. 1994. – 255–269. – ISBN 978-94-011-0800-3. – DOI: 10.1007/978-94-011-0800-3_10.

34. *Birnbaum*, *A*. Confidence curves: an omnibus technique for estimation and testing statistical hypotheses / A. Birnbaum // J. Am Stat Assoc. – 1961. – Vol. 56. – C. 246–249. – DOI:10.1080/01621459.1961.10482107.

35. *Cox*, *D. R*. Analysis of binary data / D. R. Cox, E. J. Snell. – London, New York. Chapman and Hall, 1989. – 247 p. – ISBN 0-412-30620-4.

36. *Zelen*, *M*. The analysis of several 2×2 contingency tables / M. Zelen // Biometrika. – 1971. – Vol. 58. – No. 1. – P. 129. – EDN ILOSAN.

37. *Emerson*, *J. D*. Combining estimates of the odds ratio: the state of the art / J. D. Emerson // Stat Methods in Med Research. – 1994. – Vol. 3. – P. 157–178. – DOI: 10.1177/0962280 9400300204.

38. *Hirji*, *K. F*. Exact Analysis of Discrete Data / K. F. Hirji. – Chapman & Hall/CRC. – 2006. – 521 p. – ISBN 1-58488-070-8.

39. *May*, *R. M*. Host-Parasitoid Systems in Patchy Environments: A Phenomenological Model / R. M. May // J. Animal Ecology. – 1978. – Vol. 47, N 3. – P. 833–844. – DOI: 10.2307/3674.

40. *Ross*, *G. J. S*. The Negative Binomial Distribution / G. J. S. Ross, D. A. Preece // J. Royal Stat Soc. Series D. – 1985. – Vol. 34. – N 3. – C. 323–335. – DOI: 10.2307/2987659.

41. *Amrhein*, *L*. A mechanistic model for the negative binomial distribution of single-cell mRNA counts / L. Amrhein, K. Harsha, K. Fuchs // bioRxiv. – 2019. – DOI: 10.1101/657619.

42. *Johnson*, *N. L*. Univariate Discrete Distributions / N. L. Johnson, A. W. Kemp, S. Kotz // John Wiley & Sons, Inc., Hoboken, NJ, 2005. – 677 p. – ISBN 0-471-27246-9.

43. *Tuwei*, *K. E*. Power series distributions / K. E. Tuwei // http://erepository.uonbi.ac.ke/ bitstream/hadle/11295/71878/Tuwei_Power series distribution.pdf.

44. *Rubin-Delanchy*, *P*. Meta-analysis of mid-p-values: some new results based on the convex order / P. Rubin-Delanchy, N. A. Heard, D. J. Lawson // J. Am Stat Assoc. – 2018. – Vol. 114(527). – P. 1105–1112. – DOI: 10.1080/01621459.2018.1469994.

45. *Moyé*, *L*. Statistical reasoning in Medicine: The intuitive *p*-value primer / L. Moyé – Springer-Verlag, NY, 2006. – 301 p. – ISBN 0-387-32913-7.

46. *Леман*, *Э*. Проверка статистических гипотез / Э. Леман. – М. : Наука, 1979. – 409 с.

47. *Casella*, *G*. Statistical inference / G. Casella, R. L. Berger. – Duxbury, 2002. – 660 p. – ISBN 0-534-23412-6. – EDN QJRALL.

Сведения об авторах:

Панов Владимир Григорьевич, к. ф.-м. н., доцент, ведущий научный сотрудник лаборатории математического моделирования в экологии и медицине Института промышленной экологии УрО РАН, ул. С. Ковалевской, 20, г. Екатеринбург, Россия. Эл. почта: vpanov@ecko.uran.ru

ANALYSIS OF DISCRETE DATA. BASIC NOTIONS

V. G. Panov

Institute of Industrial Ecology of UrB of RAS, Ekaterinburg, Russia

The review article is devoted to methods of discrete data analysis. The main attention is paid to exact methods based on the representation of discrete probability distributions as power series distributions. The examples presented in the article relate to medicine and biology.

Key words: discrete probability distributions; power series distributions; exact data analysis methods; statistical hypotheses; confidence regions; power of a criterion; coverage probability.

References

1. *Fletcher, R.* Clinical epidemiology: The basics of evidence-based medicine / R. H. Fletcher; S. W. Fletcher, E. Wagner; Translation from English, Eds. S. E. Bashchinsky and C. Y. Varshavsky; [Trans. Eng. A. D. Deev and others]. – 3rd ed. – Moskow : Media Spere, 2004. – 347 p. – ISBN 5-89084-011-8. – EDN QLFECB.
2. *Greenhalgh, T.* The Basics of Evidence-Based Medicine and Healthcare: Translation from English / T. Greenhalgh; Trisha Greenhalgh; Translation Eds. I. N. Denisov, K. I. Saitkulov. – 3rd ed. – Moskow : GEOTAR-Media, 2009. – 282 p. – ISBN 978-5-9704-1347-0. – EDN QLULXN.
3. *Ettinger, A. P.* What is the evidence-based medicine? / A. P. Ettinger, M. E. Zharova // Russian J. of Evidence-Based Gastroenterology. – 2021. – Vol. 10, № 1. – P. 38–48. – DOI 10.17116/dokgastro20211001138. – EDN WDYYQS.
4. General Epidemiology with Fundamentals of Evidence-Based Medicine / Eds: V. I. Pokrovsky, N. I. Briko. – 2nd ed. – M.: GEOTAR-Media, 2012. – 496 p.
5. *Glantz, S.* Primer of biostatistics / S. Glantz. Trans. Under Eds N. E. Buzikashvili and и D. V. Samoilov. – M. : Praktika, 1999. – 459 p. – ISBN 5-89816-009-4.
6. *Briko, N. I.* Epidemiology: Textbook / N. I. Briko, L. P. Zueva, V. I. Pokrovsky [et al.]. – Moskow : Publisher «Medical information agency», 2013. – 832 p. – ISBN 978-5-9986-0110-1. – EDN RTTTRD.
7. *Vlasov, V. V.* Epidemiology / V. V. Vlasov. – 2nd ed. – Moskow : GEOTAR-Media, 2006. – 462 p. – ISBN 5-9704-0265-6. – EDN QLNFJP.
8. *Rothman, K. J.* Modern epidemiology / K. J. Rothman, S. Greenland, T. L. Lash. – 3rd ed., 2011. – P. 1–758. – EDN YDXWPS.
9. *Schoenbach, V. J.* Understanding the Fundamentals of Epidemiology / V. J. Schoenbach, W. D. Rosamond. – Department of Epidemiology School of Public Health University of North Carolina at Chapel Hill, 2000. – 584 c.

10. *Ahrens*, *W*. Handbook of Epidemiology / W. Ahrens, I. Pigeot. – Springer, 2005. – 1594 p. – ISBN 3-540-00566-8.
11. *Platonov*, *A*. *E*. Statistical Analysis in Medicine and Biology: Problems, Terminology, Logic, Computer Methods / A. E. Platonov. – Moskow, 2000. – 52 p. – ISBN 5-7901-0022-8. – EDN PBDIJN.
12. *Maleta*, *Y*. *S*. Mathematical Methods of Statistical Analysis in Biology and Medicine / Y. S. Maleta, V. V. Tarasov. – M. : Moskow State University, 1982. – 179 p.
13. *Quinn*, *G*. *P*. Experimental Design and Data Analysis for Biologists / G. P. Quinn, M. J. Keough. – Cambridge University Press, 2002. – 557 p. – ISBN 0-521-00976-6.
14. *Lyubishchev*, *A*. *A*. Analysis of variance in biology / A. A. Lyubishchev. – M. : MGU, 1986. – 93 p.
15. Mathematics for Ecology and Environmental Sciences / Eds: Y. Takeuchi, Y. Iwasa, K. Sato. – Springer, 2007. – 186 p. – ISBN 10 3-540-34427-6.
16. *Legendre*, *P*. Numerical Ecology / P. Legendre, L. Legendre. – Elsevier, 1998. – 870 p. – ISBN 0-444-89250-8.
17. *Magnus*, *Y*. *R*. Econometrics. Primary course: textbook / Y. R. Magnus, P. K. Katyshev, A. A. Peresetsky. – Moskow: Publishing House «Delo» RANEPA, 2021. – 504 p. – ISBN 978-5-85006-296-5. – EDN MDSVQL.
18. *Mkhitaryan*, *V*. *S*. Econometrics : Tutorial / V. S. Mkhitaryan, M. Yu. – Moscow : Eurasian Open Institute, 2012. – 224 p. – ISBN 978-5-374-00053-5. – EDN RBAOLL.
19. *Johnston*, *J*. Econometric methods / J. Johnston. – M. : Statistika, 1980. – 440 p.
20. Handbook of Computational Econometrics / Eds. D. A. Belsley, E. J. Kontoghiorghes. – Wiley, 2009. – 516 p. – ISBN 978-0-470-74385-0.
21. *Greene*, *W*. *H*. Econometric analysis / W. H. Greene; Trans. from Eng. under Eds S. S. Sinelnikov and M. Y. Turuntseva. – M. : Publishing House «Delo», 2016. Vol. 1 – 760 p. Vol. 2 – 1476 p. – ISBN 978-5-7749-1158-5.
22. *Feller*, *W*. An Introduction to probability theory and its applications / W. Feller; Trans. of 2nd Eng. ed. – Moskow : URSS, 2009. – 751 p. – ISBN 978-5-397-01036-8. – EDN QJVXJV.
23. *Borovkov*, *A*. *A*. Probability theory / A. A. Borovkov. – 5th ed. – Moskow : URSS, 2009. – 652 p. – ISBN 978-5-397-00582-1. – EDN QJVLHP.
24. *Hajek*, *J*. Theory of rank tests / J. Hajek, Z. Shidak. – M. : Nauka, 1971. – 376 p.
25. *Hollander*, *M*. Nonparametric statistical methods / M. Hollander, D. Wolfe. – M. : Finance and Statistics, 1983. – 518 p.
26. *Corder*, *G*. *W*. Nonparametric statistics for non-statisticians. A Step-by-Step Approach / G. W. Corder, D. I. Foreman. – Wiley, 2009. – 264 p. – ISBN 978-0-470-45461-9.
27. *Fisher*, *R*. *A*. Statistical Methods for Research Workers. – 5th ed. / R. A. Fisher. – Oliver and Boyd, Edinburgh, 1934. – 336 c.
28. *Fisher*, *R*. *A*. The logic of inductive inference (with discussion) / R. A. Fisher // J. Royal Stat. Soc. – 1935. – Vol. 98. – Part I: 35–82.
29. *Irwin*, *J*. *O*. Tests of significance for differences between percentages based on small numbers / J. O. Irwin // Metron. – 1935. – Vol. 12: 83–94.
30. *Yates*, *F*. Contingency tables involving small numbers and the chi-square test / F. Yates // Supplement to the J. Royal Stat Soc, Series B. – 1934. – 1: 217–235.
31. Cytel Software Corporation. StatXact for Windows, version 6, Cytel Software Corporation, Cambridge, MA.

32. *Agresti*, *A*. Categorical data analysis / A. Agresti. – Wiley, 2002. – 732 p. – ISBN 0-471-36093-7.
33. *Sakamoto*, *Y*. Categorical data analysis by AIC / Y. Sakamoto // Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach: Vol. 2. Multivariate Statistical Modeling. 1994. – 255–269. – ISBN 978-94-011-0800-3. – DOI: 10.1007/978-94-011-0800-3_10.
34. *Birnbaum*, *A*. Confidence curves: an omnibus technique for estimation and testing statistical hypotheses / A. Birnbaum // J. Am Stat Assoc. – 1961. – Vol. 56. – C. 246–249. – DOI:10.1080/01621459.1961.10482107.
35. *Cox*, *D. R*. Analysis of binary data / D. R. Cox, E. J. Snell. – London, New York. Chapman and Hall. – 1989. – 247 c. –ISBN 0-412-30620-4.
36. *Zelen*, *M*. The analysis of several 2×2 contingency tables / M. Zelen // Biometrika. – 1971. – Vol. 58, No. 1. – P. 129. – EDN ILOSAN.
37. *Emerson*, *J. D*. Combining estimates of the odds ratio: the state of the art / J. D. Emerson // Stat Methods in Med Research. – 1994. – Vol. 3. – C. 157–178. – DOI: 10.1177/096228029400300204.
38. *Hirji*, *K. F*. Exact Analysis of Discrete Data / K. F. Hirji. – Chapman & Hall/CRC. – 2006. – 521 c. – ISBN 1-58488-070-8.
39. *May*, *R. M*. Host-Parasitoid Systems in Patchy Environments: A Phenomenological Model / R. M. May // J Animal Ecology. – 1978. – Vol. 47, N 3. – C. 833–844. – DOI: 10.2307/3674.
40. *Ross*, *G. J. S.* The Negative Binomial Distribution / G. J. S. Ross, D. A. Preece // J. Royal Stat Soc. Series D. – 1985. – Vol. 34, N 3. – C. 323–335. – DOI: 10.2307/2987659.
41. *Amrhein*, *L*. A mechanistic model for the negative binomial distribution of single-cell mRNA counts / L. Amrhein, K. Harsha, K. Fuchs // bioRxiv. – 2019. – DOI: 10.1101/657619.
42. *Johnson*, *N. L*. Univariate Discrete Distributions / N. L. Johnson, A. W. Kemp, S. Kotz // John Wiley & Sons, Inc., Hoboken, NJ, 2005. – 677 c. – ISBN 0-471-27246-9.
43. *Tuwei*, *K. E*. Power series distributions / K. E. Tuwei // http://erepository.uonbi.ac.ke/bitstream/hadle/11295/71878/Tuwei_Power series distribution.pdf.
44. *Rubin-Delanchy*, *P*. Meta-analysis of mid-p-values: some new results based on the convex order / P. Rubin-Delanchy, N. A. Heard, D. J. Lawson // J. Am Stat Assoc. – 2018. – Vol. 114(527). – P. 1105–1112. – DOI: 10.1080/01621459.2018.1469994.
45. *Moyé*, *L*. Statistical reasoning in Medicine: The intuitive *p*-value primer / L. Moyé – Springer-Verlag, NY, 2006. – 301 c. – ISBN 0-387-32913-7.
46. *Lehmann*, *E*. Testing statistical hypotheses / E. Lehmann. – M. : Nauka, 1979. – 409 p.
47. *Casella*, *G*. Statistical inference / G. Casella, R. L. Berger. – Duxbury, 2002. – 660 c. – ISBN 0-534-23412-6. – EDN QJRALL