Слабая вероятностная аксиоматика

К. В. Воронцов (vokov@forecsys.ru)

27 сентября 2009 г.

Содержание

В прикладных задачах анализа данных число наблюдений всегда конечно, тем не менее, широко используется понятие *вероятности*, подразумевающее предельный переход к бесконечной выборке. Известно, что при малых объёмах данных асимптотические методы теории вероятностей и математической статистики могут приводить к неточным или даже ошибочным выводам [13, 14]. Возникают вопросы: обосновано ли использование инфинитарного (асимптотического) понятия вероятности в задачах анализа данных? Всегда ли понятие вероятности является инфинитарным?

Рассмотрим фундаментальную задачу теории вероятностей, тесно связанную с законом больших чисел: оценить вероятность большого отклонения частоты $\nu(S,X)$ события S на конечной выборке X от вероятности $P(S)$ данного события:

$$P_\varepsilon = \mathsf{P}\big\{|\nu(S,X) - P(S)| > \varepsilon\big\}. \tag{0.1}$$

Если вероятностная мера P неизвестна, то для вычисления вероятности события $P(S)$ необходимо провести бесконечное число наблюдений, что на практике невозможно. В результате оказывается, что вероятность большого уклонения P_ε непосредственно не может быть измерена в эксперименте как частота события $\big\{X\colon |\nu(S,X) - P(S)| > \varepsilon\big\}$, поскольку само наступление этого события не может быть точно идентифицировано. Разумеется, вероятность P_ε можно оценивать математически, и не зная $P(S)$. Однако если задаться целью проверить теоретический результат экспериментально, то вероятность $P(S)$ в правой части (0.1) придётся заменять некоторой её оценкой $\hat{P}(S)$. Такая подмена означает, что оценивается вероятность совсем другого события $\big\{X\colon |\nu(S,X) - \hat{P}(S)| > \varepsilon\big\}$, для которого проверяемая теоретическая оценка уже может быть неверна. Но тогда разумнее было бы выводить теоретическую оценку именно для этого события.

Данная проблема не возникает, если с самого начала отказаться от введения вероятности $P(S)$. Она определяется как предел частоты $\nu(S,X')$ события S на произвольной выборке X' при $|X'| \to \infty$. Однако практический интерес представляет именно сама частота $\nu(S,X')$, как величина, непосредственно наблюдаемая в эксперименте. Изменим постановку задачи (0.1) и будем оценивать вероятность большого уклонения частот события S в двух различных выборках:

$$Q_\varepsilon = \mathsf{P}\big\{|\nu(S,X) - \nu(S,X')| > \varepsilon\big\}. \tag{0.2}$$

Если предполагать, что обе выборки независимы, то для определения вероятности Q_ε уже не нужно ни бесконечного числа испытаний, ни знания вероятностной меры на исходном пространстве событий. Вероятность Q_ε может быть определена как доля тех разбиений объединённой выборки $X \cup X'$ на две подвыборки, при которых имеет место большое отклонение частот. Эта вероятность является финитарной. Она может быть найдена точно, исходя из комбинаторных соображений. Более того, она может быть непосредственно измерена в эксперименте, так как идентификация события $\big\{X, X'\colon |\nu(S,X) - \nu(S,X')| > \varepsilon\big\}$ не вызывает затруднений.

Таким образом, вероятности $P(S)$ и P_ε в (0.1) имеют различную природу. Вероятность $P(S)$ принципиально инфинитарна — для её определения требуется пре-

дельный переход: $\nu(S, X') \to P(S)$ при $|X'| \to \infty$. Вероятность P_ε также инфинитарна, но после замены $P(S)$ на частоту $\nu(S, X')$ она принимает финитарный вид Q_ε, допускающий и точное вычисление, и непосредственное эмпирическое измерение.

Приведённые соображения порождают ряд вопросов. Возможно ли на уровне аксиоматики запретить использование инфинитарных вероятностей и «событий», которые не могут быть идентифицированы в эксперименте? Нельзя ли ограничиться употреблением только таких величин, которые могут быть точно определены по конечным выборкам? Возможно ли при таком ограничении построить теорию, содержащую основные фундаментальные факты теории вероятностей, математической статистики, теории информации, теории статистического обучения, относящиеся к конечным выборкам?

Современная теория вероятностей возникла из стремления объединить в рамках единого формализма частотное понятие вероятности, берущее начало от азартных игр, и континуальное, идущее от геометрических задач, таких как задача Бюффона о вероятности попадания иглы в паркетную щель. В аксиоматике Колмогорова континуальное понятие берётся за основу как более общее. Ради этой общности в теорию вероятностей привносятся гипотезы сигма-аддитивности и измеримости — технические предположения из теории меры, имеющие довольно слабые эмпирические обоснования [1]. Однако далеко не во всех задачах, связанных со случайностью, определение вероятности как континуальной меры действительно необходимо.

Обратим внимание на высказывание А. Н. Колмогорова: «представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений. На независимой ценности чисто комбинаторного подхода к теории информации я неоднократно настаивал в своих лекциях» [12, стр. 252]. Один из вариантов комбинаторно-алгебраического построения теории информации предложен в книге В. Д. Гоппы [8]. Однако высказывание А. Н. Колмогорова в значительной степени относится и к математической статистике, поскольку она также изучает свойства конечных выборок. Ученик А. Н. Колмогорова Ю. К. Беляев в предисловии к книге «Вероятностные методы выборочного контроля» пишет: «возникло глубокое убеждение, что в теории выборочных методов можно получить содержательные аналоги большинства основных утверждений теории вероятностей и математической статистики, которые к настоящему времени найдены в предположении взаимной независимости результатов измерений» [2, стр. 9].

Уместно привести ещё одно высказывание А. Н. Колмогорова: «Чистая математика благополучно развивается как по преимуществу наука о бесконечном... Весьма вероятно, что с развитием современной вычислительной техники будет понято, что в очень многих случаях разумно изучение реальных явлений вести, избегая промежуточный этап их стилизации в духе представлений математики бесконечного и непрерывного, переходя прямо к дискретным моделям.» [12, стр. 239].

В данной работе предлагается *слабая вероятностная аксиоматика*, в которой допускаются только финитарные вероятности. Понятие вероятности вводится без использования теории меры и без предельного перехода к выборкам бесконечной длины. Единственное вероятностное предположение заключается в том, что объек-

ты выборки становятся известны в случайном порядке, другими словами, что все перестановки выборки равновероятны, или что наблюдения в выборке независимы. Столь слабого вероятностного допущения оказывается достаточно, чтобы установить сходимость частот (аналог закона больших чисел), сходимость эмпирических распределений (критерий Колмогорова-Смирнова), получить многие ранговые и перестановочные критерии. Слабая аксиоматика полностью согласуется с колмогоровской, но её область применимости ограничена *анализом конечных выборок*.

В данной главе с позиций слабой аксиоматики рассматриваются классические задачи эмпирического предсказания и проверки статистических гипотез. В последующих главах будут рассматриваться задачи статистического обучения.

1 Основная аксиома

Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ — фиксированное множество попарно различных объектов, называемое *генеральной выборкой*. Обозначим через S_L группу перестановок L элементов. Всевозможные перестановки элементов генеральной выборки будем обозначать через $\tau\mathbb{X}$, $\tau \in S_L$.

Аксиома 1.1 (о независимости элементов выборки). *Все $L!$ перестановок генеральной выборки $\tau\mathbb{X}$, $\tau \in S_L$, имеют одинаковые шансы реализоваться.*

Это единственная аксиома слабой вероятностной аксиоматики. Она позволяет определить понятие вероятности как «долю перестановок выборки».

Опр. 1.1. *Пусть задан предикат $\psi\colon \mathbb{X}^L \to \{0,1\}$. Если $\psi(\tau\mathbb{X}) = 1$, то будем говорить, что событие ψ произошло на перестановке $\tau\mathbb{X}$. Вероятностью события ψ называется доля перестановок $\tau\mathbb{X}$, на которых произошло событие ψ:*

$$\mathsf{P}_\tau\, \psi(\tau\mathbb{X}) = \frac{1}{L!} \sum_{\tau \in S_L} \psi(\tau\mathbb{X}). \tag{1.1}$$

В слабой аксиоматике вероятность события зависит от состава объектов генеральной выборки $\mathbb{X}$, но не зависит от порядка их перечисления. Функция распределения и математическое ожидание также зависят от выборки.

Опр. 1.2. *Пусть $\xi\colon \mathbb{X}^L \to \mathbb{R}$ — вещественная функция. Функцией распределения величины ξ на выборке $\mathbb{X}$ будем называть функцию $F_\xi\colon \mathbb{R} \to [0,1]$ вида*

$$F_\xi(z) = \mathsf{P}_\tau\big[\xi(\tau\mathbb{X}) \leqslant z\big]. \tag{1.2}$$

Опр. 1.3. *Математическим ожиданием величины $\xi\colon \mathbb{X}^L \to \mathbb{R}$ на выборке $\mathbb{X}$ будем называть её среднее арифметическое по всем перестановкам τ:*

$$\mathsf{E}_\tau \xi(\tau\mathbb{X}) = \frac{1}{L!} \sum_{\tau \in S_L} \xi(\tau\mathbb{X}). \tag{1.3}$$

Заметим, что вероятность и матожидание формально определяются одинаково. Знаки P_τ и E_τ можно понимать как операцию среднего арифметического по всем перестановкам: $\mathsf{P}_\tau \equiv \mathsf{E}_\tau \equiv \frac{1}{L!}\sum_{\tau\in S_L}$.

Вероятность как доля разбиений выборки. Рассмотрим важный частный случай, когда предикат ψ является функцией двух подвыборок: $X \subset \mathbb{X}$ длины ℓ и её дополнения $\bar{X} = \mathbb{X}\backslash X$ длины $k = L - \ell$, причём значение предиката $\psi(\mathbb{X}) = \varphi(X, \bar{X})$ не зависит от порядка элементов в подвыборках X и $\bar{X}$. Тогда из аксиомы 1.1 следует, что все C_L^ℓ разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ имеют равные шансы реализоваться. Следовательно, в данном случае вероятность можно определять не только как долю перестановок, но и как долю разбиений выборки $\mathbb{X}$:

$$\mathsf{P}_\tau\, \psi(\tau\mathbb{X}) = \mathsf{P}\, \varphi(X, \bar{X}) = \frac{1}{C_L^\ell} \sum_{(X,\bar{X})} \varphi(X, \bar{X}).$$

Рассмотрим более общий случай, когда предикат ψ является функцией q непересекающихся подвыборок, $\psi(\mathbb{X}) = \varphi(X^{\ell_1}, \ldots, X^{\ell_q})$, где $\ell_1, \ldots, \ell_q$ — длины подвыборок, $X^{\ell_1} \sqcup \cdots \sqcup X^{\ell_q} = \mathbb{X}$. Допустим, что значение предиката φ не зависит от порядка элементов в подвыборках. Тогда вероятность определяется и в этом случае как доля разбиений, при которых реализуется событие φ:

$$\mathsf{P}\, \varphi(X^{\ell_1}, \ldots, X^{\ell_q}) = \frac{\ell_1! \cdots \ell_q!}{L!} \sum_{(X^{\ell_1}, \ldots, X^{\ell_q})} \varphi(X^{\ell_1}, \ldots, X^{\ell_q}).$$

§1.1 Задачи эмпирического предсказания

Задача эмпирического предсказания является одной из центральных в теории вероятностей и математической статистике. Она часто возникает в приложениях, связанных с прогнозированием и принятием решений. Неформально задача состоит в том, чтобы, получив выборку данных, предсказать определённые свойства аналогичных данных, пока ещё неизвестных, и заранее оценить точность предсказания.

Рассмотрим эксперимент, в котором реализуется одно из C_L^ℓ равновероятных разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$. После реализации разбиения наблюдателю сообщается подвыборка X. Не зная скрытой подвыборки $\bar{X}$, требуется предсказать значение $t = T(\bar{X}, X)$ заданной функции T, существенно зависящее от скрытой подвыборки $\bar{X}$. Требуется также оценить надёжность предсказания, то есть вероятность того, что предсказанное значение $\hat{t} = \hat{T}(X)$ будет не сильно отличаться от истинного значения t.

Рассмотрим два варианта формальной постановки задачи.

Задача 1.1. При заданной функции $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$ построить *предсказывающую функцию* $\hat{T}\colon \mathbb{X}^\ell \to R$, значение которой на наблюдаемой подвыборке $\hat{t} = \hat{T}(X)$ приближало бы неизвестное значение $t = T(\bar{X}, X)$, а также оценить надёжность предсказаний, указав невозрастающую *оценочную функцию* $\eta(\varepsilon)$ такую, что

$$\mathsf{P}\big[d\big(\hat{T}(X), T(\bar{X}, X)\big) > \varepsilon\big] \leqslant \eta(\varepsilon), \tag{1.4}$$

где $d\colon R\times R \to \mathbb{R}$ — заданная функция, характеризующая величину отклонения $d(\hat{t},t)$ предсказанного значения $\hat{t}$ от неизвестного истинного значения t.

Обозначим эту задачу через $\mathscr{P}_1\big\langle R,\ d(\hat{t},t),\ T(\bar{X},X),\ \hat{T}(X)\big\rangle$.

Параметр ε называется *точностью*, а величина $(1-\eta(\varepsilon))$ — *надёжностью* предсказания. Если в (1.4) достигается равенство, то $\eta(\varepsilon)$ называется *точной оценкой*.

Обычно предполагается, что $\varepsilon > 0$ и $0 < \eta < 1$. Если (1.4) выполняется при достаточно малых ε и η, то говорят, что в окрестности предсказываемого значения $t = T(\bar{X},X)$ имеет место *концентрация вероятности* [25].

Для упрощения обозначений условимся далее опускать второй аргумент X функции $T(\bar{X},X)$, если она зависит только от $\bar{X}$. Более того, будем иногда опускать все аргументы функций T, $\hat{T}$, подразумевая, соответственно, $T(\bar{X},X)$ и $\hat{T}(X)$.

Как станет видно далее, во многих задачах в качестве предсказывающей функции $\hat{T}(X)$ выбирается $T(X)$. Тем не менее, роль функций T и $\hat{T}$ принципиально различна. Функция T задаётся в самой постановке задачи, тогда как предсказывающую функцию $\hat{T}$ наблюдатель имеет право выбирать по собственному усмотрению.

Задача $\mathscr{P}_1$ допускает следующее естественное обобщение.

Опр. 1.4. *Семейство подмножеств $\Omega_\varepsilon(X) \subseteq R$ с параметром ε называется семейством (расширяющихся) вложенных подмножеств, если для любого $X \in \mathbb{X}^\ell$ и любых допустимых значений параметра ε,ε' из $\varepsilon \leqslant \varepsilon'$ следует $\Omega_\varepsilon(X) \subseteq \Omega_{\varepsilon'}(X)$.*

Задача 1.2. При заданной функции $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$ построить *семейство вложенных подмножеств* $\Omega_\varepsilon(X) \subseteq R$ и невозрастающую *оценочную функцию* $\eta(\varepsilon)$, для которых выполнено неравенство

$$\mathsf{P}\big[T(\bar{X},X) \notin \Omega_\varepsilon(X)\big] \leqslant \eta(\varepsilon).$$

Обозначим эту задачу через $\mathscr{P}_2\big\langle R,\ T(\bar{X},X),\ \Omega_\varepsilon(X)\big\rangle$.

Задача $\mathscr{P}_1$ является частным случаем задачи $\mathscr{P}_2$, если в качестве семейства вложенных подмножеств взять $\Omega_\varepsilon(X) = \big\{t \in R \;\big|\; d(\hat{T}(X),t) \leqslant \varepsilon\big\}$.

Примеры задач эмпирического предсказания. Выбирая множество R, функцию T и семейство Ω_ε (или функции $\hat{T}$ и d вместо Ω_ε), можно получить многие классические задачи теории вероятностей, математической статистики, статистического обучения. Приведём основные постановки, рассматриваемые в данной работе.

Задача 1.3 (оценивание частоты события). Пусть $S \subseteq \mathbb{X}$ — некоторое множество объектов; будем называть его «событием». Введём функцию *частоты события* S на произвольной конечной выборке $U \subseteq \mathbb{X}$:

$$\nu(U) = \frac{|S \cap U|}{|U|}.$$

Требуется предсказать частоту события S на скрытой выборке $\bar{X}$ по его частоте на наблюдаемой выборке X и оценить надёжность предсказания:

$$\mathsf{P}\big[|\nu(\bar{X}) - \nu(X)| \geqslant \varepsilon\big] \leqslant \eta(\varepsilon); \tag{1.5}$$

Очевидно, данная задача есть $\mathscr{P}_1\big\langle \mathbb{R},\ |t-\hat{t}|,\ \nu(\bar{X}),\ \nu(X)\big\rangle$.

Иногда (в тех случаях, когда S интерпретируется как «нежелательное событие») требуется получить одностороннюю верхнюю оценку:

$$\mathsf{P}\big[\nu(\bar{X}) - \nu(X) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.6}$$

Очевидно, данная задача есть $\mathscr{P}_1\big\langle \mathbb{R},\ (t-\hat{t}),\ \nu(\bar{X}),\ \nu(X)\big\rangle$.

Задача 1.3 имеет фундаментальное значение для теории вероятностей и тесно связана с законом больших чисел и предельными теоремами. Она возникает и в практических приложениях, таких, как выборочный контроль качества [2].

В §2.2 приводятся точные оценки для (1.5) и (1.6).

Задача 1.3′ (оценивание частоты события на генеральной выборке). Требуется предсказать частоту события $S \subseteq \mathbb{X}$ на полной выборке $\mathbb{X}$ по его частоте на наблюдаемой выборке $X \subset \mathbb{X}$ и оценить надёжность предсказания:

$$\mathsf{P}\big[|\nu(\mathbb{X}) - \nu(X)| \geqslant \varepsilon\big] \leqslant \eta(\varepsilon); \tag{1.7}$$

$$\mathsf{P}\big[\ \nu(\mathbb{X}) - \nu(X)\ \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.8}$$

Задача 1.3′ эквивалентна Задаче 1.3 в силу соотношения $\nu(\mathbb{X}) = \frac{\ell}{L}\nu(X) + \frac{k}{L}\nu(\bar{X})$, из которого следует

$$\nu(\mathbb{X}) - \nu(X) = \tfrac{k}{L}\big(\nu(\bar{X}) - \nu(X)\big).$$

Задача 1.4 (построение доверительных интервалов). Пусть задана функция $\xi\colon \mathbb{X} \to \mathbb{R}$. Требуется построить по наблюдаемой выборке X семейство вложенных *доверительных интервалов* $\Omega_\varepsilon(X) = \big[\xi_\varepsilon^-(X), \xi_\varepsilon^+(X)\big]$, такое, что значение $\xi(\bar{x})$ на скрытом объекте $\bar{x}$ попадает в $\Omega_\varepsilon(X)$ с вероятностью не менее $1-\eta(\varepsilon)$:

$$\mathsf{P}\big[\xi(\bar{x}) \notin \Omega_\varepsilon(X)\big] \leqslant \eta(\varepsilon).$$

Очевидно, данная задача есть $\mathscr{P}_2\big\langle \mathbb{R},\ \xi(\bar{x}),\ \Omega_\varepsilon(X)\big\rangle$ при единичной длине контрольной выборки, $k = |\bar{X}| = 1$.

В §4.1 приводятся точные оценки для данной задачи.

Задача 1.5 (оценивание функции распределения). Определим для произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$ и произвольной конечной выборки $U \subseteq \mathbb{X}$ *эмпирическую функцию распределения* $F_\xi\colon \mathbb{R} \to [0,1]$. Она показывает, на какой доле объектов выборки значение $\xi(x)$ не превосходит z:

$$F_\xi(z, U) = \frac{1}{|U|}\sum_{x\in U} [\xi(x) \leqslant z].$$

Требуется предсказать максимальное отклонение функции распределения на скрытой выборке $F_\xi(z,\bar{X})$ от известной функции распределения на наблюдаемой выборке $F_\xi(z,X)$ и оценить надёжность предсказания:

$$\mathsf{P}\Big[\max_{z\in\mathbb{R}}|F_\xi(z,\bar{X})-F_\xi(z,X)|>\varepsilon\Big]\leqslant\eta(\varepsilon). \tag{1.9}$$

Данная задача является частным случаем $\mathscr{P}_1$, если в качестве R взять множество всех неубывающих кусочно-постоянных функций, $R=\big\{F\colon\mathbb{R}\to[0,1]\big\}$, ввести на R равномерную (чебышевскую) метрику $d(\hat{t},t)=\max\limits_{z\in\mathbb{R}}|t(z)-\hat{t}(z)|$, и положить $T(\bar{X})(z)=F_\xi(z,\bar{X})$, $\hat{T}(X)(z)=F_\xi(z,X)$, где $z\in\mathbb{R}$.

Данная задача тесно связана с оцениванием скорости сходимости эмпирических распределений и имеет фундаментальное значение для математической статистики. На оценке (1.9) основан критерий Смирнова проверки гипотезы однородности (одинаковой распределённости) двух выборок [15, 3].

В §3.2 приводятся точные оценки для (1.9) и односторонних неравенств.

Задача 1.6 (обучение по прецедентам). Задано множество $\mathbb{A}$, элементы которого называются *алгоритмами*. Существует бинарная функция $I\colon\mathbb{A}\times\mathbb{X}\to\{0,1\}$, называемая *индикатором ошибки*. Если $I(a,x)=1$, то говорят, что алгоритм a допускает ошибку на объекте x.

Частотой ошибок алгоритма a на выборке $U\subseteq\mathbb{X}$ называется величина

$$\nu(a,U)=\frac{1}{|U|}\sum_{x\in U}I(a,x).$$

Методом обучения называется отображение $\mu\colon\mathbb{X}^\ell\to\mathbb{A}$, которое произвольной *обучающей выборке* $X\subset\mathbb{X}$ ставит в соответствие некоторый алгоритм $a=\mu X$ из $\mathbb{A}$.

Переобученностью метода μ на паре выборок X и $\bar{X}$ называется уклонение частоты ошибок алгоритма $a=\mu X$ на скрытой *контрольной* выборке $\bar{X}$ от частоты его ошибок на наблюдаемой *обучающей* выборке X:

$$\delta_\mu(X,\bar{X})=\nu(\mu X,\bar{X})-\nu(\mu X,X).$$

Требуется предсказать верхнюю границу переобученности и оценить надёжность предсказания:

$$\mathsf{P}\big[\delta_\mu(X,\bar{X})\geqslant\varepsilon\big]=\mathsf{P}\big[\nu(\mu X,\bar{X})-\nu(\mu X,X)\geqslant\varepsilon\big]\leqslant\eta(\varepsilon). \tag{1.10}$$

Очевидно, данная задача есть $\mathscr{P}_1\big\langle\mathbb{R},\ (t-\hat{t}),\ \nu(\mu X,\bar{X}),\ \nu(\mu X,X)\big\rangle$.

Переобучение является серьёзной проблемой при построении алгоритмов классификации и регрессии по конечным выборкам данных [6]. Исследованию задач обучения по прецедентам и проблеме переобучения посвящена основная часть данной диссертационной работы, начиная с главы **??**.

§1.2 Обращение оценок

Пусть для функции $\varphi\colon \mathbb{X}^\ell \times \mathbb{X}^k \to \mathbb{R}$ найдена оценка вида

$$\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] \leqslant \eta(\varepsilon), \quad \varepsilon \in \mathbb{R}. \tag{1.11}$$

Задача *обращения оценки* заключается в том, чтобы по функции $\eta(\varepsilon)$ построить функцию $\varepsilon(\eta)$, при которой (1.11) переходит в эквивалентное утверждение:

$$\varphi(X,\bar{X}) \leqslant \varepsilon(\eta) \;\text{ с вероятностью, не меньшей }\; 1-\eta, \tag{1.12}$$

где η — произвольное число из $[0,1]$. Предполагается, что η достаточно близко к нулю.

Рассмотрим несколько возможных случаев.

Непрерывная верхняя оценка. Пусть для функции φ найдена оценка вида $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] \leqslant \tilde{\eta}(\varepsilon)$, где $\tilde{\eta}(\varepsilon)$ — непрерывная строго убывающая функция.

Тогда существует *обратная* к ней функция $\varepsilon(\eta)$ такая, что $\tilde{\eta}(\varepsilon(\eta)) \equiv \eta$, и для любого $\eta \in [0,1]$ выполняется $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon(\eta)\big] \leqslant \tilde{\eta}(\varepsilon(\eta)) = \eta$. Таким образом, (1.12) выполняется, если в качестве $\varepsilon(\eta)$ взять обычную обратную функцию.

Заметим, что функция $\tilde{\eta}(\varepsilon)$ не может быть точной оценкой, иначе она была бы кусочно-постоянной и не могла бы быть строго убывающей.

Точная оценка, полунепрерывная справа. Допустим теперь, что для функции φ найдена точная оценка $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] = \eta(\varepsilon)$. Тогда функция $\eta(\varepsilon)$ монотонно не возрастает, кусочно-постоянна, полунепрерывна справа и принимает конечное множество значений $H = \big\{\eta(\varepsilon)\colon \varepsilon \in \mathbb{R}\big\}$, см. Рис. 1. Обратная к ней $\varepsilon(\eta)$ определена только при $\eta \in H$, но её можно доопределить при любом $\eta' \in \mathbb{R}$, причём двумя способами:

$$\varepsilon(\eta') = \min\{\varepsilon\colon \eta(\varepsilon) \leqslant \eta'\} = \sup\{\varepsilon\colon \eta(\varepsilon) > \eta'\}. \tag{1.13}$$

Функция $\varepsilon(\eta')$ также монотонно не возрастает, кусочно-постоянна, полунепрерывна справа. Следующая оценка справедлива при любом $\eta' \in \mathbb{R}$, обращаясь в равенство при $\eta' \in H$:

$$\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon(\eta')\big] = \eta\big(\varepsilon(\eta')\big) = \eta\big(\min\{\varepsilon\colon \eta(\varepsilon) \leqslant \eta'\}\big) \leqslant \eta'.$$

Таким образом, если обратную функцию доопределять согласно (1.13), то выполняется (1.12).

Точная оценка, полунепрерывная слева. Пусть теперь $\mathsf{P}\big[\varphi(X,\bar{X}) \geqslant \varepsilon\big] = \eta(\varepsilon)$. Тогда функция $\eta(\varepsilon)$ монотонно не возрастает, кусочно-постоянна, полунепрерывна слева и принимает конечное множество значений $H = \big\{\eta(\varepsilon)\colon \varepsilon \in \mathbb{R}\big\}$, см. Рис. 2. Обратная к ней доопределяется одним из двух способов:

$$\varepsilon(\eta') = \max\{\varepsilon\colon \eta(\varepsilon) \geqslant \eta'\} = \inf\{\varepsilon\colon \eta(\varepsilon) < \eta'\}. \tag{1.14}$$

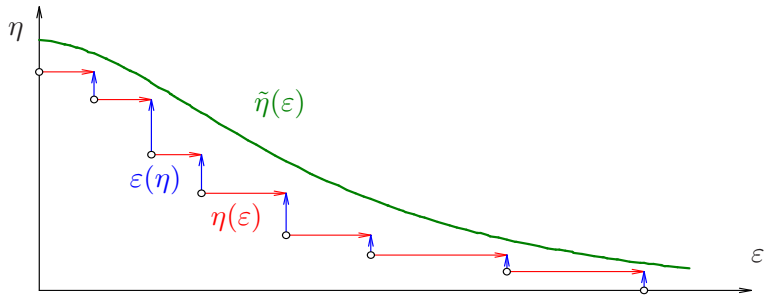

Рис. 1. Точная оценка, полунепрерывная *справа*, $\eta(\varepsilon)$ — красные горизонтальные линии, её обратная $\varepsilon(\eta)$ — синие вертикальные линии. Обе функции монотонно не возрастают, кусочно-постоянны, полунепрерывны *справа*. Строго убывающая непрерывная функция $\tilde{\eta}(\varepsilon)$ является завышенной верхней оценкой.

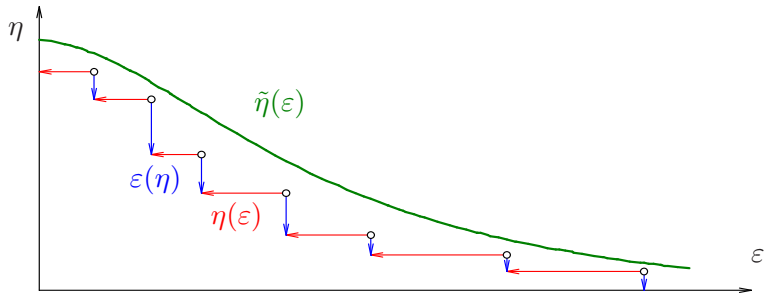

Рис. 2. Точная оценка, полунепрерывная *слева*, $\eta(\varepsilon)$ — красные горизонтальные линии, её обратная $\varepsilon(\eta)$ — синие вертикальные линии. Обе функции монотонно не возрастают, кусочно-постоянны, полунепрерывны *слева*. Строго убывающая непрерывная функция $\tilde{\eta}(\varepsilon)$ является завышенной верхней оценкой.

Функция $\varepsilon(\eta')$ также монотонно не возрастает, кусочно-постоянна, полунепрерывна слева. При любом $\eta' \in \mathbb{R}$ справедлива оценка:

$$\begin{aligned}\mathsf{P}\big[\varphi(X,\bar{X}) \geqslant \varepsilon(\eta')\big] &= \eta\big(\varepsilon(\eta')\big) = \eta\big(\max\{\varepsilon\colon \eta(\varepsilon) \geqslant \eta'\}\big) = \\ &= \begin{cases}\eta', & \eta' \in H;\\ \operatorname{next}_H(\eta'), & \eta' \notin H;\end{cases} \\ &\leqslant \operatorname{next}_H(\eta'),\end{aligned}$$

где $\operatorname{next}_H(\eta') = \min\big\{\eta \in H\colon \eta > \eta'\big\}$ — элемент множества H, следующий за η'.

Таким образом, если обратную функцию доопределять согласно (1.14), то вместо (1.12) выполняется

$\varphi(X,\bar{X}) < \varepsilon(\eta)$ с вероятностью, не меньшей $1 - \operatorname{next}_H(\eta)$.

Замечание 1.1. Из-за дополнительной операции next_H оценки, полунепрерывные слева, менее удобны, чем полунепрерывные справа. При больших длинах выборки L разностью $\operatorname{next}_H(\eta) - \eta$ можно пренебрегать как несущественной добавкой к уровню значимости. Однако при малых выборках она может быть существенной.

Замечание 1.2. Оценка, полунепрерывная справа, $\mathsf{P}\big[\varphi(X,\bar{X}) > \varepsilon\big] = \eta(\varepsilon)$, позволяет записать функцию распределения величины φ на выборке $\mathbb{X}$:

$$F_\varphi(\varepsilon) = \mathsf{P}\big[\varphi(X,\bar{X}) \leqslant \varepsilon\big] = 1 - \eta(\varepsilon).$$

Значение обратной функции $\varepsilon(\eta)$ называется *квантúлью порядка* $(1-\eta)$ для распределения величины φ.

Замечание 1.3. Непрерывные оценочные функции удобны тем, что их обращение, как правило, удаётся выполнить аналитически. Кусочно-постоянные функции задаются последовательностями точек, поэтому их обращение является вычислительной процедурой. Если мощность множества $\Phi = \{\varphi(X,\bar{X})\colon X \in [\mathbb{X}]^\ell\}$ не велика, то кусочно-постоянные функции $\eta(\varepsilon)$ и $\varepsilon(\eta)$ вычисляются довольно эффективно. Для этого множество Φ упорядочивается по возрастанию, и поиск минимального или максимального значения ε в (1.13) и (1.14) производится методом половинного деления, за $O(\log|\Phi|)$ операций. Дальнейшее повышение эффективности возможно за счёт вычисления только той относительной небольшой части множества Φ, которая соответствует достаточно малым значениям ε.

Доверительные интервалы для предсказываемой величины $T(\bar{X},X)$. Пусть в задаче $\mathscr{P}_1$ функция отклонения имеет вид $d(\hat{t},t) = t - \hat{t}$, $R = \mathbb{R}$ и имеется верхняя оценка

$$\mathsf{P}\big[T(\bar{X},X) - \hat{T}(X) > \varepsilon\big] \leqslant \eta(\varepsilon). \tag{1.15}$$

Тогда справедлива верхняя оценка (доверительный полуинтервал) для предсказываемой величины $T(\bar{X},X)$: для любого $\eta \in [0,1]$ с вероятностью не менее $1-\eta$

$$T(\bar{X},X) \leqslant \hat{T}(X) + \varepsilon(\eta),$$

где $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon)$.

При функции отклонения вида $d(\hat{t},t) = |t - \hat{t}|$ аналогичным образом выписывается двусторонняя оценка (доверительный интервал) для предсказываемой величины $T(\bar{X},X)$: для любого $\eta \in [0,1]$ с вероятностью не менее $1-\eta$

$$\hat{T}(X) - \varepsilon(\eta) \leqslant T(\bar{X},X) \leqslant \hat{T}(X) + \varepsilon(\eta).$$

Подчеркнём, что эти оценки справедливы не для всех разбиений, а только для достаточно большой их доли, то есть с вероятностью, близкой к 1. Задача оценивания надёжности эмпирических предсказаний как раз и состоит в том, чтобы находить условия, при которых значения η и ε одновременно достаточно малы.

§1.3 Наблюдаемые и ненаблюдаемые оценки

При получении оценок вида (1.15) часто оказывается, что точная оценочная функция существенно зависит от всей выборки, $\eta(\varepsilon) = \eta(\varepsilon,\mathbb{X})$. Такие оценки называются *ненаблюдаемыми* (unobservable bound) [24, 23]. Их невозможно непосредственно

использовать в задачах эмпирического предсказания, так как скрытая часть выборки в момент предсказания неизвестна.

Ниже рассматривается достаточно общий технический приём, позволяющий переходить от ненаблюдаемых оценок к *наблюдаемым* (observable bound).

Переход от ненаблюдаемой оценки к наблюдаемой. Рассмотрим случай, когда оценочная функция зависит от некоторой статистики $z(\mathbb{X})$ генеральной выборки: $\eta(\varepsilon, \mathbb{X}) = \eta(\varepsilon, z(\mathbb{X}))$. Функцию $z(\mathbb{X})$ невозможно вычислить, не зная скрытой части выборки. Возможны несколько путей решения этой проблемы.

1. Заменить точное выражение $\eta(\varepsilon, z(\mathbb{X}))$ его верхней оценкой $\max\limits_{z} \eta(\varepsilon, z)$, справедливой для любой выборки $\mathbb{X}$. Это «оценка худшего случая», и она может оказаться сильно завышенной.

2. Более тонкие результаты даёт оценивание значения $z(\mathbb{X})$ по значению этой же статистики z на наблюдаемой выборке $z(X)$. Допустим, в Задаче 1.1 получена верхняя оценка $\mathsf{P}\big[T - \hat{T} > \varepsilon\big] \leqslant \eta_T(\varepsilon, z(\mathbb{X}))$, и функция $\varepsilon_T(\eta, z)$ является обратной для $\eta_T(\varepsilon, z)$ при любом значении параметра z. Тогда с надёжностью не менее $1 - \eta_1$ справедлива верхняя оценка

$$T \leqslant \hat{T} + \varepsilon_T\big(\eta_1, z(\mathbb{X})\big).$$

Вторым шагом необходимо получить оценку для $z(\mathbb{X})$ через $z(X)$. Допустим, что функция $\varepsilon_T(\eta, z)$ монотонно не убывает по второму аргументу z, и что для z имеется оценка сверху:

$$\mathsf{P}\big[z(\mathbb{X}) - z(X) > \varepsilon\big] \leqslant \eta_z(\varepsilon).$$

Тогда с надёжностью не менее $1 - \eta_2$ справедлива оценка сверху для $z(\mathbb{X})$:

$$z(\mathbb{X}) \leqslant z(X) + \varepsilon_z(\eta_2),$$

где $\varepsilon_z(\eta_2)$ — функция, обратная для $\eta_z(\varepsilon)$. В итоге, с надёжностью не менее $1 - \eta_1 - \eta_2$ справедлива оценка сверху для T:

$$T \leqslant \hat{T} + \varepsilon_T\big(\eta_1, z(X) + \varepsilon_z(\eta_2)\big).$$

Теперь правая часть зависит только от наблюдаемой выборки. Её можно минимизировать по параметрам η_1 и η_2 при заданном $\eta = \eta_1 + \eta_2$, или для простоты положить $\eta_1 = \eta_2 = \eta/2$.

Замечание 1.4. Если статистика z выражается через статистику T, то можно обойтись без введения второго параметра η_2. Такой случай будет рассмотрен в §2.3.

Связь задач эмпирического предсказания и проверки гипотезы однородности. В прикладной статистике разработано большое количество критериев для выявления значимых отличий между двумя выборками X^ℓ и X^k, а в общем случае — между несколькими выборками [10].

В рамках сильной аксиоматики выдвигается *нулевая гипотеза* об однородности выборок: «две выборки X^ℓ и X^k случайны, независимы и получены из одного распределения». Затем для заданной *статистики* $\varphi(X^\ell, X^k)$ при условии истинности нулевой гипотезы выводится функция распределения $F_\varphi(\varepsilon) = \mathsf{P}_{X^\ell,X^k}\big[\varphi(X^\ell, X^k) \leqslant \varepsilon\big]$.

Нулевая гипотеза отвергается, если оказывается, что наблюдаемое значение $\varphi(X^\ell, X^k)$ попадает в *критическую область* маловероятных значений статистики φ — как правило, слишком больших или слишком малых.

В слабой аксиоматике проверка гипотез осуществляется аналогичным образом, с той лишь разницей, что высказывание «выборки получены из одного распределения» теперь некорректно. *Гипотеза однородности* формулируется так: «выборки X^ℓ и X^k получены в результате реализации одного из $C^\ell_{\ell+k}$ равновероятных разбиений генеральной выборки $\mathbb{X}$». При условии истинности нулевой гипотезы выводится функция распределения $F_\varphi(\varepsilon) = \mathsf{P}\big[\varphi(X^\ell, X^k) \leqslant \varepsilon\big]$. Нулевая гипотеза отвергается на *уровне значимости* η, если для заданной пары выборок (X^ℓ, X^k) выполняется условие $\varphi(X^\ell, X^k) > \varepsilon(\eta)$, где $\varepsilon(\eta)$ — квантиль порядка $1-\eta$ для распределения $F_\varphi(\varepsilon)$.

Нетрудно заметить, что одни и те же функции распределения $F_\varphi(\varepsilon)$ могут быть использованы как для оценивания надёжности эмпирических предсказаний, так и для проверки однородности. В частности, оценка (1.5) может быть использована для проверки однородности при произвольном заранее заданном событии S.

Существует принципиальное отличие между этими двумя важными классами задач анализа данных. При проверке однородности нет скрытой выборки; обе выборки являются наблюдаемыми. Поэтому оценочная функция $\eta(\varepsilon)$ имеет право зависеть (и, как правило, зависит) от всех объектов из $\mathbb{X}$, что не вызывает никаких затруднений. В задачах эмпирического предсказания ситуация существенно сложнее — оценочная функция не должна зависеть от скрытой части выборки. Это требует дополнительных усилий по переходу от ненаблюдаемой оценки к наблюдаемой, который может сопровождаться как вычислительными затратами, так и некоторой потерей точности оценки.

§1.4 Эмпирическое оценивание вероятности

В слабой аксиоматике вероятность определяется как «доля *всех* разбиений выборки», и потому может быть легко измерена эмпирически как «доля разбиений из *выбранного подмножества* разбиений». Если выбор разбиений осуществляется случайным образом, то можно говорить о применении метода Монте-Карло для эмпирического оценивания вероятности.

Пусть задан предикат $\varphi\colon \mathbb{X}^\ell \times \mathbb{X}^k \to \{0,1\}$. Вероятность $Q = \mathsf{P}\,\varphi(X, \bar{X})$ есть среднее значение φ по множеству всех разбиений N:

$$Q = \frac{1}{C^\ell_L} \sum_{(X,\bar{X})\in N} \varphi(X, \bar{X}).$$

Непосредственное вычисление величины Q по этой формуле практически осуществимо только при небольших значениях ℓ и k. В типичных случаях число разбиений

$|N| = C_L^\ell$ огромно. Рассмотрим приближённую оценку Q как среднее по некоторому подмножеству разбиений $N' \subset N$, не слишком большому, чтобы сумма вычислялась за приемлемое время:

$$\hat{Q}_{N'} = \frac{1}{|N'|} \sum_{(X,\bar{X})\in N'} \varphi(X, \bar{X}).$$

Метод Монте-Карло. Выберем случайное подмножество разбиений $N' \subset N$ из равномерного распределения на множестве всех $C_{|N|}^{|N'|}$ подмножеств мощности $|N'|$. Тогда оценивание величины $\hat{Q}_{N'}$ сводится к Задаче 1.3′ об оценивании частоты события на генеральной выборке, только теперь в качестве генеральной выборки рассматривается множество всех разбиений N.

В разделе 2 для функционала (1.8) в Задаче 1.3′ будет получена точная оценка. Сейчас предположим, что имеется верхняя оценка

$$\mathsf{P}_{N'}\{Q - \hat{Q}_{N'} > \tilde{\varepsilon}\} \leqslant \tilde{\eta}(\tilde{\varepsilon}).$$

Обращая эту оценку, получаем, что с надёжностью не менее $1 - \tilde{\eta}$ выполняется неравенство $Q_N \leqslant \hat{Q}_{N'} + \tilde{\varepsilon}(\tilde{\eta})$, где $\tilde{\varepsilon}(\tilde{\eta})$ — обратная функция для $\tilde{\eta}(\tilde{\varepsilon})$.

Конкретизируем вид предиката $\varphi(X, \bar{X})$. Рассмотрим задачу эмпирического предсказания величины $T(\bar{X}, X)$, в которой $\varphi(X, \bar{X}) = [T(\bar{X}, X) - \hat{T}(X) > \varepsilon]$. Тогда величины $Q(\varepsilon)$ и $\hat{Q}_{N'}(\varepsilon)$ являются кусочно-постоянными невозрастающими функциями параметра ε. С надёжностью не менее $1 - \tilde{\eta}$ выполняется неравенство

$$\mathsf{P}\big[T - \hat{T} > \varepsilon\big] = Q(\varepsilon) \leqslant \hat{Q}_{N'}(\varepsilon) + \tilde{\varepsilon}(\tilde{\eta}).$$

Ещё раз применяя обращение, заключаем, что с надёжностью не менее $(1 - \tilde{\eta} - \eta)$ справедлива верхняя оценка для $T(\bar{X}, X)$:

$$T(\bar{X}, X) \leqslant \hat{T}(X) + \varepsilon(\eta), \tag{1.16}$$

где $\varepsilon(\eta)$ — обратная функция для $\eta(\varepsilon) = \hat{Q}_{N'}(\varepsilon) + \tilde{\varepsilon}(\tilde{\eta})$.

Таким образом, вычисляя значения $T(\bar{X}, X)$ и $\hat{T}(X)$ по небольшому подмножеству разбиений, $(\bar{X}, X) \in N'$, можно получить верхнюю границу, которую $T(\bar{X}, X)$ не превосходит для любого разбиения $(\bar{X}, X) \in N$, с заданной надёжностью.

Описанный способ эмпирического оценивания вероятности имеет несколько существенных недостатков. Во-первых, он даёт лишь приближённые оценки. Во-вторых, он требует знания генеральной выборки $\mathbb{X}$, и потому не может быть использован непосредственно для эмпирического предсказания. В-третьих, он не позволяет получать оценки в аналитическом виде. Наконец, он может потребовать большого объёма вычислений.

Таким образом, область применимости эмпирического оценивания довольно ограничена. На практике его можно использовать для предварительного экспериментального исследования зависимости Q от некоторых параметров задачи (например, от длины выборки).

В задачах обучения по прецедентам эмпирическое оценивание называют *скользящим контролем* (cross-validation) и используют для оценивания качества метода обучения μ, а не отдельного алгоритма. Скользящий контроль незаменим в тех случаях, когда теоретические верхние оценки вероятности Q не известны или сильно завышены. В главе **??** данной работы эмпирическое оценивание применяется для анализа точности теоретических оценок и понимания причин их завышенности.

§1.5 Замечания и интерпретации

О связи с сильной вероятностной аксиоматикой. Классическая теоретико-мерная аксиоматика А. Н. Колмогорова (будем называть её сильной) основана на понятии вероятностного пространства $\langle \mathscr{X}, \Omega, \mathsf{P} \rangle$, где $\mathscr{X}$ — множество допустимых объектов, Ω — аддитивная σ-алгебра событий на $\mathscr{X}$, P — вероятностная мера, определённая на событиях из Ω. Во многих задачах статистического анализа данных предполагается, что *данные* — это конечное множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, выбранных из множества $\mathscr{X}$ случайно и независимо согласно вероятностной мере P. В реальных приложениях множество $\mathscr{X}$, как правило, бесконечно, а мера P — неизвестна.

В слабой аксиоматике множество $\mathscr{X}$ не вводится. Рассматривается только конечное множество объектов — *генеральная выборка* $\mathbb{X}$. Оно может включать в себя как объекты, наблюдавшиеся ранее, так и скрытые объекты, которые станут известны в будущем. Вероятностным пространством является конечное множество всех перестановок генеральной выборки $\mathbb{X}$, на котором задаётся равномерное распределение. Таким образом, случайными полагаются не сами объекты, а лишь порядок их появления, что соответствует предположению о независимости объектов выборки в сильной аксиоматике.

Переход от слабой аксиоматики к сильной тривиален. Допустим, что в слабой аксиоматике найдено значение вероятности

$$\mathsf{P}_\tau \, \psi(\tau \mathbb{X}) = f(\mathbb{X}). \tag{1.17}$$

Введём вероятностное пространство $\langle \mathscr{X}, \Omega, \mathsf{P} \rangle$ и примем гипотезу о независимости наблюдений в выборке $\mathbb{X}$. Тогда для любой перестановки τ выполняется $\mathsf{P}_{\mathbb{X}} \, \psi(\mathbb{X}) = \mathsf{P}_{\mathbb{X}} \, \psi(\tau \mathbb{X})$. Следовательно, перенос оценки $f(\mathbb{X})$ в сильную аксиоматику сводится к применению операции математического ожидания $\mathsf{E}_{\mathbb{X}}$ по выборке $\mathbb{X}$ к обеим частям равенства (1.17):

$$\mathsf{P}_{\mathbb{X}} \, \psi(\mathbb{X}) = \mathsf{E}_{\mathbb{X}} \, \mathsf{P}_\tau \, \psi(\tau \mathbb{X}) = \mathsf{E}_{\mathbb{X}} f(\mathbb{X}). \tag{1.18}$$

В случаях, когда оценка $f(\mathbb{X})$ не зависит от выборки $\mathbb{X}$, конечный результат — оценка в правой части (1.17) и (1.18) — будет одинаков в обеих аксиоматиках.

Если же оценка имеет вид $f(\mathbb{X}) = f(S(\mathbb{X}))$, где S — некоторая функция (статистика) полной выборки, то возможны несколько вариантов дальнейших действий.

1. *«What-if анализ»*: значение статистики $S = S(\mathbb{X})$ интерпретируется как априори задаваемый параметр, и окончательный результат представляется в виде зависимости оценки f от S.

2. Строится *оценка худшего случая* $f(\mathbb{X}) \leqslant \max\limits_{S} f(S)$, которая не зависит от выборки, но может оказаться сильно завышенной.

3. Строится *доверительный интервал* для ненаблюдаемого значения статистики $S(\mathbb{X})$ по значению той же статистики на наблюдаемой выборке $S(X)$, см. §1.3. Затем доверительный интервал для S переводится в доверительный интервал для $f(S)$.

Во всех перечисленных случаях вид оценки не меняется при переходе от слабой аксиоматики к сильной. Фактически, этот переход связан только с заменой функционала и его неформальных интерпретаций, но он никак не влияет на полученную оценку. Поэтому далее этот переход будет опускаться, и все результаты будут формулироваться в рамках слабой аксиоматики.

О частотных подходах в основаниях теории вероятностей. Предлагаемая в данной работе слабая вероятностная аксиоматика отличается не только от теоретико-мерной аксиоматики Колмогорова (являясь её специальным частным случаем), но и от известных частотных подходов к определению понятия вероятности.

Частотный подход фон Мизеса [27] является инфинитарным. Его основная цель — определение фундаментального понятия «вероятности» как предела частоты. Его основная проблема, которую до сих пор не удаётся разрешить до конца — необходимость строгой формализации понятия *иррегулярной* (т. е. бесконечной случайной) последовательности [17, 16]. В отличие от подхода фон Мизеса, в слабой аксиоматике рассматриваются только конечные последовательности.

А. Н. Колмогоров был убеждён, что «частотный подход, основанный на понятии *предельной частоты* при стремящемся к бесконечности числе испытаний, не позволяет обосновать применимость результатов теории вероятностей к практическим задачам, в которых мы имеем дело с конечным числом испытаний» [12, стр. 205]. Начиная с 1965 г., Колмогоров развивал *финитарную теорию алгоритмической случайности* [11]. В этом подходе конечная последовательность считается случайной, если длина её кратчайшего описания, называемая также *колмогоровской сложностью*, не сильно отличается от максимально возможного значения, равного $\log_2 |A|$. Здесь A — это конечное множество всевозможных последовательностей. Например, $|A| = C_L^k$ для бернуллиевских последовательностей — двоичных последовательностей длины L, состоящих из k единиц и $L-k$ нулей. Предполагается, что при разумных определениях множества A доля простых (значит, неслучайных) последовательностей крайне мала, и подавляющее большинство последовательностей являются сложными (значит, случайными). Затем строится *сложностная модель* [5]: на конечном множестве всех случайных последовательностей $A' \subset A$ вводится некоторое естественное распределение вероятностей, как правило, равномерное. В ряде важных частных случаев сложностные модели и стандартные *статистические модели* приводят к одинаковым результатам, см. ссылки на работы Е. А. Асарина в [5]. С помощью сложностных моделей возможно делать эмпирические предсказания, оценивать доверительные интервалы, проверять статистические гипотезы, и при этом вообще не вводить понятие вероятности.

Слабая аксиоматика близка к финитарной теории Колмогорова. Это становится очевидно, если заметить, что существует взаимнооднозначное соответствие между множеством всех C_L^k разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ и бернуллиевской последовательностью $A = (a_1, \ldots, a_L)$, где $a_i = [x_i \in \bar{X}]$. Однако имеется и принципиальное отличие: в слабой аксиоматике равномерное распределение вводится не на множестве случайных последовательностей A', а на множестве всех последовательностей A. Таким образом, в нашей модели неслучайное по Колмогорову разбиение имеет точно такие же шансы реализоваться, как и случайное. Если доля неслучайных разбиений крайне мала, то шансы получить какое-либо из неслучайных разбиений также крайне малы, хотя и не равны нулю. В этом случае количественные результаты, полученные в слабой аксиоматике и в финитарной теории Колмогорова, должны быть очень близки.

Вопросы о том, давать ли неслучайным разбиениям равные шансы или нулевые, а также, при какой величине *дефекта случайности* [5] разбиение должно считаться неслучайным, разрешаются априори и находятся за пределами математики.

В отличие от подходов фон Мизеса и Колмогорова, в слабой аксиоматике не предпринимается никаких попыток определить понятие случайной последовательности. Проблема в том, что все известные критерии случайности неконструктивны и слишком тяжелы для практической реализации. Например, длину кратчайшего описания возможно оценить только приближённо, применяя какой-либо алгоритм сжатия данных. Простой способ обойти эту проблему — считать, что почти все разбиения выборки, за исключением малой доли η, случайны. При этом ничего не утверждается о том, *какие именно* разбиения неслучайны. Априори задаётся только величина η — *уровень значимости* или *надёжность*, с которой будут справедливы статистические выводы. Если доля неслучайных разбиений $\frac{|A \setminus A'|}{|A|}$ мала, то выводы, справедливые для большинства разбиений из A, будут также справедливы и для большинства случайных (по Колмогорову) разбиений из A'. Качество выводов на оставшейся малой доле разбиений может быть сколь угодно плохим, однако *именно эти разбиения можно считать неслучайными для данной задачи.*

Таким образом, слабая аксиоматика фактически представляет собой компромиссное упрощение финитарной теории Колмогорова, не требующее оценивания колмогоровских сложностей и дефектов случайности, и тем самым лучше приспособленное для непосредственного практического применения в задачах анализа данных.

Об уровнях значимости. Обычно уровень значимости устанавливается экспертом, исходя из субъективных представлений о том, какой должна быть «вероятность маловероятного события». При такой интерпретации возникает искушение установить уровень значимости поменьше, особенно в тех приложениях, где требуется высокая надёжность. В слабой аксиоматике понимание, почему этого не стоит делать, возникает благодаря естественной интерпретация уровня значимости как *доли разбиений (в общем случае — перестановок) выборки, которые нельзя считать случайными.*

В Задаче 1.3 о предсказании частоты события S неслучайными являются, в частности, те разбиения, при формировании которых используется информация о самом

событии S. Например, когда в скрытую выборку преднамеренно включаются только элементы из S. Очевидно, это приводит к нарушению основной Аксиомы 1.1, поскольку разбиения перестают быть равновероятными.

В Задаче 1.6 (обучения по прецедентам) неслучайными можно считать разбиения, при формировании которых используется информация об алгоритмах. Например, фиксируется некоторый алгоритм $a_0 \in \mathbb{A}$, и в скрытую (контрольную) выборку преднамеренно включаются все объекты, на которых данный алгоритм $a_0 \in \mathbb{A}$ допускает ошибку. Другой пример: пусть $\mathbb{X}$ — это точки линейного векторного пространства, заранее разделённого некоторой гиперплоскостью на два полупространства; в наблюдаемую (обучающую) выборку преднамеренно включаются только объекты из первого полупространства, а в скрытую — только из второго. В подобных случаях алгоритм, лучший на обучающей выборке, может оказаться сколь угодно плохим на контрольной выборке в силу того, что две выборки, взятые из разных областей пространства, могут практически не нести информации друг о друге.

Заметим, что в примере с гиперплоскостью доля неслучайных разбиений возрастает с ростом размерности пространства $\mathbb{X}$. Стало быть, понятие «неслучайного разбиения» может зависеть от особенностей конкретной задачи. Возможно, именно по этой причине в частотной теории фон Мизеса так и не удалось выработать окончательного определения иррегулярной последовательности.

В общем случае неизвестно, какие именно разбиения являются неслучайными, однако можно считать, что их доля невелика, и априори оценивать её уровнем значимости η. Можно следовать обычной практике и назначать уровень значимости эвристически, в частности, полагать $\eta = 0.05$. Более обоснованно уровень значимости можно было бы определять путём комбинаторного подсчёта доли разбиений выборки, которые в данной конкретной задаче нельзя считать случайными. Однако при попытке практической реализации этой идеи возникают те же трудности, что и в подходах фон Мизеса и Колмогорова.

Нет никакого смысла устанавливать уровень значимости существенно ниже доли неслучайных разбиений. Искусственное занижение уровня значимости η понижает точность ε эмпирических предсказаний. Невозможно заставить функцию $\hat{T}(X)$ давать хорошие предсказания в тех ситуациях, когда разбиение не случайно (преднамеренно) подобрано так, чтобы хорошее предсказание было невозможно.

О трансдукции. Предсказание некоторого свойства выборки на основании свойств другой выборки называется *трансдукцией* или переходом от частного к частному. Принято считать, что трансдукция более примитивна и ограничена, чем *индукция* — переход от частного к общему. В нашем случае это не совсем так. Допустим, в правой части (1.4) удалось получить оценку $\eta(\varepsilon)$, не зависящую от генеральной выборки $\mathbb{X}$. Тогда с помощью техники обращения выводится оценка и для $T(\bar{X}, X)$. Поскольку она будет справедлива *для любой* скрытой выборки $\bar{X}$, трансдукция в данном случае становится не менее общей, чем индукция.

Заметим, что в машинном обучении под *трансдуктивным обучением* (transductive learning) имеется в виду совсем другая постановка задачи — там это специальный

тип задач классификации, в которых объекты контрольной выборки $\bar{X}$ известны ещё до решения задачи, а неизвестными являются только ответы на этих объектах.

О скользящем контроле. В Задаче 1.6 (обучения по прецедентам) эмпирическое оценивание $\hat{Q}_{N'} \approx Q$ по подмножеству разбиений N' принято называть *скользящим контролем* или *кросс-проверкой* (cross-validation, CV). В зависимости от способа формирования подмножества разбиений N' различают несколько разновидностей скользящего контроля [22].

Если берётся одно случайное разбиение, $|N'| = 1$, говорят об *оценке на отдельной тестовой выборке* (hold-out estimate).

Если берутся все разбиения при длине контрольной выборки $k = 1$, то говорят об *оценке с одним отделяемым объектом* (leave-one-out estimate, LOO).

Если используются все разбиения с контрольной выборкой фиксированной, но не обязательно единичной, длины, то говорят об *оценке полного скользящего контроля* (complete cross-validation) [26].

Если производится случайная независимая выборка L объектов из $\mathbb{X}$ с возвращениями при фиксированной длине контроля k, то говорят о *бутстреп-оценке* (bootstrap estimate) [18].

Если множество разбиений N' образуется q непересекающимися контрольными выборками, в объединении дающими генеральную выборку $\mathbb{X}$, то говорят о *q-кратном скользящем контроле* (q-fold cross-validation).

Многие практические методики скользящего контроля стремятся уменьшить дисперсию оценки и вычислительные затраты, выбирая разбиения «более равномерно», а не просто случайно и независимо, как в методе Монте-Карло.

В прикладной статистике скользящий контроль активно применяется, начиная с основополагающих работ Б. Эфрона [21, 18], и относится к «нетрадиционным методам» многомерного статистического анализа. В машинном обучении скользящий контроль признан де факто стандартной методикой эмпирического оценивания *обобщающей способности* (generalization ability), сравнения и выбора методов обучения [22].

В данной диссертационной работе *полный скользящий контроль* фактически принимается за определение обобщающей способности, а в слабой аксиоматике закладывается в само понятие вероятности.

Преимущества слабой аксиоматики проявляются в задачах анализа данных. Ещё раз резюмируем основные из них.

1. В задачах анализа данных выборки всегда конечны, будь то уже известные наблюдаемые данные или скрытые данные, которые станут известны в будущем. В некоторых практических задачах число предсказаний k настолько мало́, что оценивать «вероятность ошибки», которая есть предел частоты ошибок при $k \to \infty$, просто некорректно. В слабой аксиоматике рассматриваются только *статистики* — величины, которые могут быть определены непосредственно по конечной выборке.

2. Некоторые предположения сильной аксиоматики имеют недостаточное эмпирическое обоснование. Для проверки случайности, независимости и одинаковой

распределённости выработаны специальные статистические тесты. Однако гипотеза о существовании адекватной σ-аддитивной меры P на множестве объектов $\mathscr{X}$ эмпирической проверке не поддаётся [1]. Слабая аксиоматика обходится без теории меры, предъявляя существенно более скромные требования к пространству объектов и исходным данным. Об объектах вне конечной выборки $\mathbb{X}$ вообще не делается никаких предположений.

3. Вероятности вида (1.1) легко измеряются эмпирически по подмножеству разбиений, в частности, методом Монте-Карло. Поэтому результаты, полученные в слабой аксиоматике, всегда могут быть проверены экспериментально. В задачах статистического обучения такая проверка реализуется скользящим контролем.

Недостатки слабой аксиоматики связаны, главным образом, с наложением запретов на применение многих классических методов и приёмов теории вероятностей, и даже на употребление самого понятия «вероятность».

1. Нельзя сказать «вероятность ошибки», но можно сказать «вероятность того, что число ошибок на скрытой выборке превысит n». К этому трудно привыкать.

2. В отличие от асимптотических оценок, точные оценки часто представляют собой сложные комбинаторные формулы, требующие значительных объёмов вычислений. Упрощение или асимптотический анализ этих формул требует дополнительных математических усилий.

3. Многие континуальные вероятностные модели в физике, биологии, экономике и других естественных науках существенно опираются на асимптотическое понятие вероятности. Привлечение слабой аксиоматики в эти области вряд ли целесообразно, и приведёт лишь к усложнению математического аппарата.

О границах применимости слабой аксиоматики. Подсчёт доли разбиений (или перестановок) выборки, как технический приём, широко применяется и в рамках сильной аксиоматики. С его помощью получены многие содержательные результаты теории вероятностей, математической статистики, теории статистического обучения. Эти результаты легко переносятся в слабую аксиоматику. Возникает интригующий вопрос: насколько существенную часть этих математических теорий можно построить, пользуясь *только этим приёмом*, то есть в рамках слабой аксиоматики?

Основная цель последующих параграфов данной главы — показать примеры классических задач теории вероятностей и математической статистики, которые могут быть переформулированы и решены в терминах слабой аксиоматики. Вторая цель — наработать математическую технику, необходимую для рассмотрения более сложных задач статистического обучения в последующих главах.

2 Оценивание частоты события

Рассмотрим Задачу 1.3 о предсказании частоты события $S \subseteq \mathbb{X}$. Будем обозначать через $n(U) = |S \cap U|$ *число элементов* события S на произвольной конечной выборке $U \subseteq \mathbb{X}$.

Лемма 2.1. *Если $n(\mathbb{X}) = m$, то число элементов события S в наблюдаемой подвыборке $n(X)$ и в скрытой подвыборке $n(\bar{X})$ подчиняются гипергеометрическому распределению:*

$$\mathsf{P}\big[n(X) = s\big] = \mathsf{P}\big[n(\bar{X}) = m - s\big] = h_L^{\ell,m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \tag{2.1}$$

где s принимает значения от $s_0 = \max\{0, m-k\}$ до $s_1 = \min\{\ell, m\}$.

Доказательство. Отобрать s элементов события S в наблюдаемую подвыборку можно C_m^s различными способами. Для каждого из этих способов имеется $C_{L-m}^{\ell-s}$ способов сформировать оставшуюся часть наблюдаемой подвыборки из объектов, не принадлежащих S. Значит, $C_m^s C_{L-m}^{\ell-s}$ — число разбиений, при которых s элементов множества S попадают в наблюдаемую подвыборку, остальные $(m-s)$ — в скрытую. Их доля в общем числе разбиений $N = C_L^\ell$ как раз и составляет $h_L^{\ell,m}(s)$. ■

Замечание 2.1. Если не выполняется одно из условий $0 \leqslant s \leqslant m$, $0 \leqslant \ell - s \leqslant L - m$, или, что то же самое, не выполняется условие $s_0 \leqslant s \leqslant s_1$, то соответствующее число разбиений равно нулю. В этом случае будем полагать $h_L^{\ell,m}(s) = 0$.

§2.1 Свойства гипергеометрического распределения

Гипергеометрическое распределение носит фундаментальный характер и возникает, как мы увидим далее, в большинстве задач эмпирического предсказания. В данном разделе перечисляются в справочном порядке основные свойства гипергеометрического распределения [2, 3].

1. При фиксированных L и ℓ функция $h_L^{\ell,m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ определена на множестве пар целых чисел (m, s): $0 \leqslant m \leqslant L$, $\max\{0, m-k\} = s_0 \leqslant s \leqslant s_1 = \min\{\ell, m\}$. Это множество имеет форму параллелограмма, Рис. 3. Вне этой области принято полагать $h_L^{\ell,m}(s) = 0$.

2. Введём следующие обозначения для сумм крайних левых и крайних правых членов гипергеометрического распределения:

$$H_L^{\ell,m}(s) = \sum_{t=s_0}^{s} h_L^{\ell,m}(t)\,; \qquad \bar{H}_L^{\ell,m}(s) = \sum_{t=s}^{s_1} h_L^{\ell,m}(t)\,.$$

Справедлива формула полной вероятности:

$$\sum_{s=s_0}^{s_1} h_L^{\ell,m}(s) = H_L^{\ell,m}(s_1) = \bar{H}_L^{\ell,m}(s_0) = 1.$$

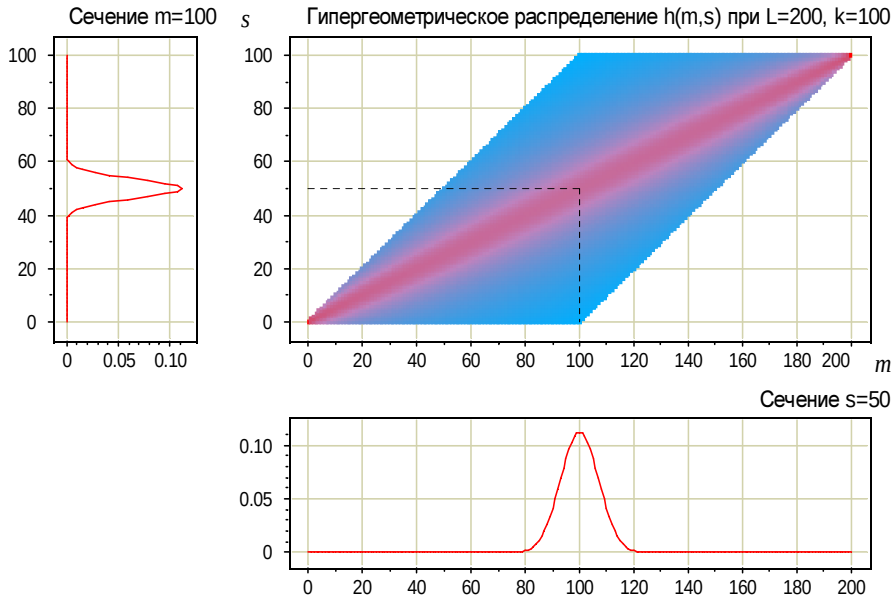

Рис. 3. Гипергеометрическое распределение $h_L^{\ell,m}(s)$ при $L=200$, $\ell=k=100$, $m=30$.

При фиксированных L, ℓ и m функция $h(s)=h_L^{\ell,m}(s)$ является одномерным дискретным распределением. Для примера на Рис. 3 слева показана функция $h(s)$ при фиксированном $m=100$. Функция $h'(m)=h_L^{\ell,m}(s)$ распределением, вообще говоря, не является, так как не удовлетворяет условию нормировки $\sum_m h'(m)\neq 1$. На Рис. 3 снизу показана функция $h'(m)$ при фиксированном $s=50$.

3. Параметры ℓ и m можно переставлять местами: $h_L^{\ell,m}(s)=h_L^{m,\ell}(s)$.
4. Параметры m и s можно заменять разностями: $h_L^{\ell,m}(s)=h_L^{\ell,L-m}(\ell-s)$.
5. Справедливы тождества:

$$h_L^{\ell,m}(s)=h_L^{\ell,L-m}(\ell-s)=h_L^{m,\ell}(s)=h_L^{m,k}(m-s)=h_L^{k,m}(m-s)\,.$$

6. Отсюда вытекают тождества для функций H и $\bar{H}$:

$$H_L^{\ell,m}(s)=\sum_{j=s_0}^{s}h_L^{\ell,m}(j)=\sum_{j=s_0}^{s}h_L^{k,m}(m-j)=\bar{H}_L^{k,m}(m-s)\,.$$

7. Распределение $h(s)$ является унимодальным (имеет форму пика). Максимальное значение достигается при $s^*=\frac{(m+1)(\ell+1)}{L+2}$, с точностью до округления.

8. Таблица гипергеометрического распределения содержит ℓk ненулевых значений. Её можно эффективно вычислить с помощью рекуррентных соотношений:

$$\begin{aligned}
h_L^{\ell,0}(0)&=1;\\
h_L^{\ell,m+1}(s)&=h_L^{\ell,m}(s)\,\frac{m+1}{m+1-s}\cdot\frac{k-m+s}{L-m};\\
h_L^{\ell,m}(s+1)&=h_L^{\ell,m}(s)\,\frac{m-s}{s+1}\cdot\frac{\ell-s}{k-m+s+1};\\
h_L^{\ell,m}(s-1)&=h_L^{\ell,m}(s)\,\frac{s}{m-s+1}\cdot\frac{k-m+s}{\ell-s+1}.
\end{aligned}\qquad(2.2)$$

Чтобы избежать накопления вычислительных погрешностей, значения $h_L^{\ell,m}(s)$ вычисляются последовательно для всех $m = 0,\ldots,L$. Для каждого m первым вычисляется максимальное значение или близкое к максимальному (достаточно взять $s = s_{\max}$), затем меньшие значения вычисляются через бóльшие.

9. Матожидание величины s равно $\lambda = \frac{\ell m}{L}$.

10. Дисперсия величины s равна $\sigma^2 = \lambda\frac{k(L-m)}{L(L-1)}$.

11. При больши́х значениях параметров L, ℓ, m предельными распределениями для $h(s) = h_L^{\ell,m}(s)$ могут быть только распределения одного из четырёх типов:

— при $\lambda \to \infty$ нормальное распределение $h(s) \to \frac{1}{\sqrt{2\pi}}\exp\bigl(-\frac{(s-\lambda)^2}{2\sigma^2}\bigr)$;
— если λ имеет конечный предел:
  — при $\frac{m}{L} \to p$ биномиальное распределение $h(s) \to C_\ell^s\, p^s(1-p)^{\ell-s}$;
  — при $\frac{\ell}{L} \to p$ биномиальное распределение $h(s) \to C_m^s\, p^s(1-p)^{m-s}$;
  — при $\frac{m}{L} \to 0$ или $\frac{\ell}{L} \to 0$ распределение Пуассона $h(s) \to e^{-\lambda}\lambda^s/s!$;
— при $\lambda \to 0$ вырожденное распределение $s = 0$.

12. Гипергеометрическое распределение довольно точно приближается с помощью аппроксимации Моленара:

$$h(s) \approx C_\ell^s\, \tilde{p}^s(1-\tilde{p})^{\ell-s}, \quad \tilde{p} = \frac{2m-s}{2L-\ell+1}.$$

§2.2 Закон больших чисел в слабой аксиоматике

Продолжим рассмотрение Задачи 1.3 о предсказании частоты события $S \subseteq \mathbb{X}$. Введём сокращённые обозначения для частот события S на выборках X и $\bar{X}$:

$$\nu = \frac{n(X)}{\ell}, \quad \bar{\nu} = \frac{n(\bar{X})}{k}.$$

Теорема 2.2. *Для любого $\varepsilon \in [0,1)$ справедливы точные оценки:*

$$\mathsf{P}\bigl[\nu \leqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(\lfloor \varepsilon\ell \rfloor\bigr); \tag{2.3}$$

$$\mathsf{P}\bigl[\bar{\nu} \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(\lfloor m - \varepsilon k \rfloor\bigr); \tag{2.4}$$

$$\mathsf{P}\bigl[\bar{\nu} - \nu \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(s_m^-(\varepsilon)\bigr), \qquad s_m^-(\varepsilon) = \bigl\lfloor \tfrac{\ell}{L}(m - \varepsilon k)\bigr\rfloor; \tag{2.5}$$

$$\mathsf{P}\bigl[|\bar{\nu} - \nu| \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(s_m^-(\varepsilon)\bigr) + \bar{H}_L^{\ell,m}\bigl(s_m^+(\varepsilon)\bigr), \qquad s_m^+(\varepsilon) = \bigl\lceil \tfrac{\ell}{L}(m + \varepsilon k)\bigr\rceil. \tag{2.6}$$

Доказательство. Первые два неравенства являются непосредственным следствием (2.1). Третье следует из первого, если заметить, что $\bar{\nu} - \nu \geqslant \varepsilon$ равносильно $\frac{m-s}{k} - \frac{s}{\ell} \geqslant \varepsilon$, откуда элементарными преобразованиями получаем $s \leqslant \frac{\ell}{L}(m - \varepsilon k)$.

Двусторонняя оценка (2.6) доказывается аналогично, если представить множество разбиений в виде объединения двух непересекающихся подмножеств:

$$\mathsf{P}\bigl[|\bar{\nu} - \nu| \geqslant \varepsilon\bigr] = \mathsf{P}\bigl[\bar{\nu} - \nu \geqslant \varepsilon\bigr] + \mathsf{P}\bigl[\nu - \bar{\nu} \geqslant \varepsilon\bigr] = H_L^{\ell,m}\bigl(s_m^-(\varepsilon)\bigr) + \bar{H}_L^{\ell,m}\bigl(s_m^+(\varepsilon)\bigr).$$

Теорема доказана. ■

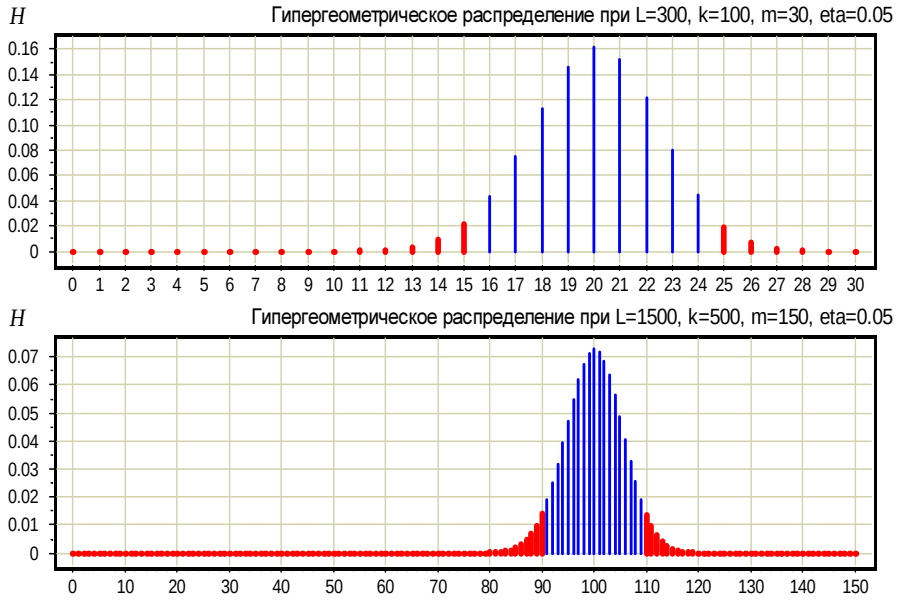

Рис. 4. Гипергеометрическое распределение $h_L^{\ell,m}(s)$. Верхний график получен при $L = 300$, $\ell = 200$, $m = 30$. Выделены крайние левые, $[s_0, s_m^-(\varepsilon)] = [0, 15]$, и крайние правые $[s_m^-(\varepsilon), s_1] = [25, 30]$, члены распределения, соответствующие значению надёжности $\eta = 0.05$. Нижний график получен при значениях L, ℓ, m, пропорционально увеличенных в 5 раз.

Замечание 2.2. В условии теоремы под $\lfloor z \rfloor$ понимается целая часть действительного числа z, то есть наибольшее целое, *меньшее или равное* z. Аналогично, $\lceil z \rceil$ — наименьшее целое, *большее или равное* z. Если в левой части поменять нестрогие неравенства на строгие, то все оценки останутся в силе с одной оговоркой: $\lfloor z \rfloor$ надо будет понимать как наибольшее целое, *меньшее* z; оно отличается от функции целой части только тем, что $\lfloor z \rfloor = z - 1$ при целых z. Соответственно, и $\lceil z \rceil$ надо будет понимать как наименьшее целое, *большее* z, тогда $\lceil z \rceil = z + 1$ при целых z.

О законе больших чисел. При пропорциональном увеличении L, ℓ и m относительная ширина гипергеометрического пика уменьшается, см. Рис. 4. В пределе при $L, \ell, m \to \infty$ это позволяет сколь угодно точно предсказывать частоту события S в скрытой выборке $\bar{\nu}$ по его частоте на наблюдаемой выборке ν. Равенство (2.6) даёт точную оценку скорости сходимости частот события в двух выборках.

Классический закон больших чисел утверждает сходимость частоты события к её вероятности. Однако в слабой аксиоматике понятие «вероятности события S» не определено. Поэтому (2.6) можно интерпретировать как *аналог закона больших чисел* в слабой аксиоматике. Основанием для такой интерпретации служит тот факт, что два функционала — (а) вероятность большого отклонения частот события в двух выборках и (б) вероятность большого отклонения частоты события от его вероятности — оцениваются сверху друг через друга [4]. По сути, эти две оценки отличаются не принципиально.

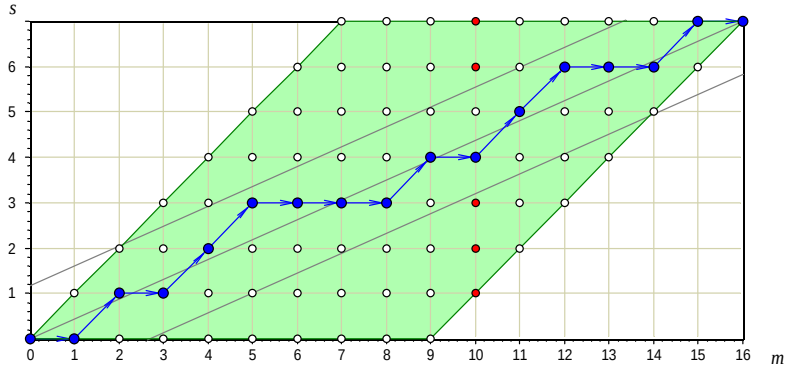

Рис. 5. Траектория, соответствующая бинарной последовательности $\{b_i\}_{i=0}^{L} =$ $= 0101100010110010$, при $L = 16$, $\ell = 7$, $\varepsilon = 0.3$. Проведены линии $s = \frac{\ell}{L}m$ и $s = \frac{\ell}{L}(m \pm \varepsilon k)$. При $m = 10$ выделены точки выше верхней линии, $s \geqslant s_m^+(\varepsilon)$, и ниже нижней линии, $s \leqslant s_m^-(\varepsilon)$.

Классические неравенства Чебышёва, Чернoффа, Бернштейна, Хёффдинга [25] позволяют количественно оценивать скорость сходимости в законе больших чисел. Однако все они являются асимптотическими и дают несколько завышенные оценки вероятности большого уклонения. Выражение (2.6) является точной (не завышенной, не асимптотической) оценкой, и потому его можно считать наиболее точным выражением закона больших чисел.

Геометрическая интерпретация соотношений (2.1), (2.6) и (2.6). Рассмотрим прямоугольную сетку $\{0,\ldots,L\} \times \{0,\ldots,\ell\}$, см. Рис. 5. Положим $b_i = [x_i \in X]$, то есть $b_i = 1$ означает, что при разбиении $\mathbb{X} = X \sqcup \bar{X}$ объект x_i попадает в наблюдаемую часть выборки. Договоримся отображать выборку в виде траектории, проходящей по узлам сетки из точки $(0,0)$ в точку (L,ℓ) согласно правилу: если $b_i = 1$, то смещаемся на единицу вправо-вверх; если $b_i = 0$, то смещаемся на единицу вправо. *Допустимыми* являются те и только те траектории, которые не выходят за пределы параллелограмма, выделенного на Рис. 5. Множество всех допустимых траекторий изоморфно множеству разбиений выборки $\mathbb{X} = X \sqcup \bar{X}$, и оба они изоморфны множеству L-мерных бинарных векторов $(b_1,\ldots,b_L)$, содержащих ровно ℓ единиц. Поэтому для подсчёта числа разбиений, удовлетворяющих некоторому свойству, достаточно найти число соответствующих траекторий.

Чтобы вывести (2.1), пронумеруем объекты выборки так, чтобы первые m объектов принадлежали множеству S. Тогда задача сведётся к подсчёту доли допустимых траекторий, проходящих через точку (m,s). Число допустимых траекторий на отрезке от $(0,0)$ до (m,s) есть C_m^s, и для каждой такой траектории существует $C_{L-m}^{\ell-s}$ вариантов её продолжения от (m,s) до (L,ℓ). Всего $C_m^s C_{L-m}^{\ell-s}$ траекторий. Разделив на общее число траекторий C_L^ℓ, получаем требуемое $h_L^{\ell,m}(s)$.

Чтобы вывести (2.5), необходимо подсчитать число траекторий, проходящих через любую точку (m,s), лежащую ниже диагонали на $\varepsilon\frac{\ell k}{L}$ или более. Для этого суммируется число траекторий $C_m^s C_{L-m}^{\ell-s}$ по всем $s = s_0,\ldots,s_m^-(\varepsilon)$.

Для вывода двусторонней оценки (2.6) подсчитывается число траекторий, отстоящих от диагонали на $\varepsilon\frac{\ell k}{L}$ или более. В этом случае суммирование идёт по всем $s = s_0, \ldots, s_m^-(\varepsilon)$, затем по всем $s = s_m^+(\varepsilon), \ldots, s_1$.

Задача выборочного контроля качества является классическим примером прикладной задачи, в которой оценки Теоремы 2.2 применяются непосредственно [2]. Пусть $\mathbb{X}$ — множество изделий, $S \subset \mathbb{X}$ — подмножество дефектных изделий. Изготовлена партия изделий $\mathbb{X}$, из них m оказались дефектными. Число m неизвестно. Проверить всю партию поштучно не представляется возможным. Поэтому делается *выборочный контроль*: случайно, независимо, без возвращений выбирается подмножество $X \subset \mathbb{X}$, что равносильно случайному равномерному выбору разбиения $X \sqcup \bar{X} = \mathbb{X}$. Зная долю дефектов в наблюдаемой подвыборке ν, требуется предсказать долю дефектов в скрытой подвыборке $\bar{\nu}$. Если при заданной точности ε и надёжности η имеет место оценка $\mathsf{P}[\bar{\nu} \geqslant \varepsilon] < \eta$, то партия $\mathbb{X}$ принимается, иначе она целиком бракуется. Параметры ε и η подбираются из экономических соображений — с учётом стоимости контроля одного изделия и величины потерь от использования дефектного изделия.

§2.3 Проблема неизвестного m и наблюдаемые оценки

Оценочные функции (2.3)–(2.6) зависят от числа элементов m события S в генеральной выборке $\mathbb{X}$, которое невозможно узнать, пока неизвестна скрытая часть данных. Таким образом, оценки (2.3)–(2.6) являются ненаблюдаемыми.

Прежде чем применить описанный в §1.3 (стр. 11) переход от ненаблюдаемой оценки к наблюдаемой, покажем, что известные альтернативные подходы либо не дают достаточно точных оценок, либо требуют привлечения субъективной дополнительной информации.

Верхняя оценка. Простейшее решение проблемы неизвестного m заключается в том, чтобы взять максимум по m и получить вместо точной оценки завышенную верхнюю оценку:

$$\mathsf{P}\big[\bar{\nu} - \nu \geqslant \varepsilon\big] \leqslant \max_{m=0,\ldots,L} H_L^{\ell,m}\left(s_m^-(\varepsilon)\right) \equiv \Gamma_L^\ell(\varepsilon). \tag{2.7}$$

Здесь максимум достаточно взять по всем m от $\lceil \varepsilon k \rceil$ до $\lfloor L - \varepsilon\ell \rfloor$, так как при остальных значениях m левая часть неравенства равна нулю.

К сожалению, (2.7) — довольно грубая оценка при малых m, см. Рис. 6. По этой причине данный подход не приемлем для задач выборочного контроля качества, обучения по прецедентам, и других случаев, когда именно малые значения m представляют большой практический интерес.

Известна верхняя оценка «хвоста» гипергеометрического распределения [20], с помощью которой можно оценить сверху правую часть (2.7): для любого $\varepsilon > 0$

$$\Gamma_L^\ell(\varepsilon) \leqslant \exp\left(-2\varepsilon^2\frac{\ell k^2}{L^2}\right).$$

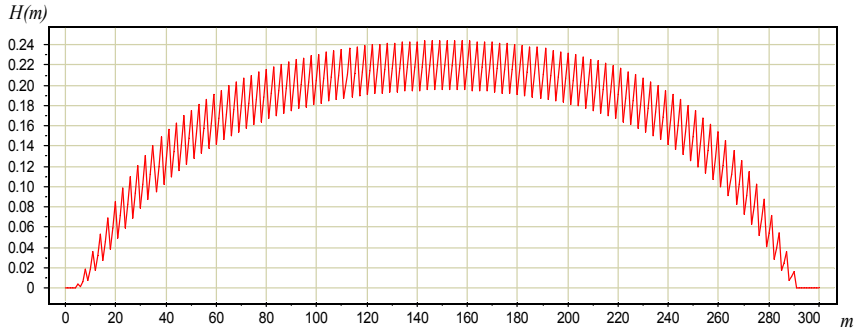

Рис. 6. График зависимости $H(m) = H_L^{\ell,m}\left(s_m^-(\varepsilon)\right)$ от m при $L = 300,\ \ell = 200,\ \varepsilon = 0.05$.

Эта оценка ещё более грубая (на Рис. 6 ей соответствовала бы горизонтальная линия с ординатой 0.89, но она не показана). Асимптотически эта оценка сходится к нулю при одновременном стремлении ℓ и k к бесконечности, что ещё раз подтверждает связь точных оценок (2.5) и (2.6) с законом больших чисел.

Байесовский подход. В теории выборочного контроля качества m — это критически важная величина — число бракованных изделий в общей партии из L штук изделий. Возникают вопросы: является ли m случайной величиной или неслучайным параметром; и если это случайная величина, то каково её распределение? Окончательного ответа на них в [2] не даётся. Вместо этого предлагается несколько альтернативных подходов, приводящих, вообще говоря, к разным результатам.

Пусть m — случайная величина с заданным априорным распределением $p(m)$. Тогда, зная число s элементов события S в наблюдаемой выборке и зная распределение $p(s|m) = H_L^{\ell,m}(s)$, можно оценить апостериорное распределение $p(m|s)$ по формуле Байеса. В [2] рассмотрено несколько вариантов задания априорного распределения: равномерное, биномиальное, гипергеометрическое, и даже смеси биномиальных или гипергеометрических распределений.

В слабой аксиоматике, чтобы применить байесовский подход, необходимо определить вероятность $p(m)$ через долю разбиений выборки. Для этого придётся ввести расширенную генеральную выборку $X^{\mathbb{L}}$, в которой M элементов принадлежат событию S, и предположить, что выборка $\mathbb{X}$ равновероятна среди всех $C_{\mathbb{L}}^L$ разбиений расширенной выборки $X^{\mathbb{L}}$. Тогда величина m будет подчиняться гипергеометрическому распределению $p(m) = h_{\mathbb{L}}^{L,M}(m)$ с неизвестным параметром M. Однако вопрос «чему равно M для выборки $X^{\mathbb{L}}$», ничем не отличается от исходного вопроса «чему равно m для выборки $\mathbb{X}$». Таким образом, априорная вероятность $p(m)$ не имеет частотной интерпретации, и её следует понимать как *субъективную вероятность*.

Недостаток байесовского подхода в том, что результат существенно зависит от $p(m)$, однако привнесение субъективной дополнительной информации представляется неоправданным в большинстве приложений.

Из приведённых рассуждений следует также, что неизвестную величину m целесообразно рассматривать как неслучайный параметр, значение которого необходимо оценить по случайной наблюдаемой выборке X.

Переход от ненаблюдаемой оценки к наблюдаемой не требует привлечения дополнительной информации и приводит к точным (достигаемым) верхним и нижним оценкам для $n(\mathbb{X})$ и $n(\bar{X})$, которые могут быть вычислены по наблюдаемому значению числа ошибок $s = n(X)$.

Теорема 2.3. *Если $s = n(X)$ — число элементов события S в наблюдаемой выборке, то для числа элементов события S в полной выборке с вероятностью $(1-\eta)$ справедлива верхняя оценка:*

$$n(\mathbb{X}) \leqslant \max\bigl\{m = m_0, \ldots, L \;\bigm|\; H_L^{\ell,m}(s) \geqslant \eta\bigr\}, \quad \text{где } m_0 = \bigl\lceil s\tfrac{L+2}{\ell+1} - 1\bigr\rceil. \tag{2.8}$$

Доказательство. Рассмотрим одностороннюю точную оценку (2.4), обозначив правую её часть через $H(\varepsilon, m)$, где $m = n(\mathbb{X})$ — неизвестная величина:

$$\mathsf{P}\bigl[\bar{\nu} \geqslant \varepsilon\bigr] = H_L^{\ell,m}(\lfloor m - \varepsilon k \rfloor) = H(\varepsilon, m). \tag{2.9}$$

Тогда с вероятностью $(1-\eta)$ справедлива оценка сверху $\bar{\nu} < E(\eta, m)$, где $E(\eta, m)$ — обратная функция от $H(\varepsilon, m)$. Обращение производится по первому аргументу при каждом значении второго аргумента m, который выступает в роли параметра. Поскольку функция $E(\eta, m)$ не возрастает по первому аргументу, из оценки $\bar{\nu} < E(\eta, m)$ следует неравенство $H(\bar{\nu}, m) \geqslant \eta$. Подставляя $\bar{\nu} = \frac{m-s}{k}$ в функцию $H(\bar{\nu}, m)$, определяемую согласно (2.9), получаем, что с вероятностью $(1-\eta)$ справедливо неравенство $H_L^{\ell,m}(s) \geqslant \eta$. Чтобы разрешить данное неравенство относительно m при фиксированном s, найдём максимальное значение m, при котором оно выполнено. При максимальном значении m значение s должно находится левее точки максимума гипергеометрического распределения $s^* = \frac{(m+1)(\ell+1)}{L+2}$, Следовательно, $s(L+2) < (m+1)(\ell+1)$. Поэтому для нахождения максимального значения m достаточно перебрать значения m, не меньшие $s\frac{L+2}{\ell+1} - 1$. Теорема доказана. ■

Замечание 2.3. Аналогичным образом можно оценить частоту на скрытой выборке $\bar{\nu}$, зная только частоту на наблюдаемой выборке ν: с вероятностью $(1-\eta)$ выполнено неравенство

$$\bar{\nu} \leqslant \tfrac{1}{k}\max\bigl\{t = t_0, \ldots, k \;\bigm|\; H_L^{\ell,\nu\ell+t}(\nu\ell) \geqslant \eta\bigr\}, \quad t_0 = \bigl\lceil s\tfrac{k+1}{\ell+1} - 1\bigr\rceil. \tag{2.10}$$

Замечание 2.4. Аналогичным образом строятся и нижние оценки: с вероятностью $(1-\eta)$ выполнены неравенства

$$n(\mathbb{X}) \geqslant \min\bigl\{m = 0, \ldots, m_0 \;\bigm|\; \bar{H}_L^{\ell,m}(s) \geqslant \eta\bigr\};$$
$$\bar{\nu} \geqslant \tfrac{1}{k}\min\bigl\{t = 0, \ldots, t_0 \;\bigm|\; \bar{H}_L^{\ell,\nu\ell+t}(\nu\ell) \geqslant \eta\bigr\},$$

Замечание 2.5. Вычисление полученных верхних и нижних оценок с использованием рекуррентных соотношений (2.2) требует порядка $O\bigl(n(X)n(\bar{X})\bigr)$ операций.

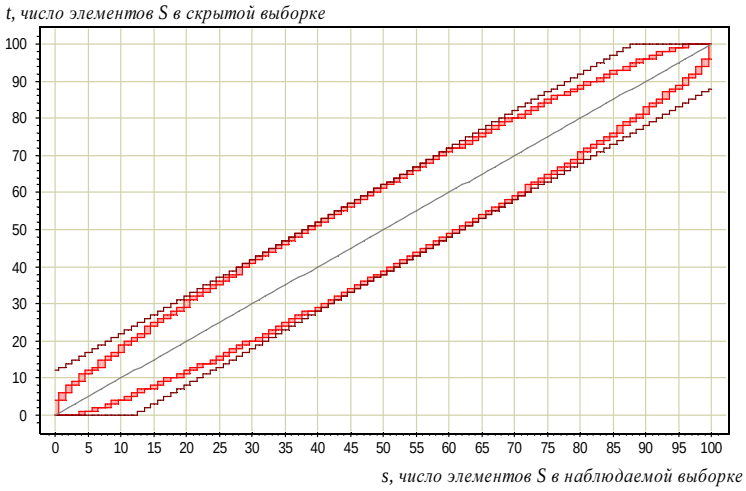

Рис. 7. Точные верхние и нижние оценки числа $t = n(\bar{X})$ элементов события S в скрытой выборке в зависимости от их числа $s = n(X)$ в наблюдаемой выборке. Условия эксперимента: $L = 200,\ \ell = k = 100,\ \eta = 0.05$.

Замечание 2.6. Достижимость полученных выше оценок надо понимать в следующем смысле. Когда говорят, что неравенство выполнено с вероятностью $(1-\eta)$, это означает, что оно выполнено при $(1-\eta)C_L^\ell$ разбиениях и не выполнено при ηC_L^ℓ разбиениях. При этом существует хотя бы одно разбиение, при котором неравенство выполнено как равенство.

На Рис. 7 показаны верхние и нижние оценки числа элементов события S в скрытой выборке $t = n(\bar{X})$ в зависимости от их числа в наблюдаемой выборке $s = n(X)$. Толстые ступенчатые линии — граничные области, в которых выполняется равенство $H_L^{\ell,s+t}(s) = \eta$. Точная верхняя оценка совпадает с верхней границей верхней области, точная нижняя — с нижней границей нижней области. Вместе они определяют $1-2\eta = 90\%$-й доверительный интервал для числа t при каждом значении s. Тонкие ступенчатые линии — оценки по наихудшему m (2.7), их точность падает по мере приближения m к 0 или к L.

О вероятности нуль-события. В работе С. И. Гурова [9] ставится задача точечного оценивания вероятности p события, ни разу не наблюдавшегося в заданной выборке длины ℓ. Рассматриваются следующие типы оценок (предполагается, что коэффициент доверия $\eta \in (0,1)$ задаётся априори и достаточно близок к единице):

1. Оценка максимального правдоподобия вырождена и равна нулю.

2. Верхняя граница доверительного интервала согласно классическому частотному подходу равна $\hat{p} = 1 - \sqrt[\ell]{1-\eta}$ и представляется сильно завышенной.

3. Байесовские оценки по математическому ожиданию $\left(\hat{p}_B = \frac{1}{\ell+2}\right)$ или по медиане $\left(\hat{p} = 1 - \sqrt[\ell]{0.5}\right)$ апостериорного распределения представляются завышенными при больших ℓ.

4. Наконец, предлагается оценка $\hat{p}_0(\ell) = 1 - \sqrt[\ell]{\eta}$, получаемая несколькими способами, в том числе путём замены наблюдавшейся выборки некоторой гипотетиче-

ской выборкой, в которой событие произошло хотя бы один раз. Замена делается таким образом, чтобы для данной пары выборок принималась гипотеза однородности. Затем к новой выборке применяется оценка максимального правдоподобия. В качестве критерия однородности используется *точный тест Фишера*, основанный на гипергеометрическом распределении.

Основные выводы [9] следующие: при малых выборках (ℓ от 4 до 20–30 объектов) рекомендуется использовать байесовскую оценку $\hat{p}_B$; при больших выборках (ℓ более 20–30 объектов) — оценку $\hat{p}_0(\ell)$. Констатируется наличие резкого скачка оценки при переходе от «малой» выборки к «большой».

В описанном подходе вызывает неудовлетворённость как большое количество методик, дающих противоречивые результаты, так и искусственный, эвристический характер этих методик. Остаётся не ясно, сколько ещё эвристик можно изобрести, и дадут ли они схожие оценки. Но, главное, вызывает сомнение правомерность самого понятия «вероятности редкого события», а также возможность и полезность его точечного оценивания, особенно в условиях малых выборок.

В рамках слабой аксиоматики задача точечного оценивания вероятности нуль-события неправомерна. В то же время, вполне правомерна задача интервального оценивания частоты нуль-события на произвольной скрытой выборке заданной длины k. Это частный случай задачи эмпирического предсказания 1.3 (стр. 6) при условии, что $\nu(X) = 0$. Данная задача имеет точное решение (2.10) при $\nu = 0$. На Рис. 7 видно, что при $s = 0$ и длине наблюдаемой выборки $\ell = 100$ число событий в скрытой выборке длины $k = 100$ не превзойдёт 4 с вероятностью $1 - \eta = 95\%$.

Таким образом, в слабой аксиоматике решение является математически строгим и единственным; отпадает необходимость изобретения эвристических оценок.

3 Оценивание функции распределения

Рассмотрим Задачу 1.5 об оценивании функции распределения. Для произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$ и произвольного разбиения $X \sqcup \bar{X} = \mathbb{X}$ определим одностороннее и двустороннее равномерное отклонение эмпирических функций распределения:

$$D^+(X, \bar{X}) = \max_{z\in\mathbb{R}} \big(F_\xi(z, \bar{X}) - F_\xi(z, X)\big);$$
$$D^-(X, \bar{X}) = \max_{z\in\mathbb{R}} \big(F_\xi(z, X) - F_\xi(z, \bar{X})\big);$$
$$D(X, \bar{X}) = \max_{z\in\mathbb{R}} \big|F_\xi(z, \bar{X}) - F_\xi(z, X)\big|.$$

В сильной аксиоматике имеет место следующая теорема [15].

Теорема 3.1 (Н. В. Смирнов). *Если $X, \bar{X} \subseteq \mathscr{X}$ — случайные, независимые, одинаково распределённые выборки; $\xi\colon \mathscr{X} \to \mathbb{R}$ — случайная величина с непрерывным*

распределением, то справедливы асимптотические оценки

$$\lim_{\ell,k\to\infty} \mathsf{P}\big\{D^{\pm}(X,\bar{X}) \geqslant \varepsilon\big\} = \exp\left(-2\varepsilon^2\tfrac{\ell k}{\ell+k}\right); \tag{3.1}$$

$$\lim_{\ell,k\to\infty} \mathsf{P}\big\{D(X,\bar{X}) \geqslant \varepsilon\big\} = 2\sum_{i=1}^{\infty}(-1)^{i+1}\exp\left(-2\varepsilon^2\tfrac{\ell k}{\ell+k}i^2\right); \tag{3.2}$$

Заметим, что правая часть (3.2) представима также в виде $1-K\left(\varepsilon\sqrt{\frac{\ell k}{\ell+k}}\right)$, где $K(z)=\sum\limits_{i=-\infty}^{\infty}(-1)^i e^{-2z^2i^2}$ — функция распределения Колмогорова.

Известны и неасимптотические точные оценки, но они имеют достаточно громоздкий вид [7]. Мы покажем, что точные оценки могут быть выражены более элегантно через усечённый треугольник Паскаля [19]. Доказательство проводится в рамках слабой аксиоматики и имеет прозрачный геометрический смысл.

§3.1 Усечённый треугольник Паскаля

Пусть $g_m^-,\ g_m^+,\ m=0,\ldots,L$ — две неубывающие последовательности, удовлетворяющие условию $0\leqslant g_m^-\leqslant g_m^+\leqslant m$.

Опр. 3.1. *Усечённым треугольником Паскаля с нижней границей g_m^- и верхней границей g_m^+ будем называть целочисленную функцию $G_m^s=G_m^s[g_m^-,g_m^+]$, определённую рекуррентными соотношениями $G_0^s=[s=0]$ и*

$$G_m^s = \big(G_{m-1}^s + G_{m-1}^{s-1}\big)\big[g_m^- \leqslant s \leqslant g_m^+\big], \quad m\in\mathbb{N}, \quad s\in\mathbb{Z}. \tag{3.3}$$

Усечённый треугольник Паскаля вычисляется по тому же рекуррентному правилу, что и классический треугольник Паскаля, если в нём обнулить все элементы, лежащие за пределами границ $[g_m^-,g_m^+]$. «Неусечённый» треугольник Паскаля $G_m^s[0,m]$ совпадает с классическим и даёт биномиальные коэффициенты C_m^s.

При начальном условии $G_0^s=[s=0]$ ненулевыми могут быть только элементы G_m^s при $0\leqslant s\leqslant m$. Другие начальные условия приводят к неклассическим обобщениям треугольника Паскаля, которые в данной работе не рассматриваются.

Определим для произвольных $\varepsilon>0,\ m=0,1,2,\ldots,$ следующие функции (в дальнейшем аргумент ε будем опускать):

$$g_m^+(\varepsilon) = \tfrac{\ell}{L}(m+\varepsilon k);$$
$$g_m^-(\varepsilon) = \tfrac{\ell}{L}(m-\varepsilon k).$$

На Рис. 8 и 9 приведены четыре возможных варианта усечённых треугольников Паскаля с такими границами. В отличие от принятого способа изображения здесь они «положены на бок» путём поворота на 90° против часовой стрелки.

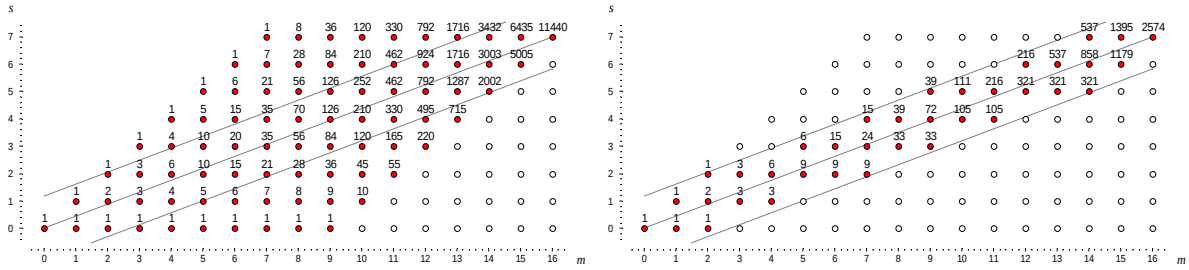

Рис. 8. Треугольники Паскаля: классический $C_m^s = G_m^s[0,m]$ и усечённый $G_m^s\big[g_m^-(\varepsilon), g_m^+(\varepsilon)\big]$, при $L=16$, $\ell=7$, $\varepsilon=0.3$.

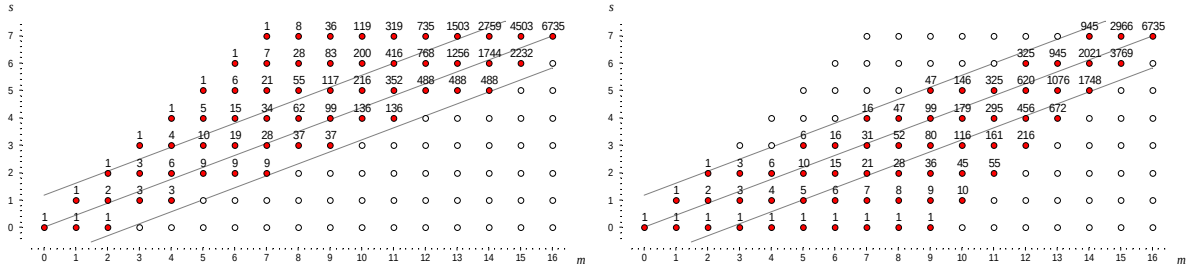

Рис. 9. Треугольники Паскаля: усечённый слева $G_m^s\big[g_m^-(\varepsilon), m\big]$ и усечённый справа $G_m^s\big[0, g_m^+(\varepsilon)\big]$, при $L=16$, $\ell=7$, $\varepsilon=0.3$.

§3.2 Теорема Смирнова в слабой аксиоматике

Теорема 3.2. *В слабой аксиоматике для произвольной конечной выборки $\mathbb{X}$ и произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$, значения которой попарно различны на элементах выборки $\mathbb{X}$, справедливы точные оценки:*

$$\mathsf{P}\big[D^+(X,\bar{X}) \leqslant \varepsilon\big] = G_L^\ell\big[0, g_L^+(\varepsilon)\big]/C_L^\ell; \tag{3.4}$$

$$\mathsf{P}\big[D^-(X,\bar{X}) \leqslant \varepsilon\big] = G_L^\ell\big[g_L^-(\varepsilon), m\big]/C_L^\ell; \tag{3.5}$$

$$\mathsf{P}\big[D(X,\bar{X}) \leqslant \varepsilon\big] = G_L^\ell\big[g_L^-(\varepsilon), g_L^+(\varepsilon)\big]/C_L^\ell. \tag{3.6}$$

Доказательство. 1. Составим вариационный ряд значений функции $\xi(x)$ на элементах выборки: $\xi(x^{(1)}) < \xi(x^{(2)}) < \cdots < \xi(x^{(L)})$. Здесь все неравенства строгие в силу условия попарной различности.

Обозначим $b_i = [x^{(i)} \in X]$. Бинарная последовательность $b_1, \ldots, b_L$ содержит ровно ℓ единиц и k нулей. Покажем, что равномерное отклонение эмпирических распределений на выборках X и $\bar{X}$ выражается через равномерное отклонение числа единиц в первых m членах последовательности $B_m = b_1 + \cdots + b_m$ от «ожидаемого»

числа единиц $m\ell/L$:

$$\begin{aligned}
D(X,\bar{X}) &= \max_{z\in\mathbb{R}} \left|F_\xi(z,X^k) - F_\xi(z,X^\ell)\right| = \\
&= \max_{z\in\mathbb{R}} \left|\frac{1}{k}\sum_{i=1}^{L}[x_i\in\bar{X}][\xi(x_i)<z] - \frac{1}{\ell}\sum_{i=1}^{L}[x_i\in X][\xi(x_i)<z]\right| = \\
&= \max_{m=1..L} \left|\frac{1}{k}\sum_{i=1}^{m}\underbrace{[x^{(i)}\in\bar{X}]}_{1-b_i} - \frac{1}{\ell}\sum_{i=1}^{m}\underbrace{[x^{(i)}\in X]}_{b_i}\right| = \\
&= \max_{m=1..L} \left|\frac{m}{k} - \frac{\ell+k}{\ell k}\sum_{i=1}^{m} b_i\right| = \max_{m=1..L} \frac{L}{\ell k}\left|B_m - \frac{m\ell}{L}\right|.
\end{aligned}$$

Теперь запишем долю разбиений выборки $\mathbb{X}$, при которых равномерное отклонение эмпирических распределений не превышает пороговую точность ε:

$$\begin{aligned}
&\mathsf{P}\big[D(X,\bar{X})\leqslant\varepsilon\big] = \mathsf{P}\left[\max_{m=1..L}\left|B_m - \frac{m\ell}{L}\right| \leqslant \frac{\varepsilon\ell k}{L}\right] = \\
&= \mathsf{P}\left[\max_{m=1..L}\Big(-B_m + \underbrace{\big(\tfrac{m\ell}{L} - \tfrac{\varepsilon\ell k}{L}\big)}_{g_m^-(\varepsilon)}\Big)\leqslant 0\right]\left[\max_{m=1..L}\Big(B_m - \underbrace{\big(\tfrac{m\ell}{L} + \tfrac{\varepsilon\ell k}{L}\big)}_{g_m^+(\varepsilon)}\Big)\leqslant 0\right] = \\
&= \frac{1}{N}\sum_{n=1}^{N}\prod_{m=1}^{L}\big[g_m^-(\varepsilon)\leqslant B_m\leqslant g_m^+(\varepsilon)\big]. \qquad (3.7)
\end{aligned}$$

Последнее равенство следует из тождества $\big[\max\limits_m A_m\leqslant 0\big] = \prod\limits_m \big[A_m\leqslant 0\big]$.

2. Рассмотрим выборку $X^m = \{x^{(1)},\ldots,x^{(m)}\}$, состоящую из первых m членов вариационного ряда. Возьмём максимальное (по включению) подмножество разбиений $N'\subseteq\{1,\ldots,N\}$, удовлетворяющих двум условиям:

1) они индуцируют попарно различные разбиения выборки X^m;

2) ровно s объектов из X^m попадают в X^ℓ.

Очевидно, число этих разбиений $|N'| = C_m^s$. Представим множество разбиений N' в виде объединения непересекающихся подмножеств $N_0' = \{n\in N' \,|\, b_m^n = 0\}$ и $N_1' = \{n\in N' \,|\, b_m^n = 1\}$. Очевидно, $|N_0'| = C_{m-1}^s$, $|N_1'| = C_{m-1}^{s-1}$.

Нас будет интересовать выражение $H_m^s = \sum\limits_{n\in N'}\prod\limits_{r=1}^{m}\big[g_r^- \leqslant B_r^n \leqslant g_r^+\big]$, поскольку правая часть (3.7) есть ни что иное, как $\frac{1}{N}H_L^\ell$. Разобьём в этом выражении сумму по N' на две суммы — по N_0' и по N_1', и ещё заметим, что $B_m^n = s$ для всех $n\in N'$:

$$\begin{aligned}
H_m^s &= \underbrace{\sum_{n\in N_0'}\prod_{r=1}^{m-1}\big[g_r^-\leqslant B_r^n\leqslant g_r^+\big]}_{H_{m-1}^s}\big[g_m^-\leqslant s\leqslant g_m^+\big] + \\
&+ \underbrace{\sum_{n\in N_1'}\prod_{r=1}^{m-1}\big[g_r^-\leqslant B_r^n\leqslant g_r^+\big]}_{H_{m-1}^{s-1}}\big[g_m^-\leqslant s\leqslant g_m^+\big] = \big(H_{m-1}^s + H_{m-1}^{s-1}\big)\big[g_m^-\leqslant s\leqslant g_m^+\big].
\end{aligned}$$

Таким образом, получена рекуррентная формула для H_m^s, формально совпадающая с формулой усечённого треугольника Паскаля (3.3). Осталось только проверить граничные случаи.

При $m = 1$ и фиксированном $s \in \{0, 1\}$ имеется только одно разбиение, $|N'| = 1$, следовательно, $H_1^s = \big[g_m^- \leqslant s \leqslant g_m^+\big]$, что совпадает с (3.3).

При $s = 0$ и произвольном $m = 1, \ldots, k$ имеется только одно разбиение, $|N'| = 1$, причём ни один объект из X^m не попадает в X^ℓ. Это означает, что $B_r^n = 0$ при всех $r = 1, \ldots, m$. Но тогда $H_m^0 = \prod_{r=1}^m \big[g_r^- \leqslant s \leqslant g_r^+\big]$, что, опять-таки, совпадает с (3.3).

Заметим также, что при $s = 0$ запись $G_{m-1}^{s-1} = 0$ по определению корректна, в то же время $H_{m-1}^{s-1} = 0$, поскольку $N_1' = \varnothing$. Аналогично, при $s = m$ имеем $N_0' = \varnothing$, следовательно, $H_{m-1}^s = 0 = G_{m-1}^s$.

3. Односторонние оценки (3.4) и (3.5) выводится аналогично. Различие в том, что для них выражение под знаком произведения в (3.7) принимает вид, соответственно, либо $\big[0 \leqslant B_m^n \leqslant g_m^+(\varepsilon)\big]$, либо $\big[g_m^-(\varepsilon) \leqslant B_m^n \leqslant m\big]$. Изменяется только форма границы в усечённом треугольнике Паскаля, соответственно, либо нижней $g_m^-(\varepsilon) = 0$, либо верхней $g_m^+(\varepsilon) = m$, и все дальнейшие рассуждения остаются в силе. ■

Геометрическая интерпретация. Каждое разбиение $X \sqcup \bar{X} = \mathbb{X}$ взаимно однозначно соответствует бинарному вектору $b = (b_1, \ldots, b_L)$, состоящему из ℓ единиц и k нулей, и в то же время, некоторой траектории, проходящей из точки $(0, 0)$ в точку (L, ℓ) согласно правилу: если $b_i = 1$, то сместиться вправо и вверх на 1; если $b_i = 0$, то сместиться вправо на 1, см. Рис. 5. Выполнение совокупности условий $\big[g_m^- \leqslant B_m \leqslant g_m^+\big]$ при всех $m = 1, \ldots, L$ означает, что траектория не может проходить ниже границы g_m^- или выше границы g_m^+. На Рис. 5 эти границы показаны линиями. Согласно (3.7) функционал $\mathsf{P}\big[D(X, \bar{X}) \leqslant \varepsilon\big]$ в точности равен доле таких траекторий. Будем называть их допустимыми. Обозначим через H_m^s число допустимых траекторий, проходящих из точки $(0, 0)$ в точку (m, s). Допустимая траектория может придти в (m, s) либо из $(m-1, s-1)$, либо из $(m-1, s)$. Отсюда следует рекуррентная формула для числа допустимых траекторий: $H_m^s = H_{m-1}^s + H_{m-1}^{s-1}$. Однако, если $s \notin \big[g_m^-, g_m^+\big]$, то все эти траектории уже не будут допустимыми, поэтому окончательная формула принимает вид $H_m^s = \big(H_{m-1}^s + H_{m-1}^{s-1}\big)\big[g_m^- \leqslant s \leqslant g_m^+\big]$, что совпадает с определением усечённого треугольника Паскаля: $H_m^s \equiv G_m^s$.

Практическое вычисление по рекуррентным соотношениям (3.3) сталкивается с проблемой переполнения: значения G_m^s выходят за пределы разрядной сетки современных компьютеров при L порядка нескольких сотен. Проблема снимается, если вывести рекуррентную формулу для отношений $\varphi_m^s = G_m^s / C_m^s$, которые принимают значения из отрезка $[0, 1]$. Применив тождества $C_m^s = \frac{m}{m-s} C_m^{s-1} = \frac{m}{s} C_{m-1}^{s-1}$, получим:

$$\varphi_m^s = \frac{m-s}{m} \varphi_{m-1}^s + \frac{s}{m} \varphi_{m-1}^{s-1}.$$

Усечённый треугольник Паскаля оказывается полезной концепцией не только при выводе точного выражения для критерия Смирнова, но во многих задачах, свя-

занных со случайными блужданиями при ограничениях. Упомянем только выборочный контроль качества [2] и анализ выживаемости [19].

§3.3 Обобщение на случай вариационного ряда со связками

В Теореме 3.1 (Смирнова) требование непрерывности функции распределения является существенным. В сильной аксиоматике оно гарантирует, что с вероятностью 1 вариационный ряд $\xi(x^{(1)}) < \xi(x^{(2)}) < \cdots < \xi(x^{(L)})$ не содержит одинаковых элементов. Это нужно для того, чтобы ранжировка была определена единственным образом, иначе доказательство теоремы наталкивается на значительные технические трудности. Когда условие непрерывности нарушается, неравенства (3.1) и (3.2) могут не выполняться [3].

В Теореме 3.2 требование попарной различности формулировалось в явном виде. Покажем в слабой аксиоматике, что отказ от этого требования не сильно меняет вид результата — в Теореме 3.2 изменяются только границы усечения $[g_m^-, g_m^+]$.

Теорема 3.3. *Пусть $\xi\colon \mathbb{X} \to \mathbb{R}$ — произвольная функция, $\mathbb{X}$ — произвольная конечная выборка, вариационный ряд значений $\xi(x_i)$ состоит из H связок:*

$$\underbrace{\xi(x^{(1)}) = \cdots = \xi(x^{(i_2-1)})}_{\text{1-я связка}} < \underbrace{\xi(x^{(i_2)}) = \cdots = \xi(x^{(i_3-1)})}_{\text{2-я связка}} < \cdots < \underbrace{\xi(x^{(i_H)}) = \cdots = \xi(x^{(L)})}_{\text{H-я связка}}.$$

Тогда в слабой аксиоматике справедливы точные оценки (3.4), (3.5), (3.6), *если взять границы усечённого треугольника Паскаля $[\tilde g_m^-, \tilde g_m^+]$:*

$$\tilde g_m^+(\varepsilon) = \min\bigl\{g_{i_h}^+(\varepsilon) + m - i_h,\ g_{i_{h+1}}^+(\varepsilon)\bigr\};$$
$$\tilde g_m^-(\varepsilon) = \max\bigl\{g_{i_h}^-(\varepsilon),\ g_{i_{h+1}}^-(\varepsilon) + m - i_{h+1}\bigr\};$$

для всех $m = i_h, \ldots, i_{h+1} - 1$, где h пробегает значения от 1 до H, и $i_{H+1} = L + 1$.

Доказательство. ■

Замечание 3.1. Если все связки одноэлементные, $\{i_1, \ldots, i_H\} \equiv \{1, \ldots, L\}$, то $\tilde g_m^+(\varepsilon) = g_m^+(\varepsilon)$, $\tilde g_m^-(\varepsilon) = g_m^-(\varepsilon)$, и Теорема 3.3 переходит в Теорему 3.2.

Замечание 3.2. Полученные оценки являются точными, но ненаблюдаемыми. Модифицированные границы $[\tilde g_m^-, \tilde g_m^+]$ существенно зависят от последовательности $i_1, \ldots, i_H$, которая строится по всей генеральной выборке $\mathbb{X}$; её невозможно знать, имея лишь наблюдаемую выборку X. Это означает, что Теорему 3.3 можно применять для проверки статистических гипотез, однако непосредственно она не годится для эмпирического предсказания. В частности, для предсказания эмпирических распределений дискретнозначных функций ξ необходимо разрабатывать другие методы.

4 Некоторые ранговые статистики и критерии

§4.1 Доверительное оценивание

Рассмотрим Задачу 1.4 (стр. 7) о доверительном оценивании.

Задана функция $\xi\colon \mathbb{X} \to \mathbb{R}$. Требуется построить по наблюдаемой выборке X семейство вложенных *доверительных интервалов* $\Omega_\varepsilon(X) = \big[\xi_\varepsilon^-(X), \xi_\varepsilon^+(X)\big]$ такое, что для произвольного скрытого объекта $\bar{x}$ выполняется $\xi(\bar{x}) \in \Omega_\varepsilon(X)$ с вероятностью не менее $1 - \eta(\varepsilon)$.

Семейство вложенных интервалов построим следующим образом.

Вариационным рядом функции ξ на выборке $U = \{u_1, \ldots, u_t\} \subseteq \mathbb{X}$ называется последовательность значений $\xi(u_1), \ldots, \xi(u_t)$, упорядоченная по возрастанию. Обозначим s-е значение вариационного ряда ξ на U через $\xi_U^{(s)}$, тогда $\xi_U^{(1)} < \cdots < \xi_U^{(t)}$.

Теорема 4.1. *Определим семейство вложенных отрезков* $\Omega_\varepsilon(X) = \big[\xi_X^{(\ell-\varepsilon+1)}, \xi_X^{(\varepsilon)}\big]$, *где* $\varepsilon = \lceil \ell/2 \rceil, \ldots, \ell$. *Тогда справедлива точная оценка:*

$$\mathsf{P}\big[\xi(\bar{x}) \notin \Omega_\varepsilon(X)\big] = 2(1 - \varepsilon/L), \quad \varepsilon = \lceil \ell/2 \rceil, \ldots, \ell. \tag{4.1}$$

Доказательство. Всего имеется $N = C_L^1 = L$ разбиений. Величина $\xi(\bar{x})$ превосходит $\xi_X^{(\varepsilon)}$ на тех разбиениях, при которых правее $\xi(\bar{x})$ в вариационном ряду находится менее $L-\varepsilon$ объектов. Таких разбиений ровно $L-\varepsilon$. Аналогично, $\xi(\bar{x}) < \xi_X^{(L-\varepsilon+1)}$ на тех разбиениях, при которых левее $\xi(\bar{x})$ находится менее $L-\varepsilon$ объектов. Таких разбиений также ровно $L - \varepsilon$. Итак, доля разбиений, при которых значение $\xi(\bar{x})$ попадает вне отрезка $\Omega_\varepsilon(X)$, составляет $2(L - \varepsilon)/L$. ■

Аналогично, справедлива точная верхняя оценка:

$$\mathsf{P}\big[\xi(\bar{x}) > \xi_X^{(\varepsilon)}\big] = 1 - \varepsilon/L, \quad \varepsilon = \lceil \ell/2 \rceil, \ldots, \ell. \tag{4.2}$$

Полагая в (4.1) $\varepsilon = \ell$, заключаем, что скрытое значение $\xi(\bar{x})$ выходит за пределы диапазона наблюдаемых значений $\big[\xi_X^{(1)}, \xi_X^{(\ell)}\big]$ с вероятностью $\frac{2}{L}$. Для предсказания $\xi(\bar{x})$ с надёжностью η достаточно иметь $\frac{1}{\eta} - 1$ объектов в случае односторонней оценки, и примерно вдвое больше, $\frac{2}{\eta} - 1$, для двусторонней. В частности, 19 объектов достаточно для получения верхней оценки с надёжностью 0.95.

§4.2 Доверительный интервал для медианы

§4.3 Критерий Вилкоксона–Манна–Уитни

§4.4 Критерий знаков

§4.5 Критерий серий

Литература

[1] *Алимов Ю. И.* Альтернатива методу математической статистики. — Знание, 1980.

[2] *Беляев Ю. К.* Вероятностные методы выборочного контроля. — М.: Наука, 1975.

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

[4] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

[5] *Вовк В. Г., Шейфер Г. Р.* Вклад А. Н. Колмогорова в основания теории вероятностей // *Проблемы передачи информации.* — 2003. — Т. 39, № 1. — С. 24–35.

[6] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.
http://www.ccas.ru/frc/papers/voron04mpc.pdf.

[7] *Гаек Я., Шидак З.* Теория ранговых критериев. — М.: Наука, 1971.

[8] *Гоппа В. Д.* Введение в алгебраическую теорию информации. — М.: Наука, 1995.

[9] *Гуров С. И.* Точечная оценка вероятности 0-события // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 22–25.
http://www.mmro.ru.

[10] *Кобзарь А. И.* Прикладная математическая статистика. — М.: Физматлит, 2006.

[11] *Колмогоров А. Н.* Комбинаторные основания теории информации и исчисления вероятностей // *Успехи математических наук.* — 1983. — Т. 38, № 4. — С. 27–36.

[12] *Колмогоров А. Н.* Теория информации и теория алгоритмов / Под ред. Ю. В. Прохоров. — М.: Наука, 1987. — 304 с.

[13] *Орлов А. И.* Эконометрика: Учебник для вузов. — М.: Экзамен, 2003. — 576 с.

[14] *Орлов А. И.* Нечисловая статистика. — М.: МЗ-Пресс, 2004.

[15] *Смирнов Н. В.* Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // *Бюлл. Московского ун-та, серия А.* — 1939. — № 2. — С. 3–14.

[16] *Успенский В. А.* Четыре алгоримтических лица случайности. — М.: Изд-во МЦНМО, 2009. — 48 с.

[17] *Шень А. Х.* Частотный подход к определению понятия случайной последовательности // Семиотика и информатика. — М.: ВИНИТИ, 1982. — Т. 18. — С. 14–42.

[18] *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. — М: Финансы и статистика, 1988.

[19] *Bauer M., Godreche C., Luck J. M.* Statistics of persistent events in the binomial random walk: Will the drunken sailor hit the sober man? // *J.STAT.PHYS.* — 1999. — Vol. 96. — P. 963.
http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/9905252.

[20] *Chvátal V.* The tail of the hypergeometric distribution // *Discrete Mathematics.* — 1979. — Vol. 25, no. 3. — Pp. 285–287.

[21] *Efron B.* The Jackknife, the Bootstrap, and Other Resampling Plans. — SIAM, Philadelphia, 1982.

[22] *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // 14th International Joint Conference on Artificial Intelligence, Palais de Congres Montreal, Quebec, Canada. — 1995. — Pp. 1137–1145.

http://citeseer.ist.psu.edu/kohavi95study.html.

[23] *Langford J.* Quantitatively Tight Sample Complexity Bounds: Ph.D. thesis / Carnegie Mellon Thesis. — 2002.
http://citeseer.ist.psu.edu/langford02quantitatively.html.

[24] *Langford J., McAllester D.* Computable shell decomposition bounds // Proc. 13th Annu. Conference on Comput. Learning Theory. — Morgan Kaufmann, San Francisco, 2000. — Pp. 25–34.
http://citeseer.ist.psu.edu/langford00computable.html.

[25] *Lugosi G.* On concentration-of-measure inequalities. — Machine Learning Summer School, Australian National University, Canberra. — 2003.
http://citeseer.ist.psu.edu/lugosi98concentrationmeasure.html.

[26] *Mullin M., Sukthankar R.* Complete cross-validation for nearest neighbor classifiers // Proceedings of International Conference on Machine Learning. — 2000. — Pp. 639–646.
http://citeseer.ist.psu.edu/309025.html.

[27] *von Mises R.* Mathematical Theory of Probability and Statistics. — New York, Academic Press, 1964.