МЕТОДЫ АНАЛИЗА ДАННЫХ
(Вычислительные системы)
1985 год Выпуск III

УДК 519

О НЕПАРАМЕТРИЧЕСКОЙ СТАТИСТИКЕ

К.Ф. Самохвалов

Статья носит методологический характер. Она ориентирована на инженеров. Ее цель - подчеркнуть те особенности непараметрической статистики, знать которые - в качестве общих ориентиров - полезно прикладнику, выбирающему подходящий для своих целей статистический аппарат. С этой точки зрения статья может рассматриваться как некий комментарий к первой (методологической) главе книги Ф.П.Тарасенко "Непараметрическая статистика" [1].

Сразу заметим, что на протяжении всей работы статистика понимается как прикладная наука, и поэтому ниже уделяется внимание не только формулам, но и тому нематематическому содержанию, которое эти формулы могут выражать при разумном соглашении о способах их интерпретации. Такое соглашение определяет область применимости статистики как прикладной дисциплины, и, если не гнаться за чрезмерной общностью, его можно описать следующим образом.

Пусть R^1 - евклидово пространство размерности 1. Если R^m - множество значений какой-то ненаблюдаемой (но интересующей нас) m-мерной величины ξ, то R^m назовем *пространством возможных состояний* изучаемого объекта. Если R^n - множество значений наблюдаемой n-мерной величины η, то R^n будем называть *пространством выборок (длины n)*. Пусть, далее, Φ_{mn} - класс всех функций $F(x_1,\ldots, x_{m+n})$, определенных на R^{m+n} и таких, что:

1) $F(x_1,\ldots, x_{m+n})$ всюду непрерывна справа по каждому из $m+n$ своих аргументов;

2) $0 \le F(x_1,\ldots, x_{m+n}) \le 1$;

3) $F(-\infty, x_2,\ldots, x_{m+n}) = \ldots = F(x_1,\ldots, x_{m+n-1}, -\infty) = 0$;

4) $F(+\infty,\ldots,+\infty) = 1$;

5) $\Delta_1\Delta_2\ldots\Delta_{m+n}F(x_1,\ldots,x_{m+n})\geq 0$ для любых $h_1,\ldots,h_{m+n}\geq 0$. (Здесь $\Delta_i G(x_1,\ldots,x_{m+n}) = G(x_1,\ldots,x_{i-1},x_i+h_i,x_{i+1},\ldots,x_{m+n}) - G(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_{m+n})$, G - произвольная функция из R^{m+n} в R^1.) Если $F(x_1,\ldots,x_{m+n})\in\Phi_{mn}$, то $F(x_1,\ldots,x_{m+n})$ назовем функцией (совместного) распределения (случайных векторов ξ и η) (ср. [2], с.79 или [3], с.95). Всякое отображение вида $\varphi: X_1\times\ldots\times X_k\to Y$ будем называть вероятностным функционалом, если хотя бы одно из множеств $X_1,\ldots,X_k$ есть Φ_{mn} для подходящих m,n (в остальном природа $X_1,\ldots,X_k,Y$ произвольна).

Подчеркнем, что, согласно определению, вероятностный функционал - это просто некий математический объект и только. Никаких идей, связанных с пониманием вероятности как меры случайности, с этим названием не ассоциируется. Точно так же не ассоциируется никаких идей о стохастической природе величин ξ и η с названием "функция распределения случайных векторов ξ и η".

Рассматриваемое соглашение состоит в том, чтобы считать относящимися к статистике лишь те эмпирические проблемы, которые удается математически истолковать как задачи на исследование тех или иных свойств тех или иных вероятностных функционалов. И наоборот, относится ли данная математическая задача, поставленная в терминах определенных вероятностных функционалов (как математических объектов), к статистике (в нашем понимании), зависит от того, удается или нет истолковать эти функционалы в терминах каких-то практических полезностей; да еще и так, чтобы вся задача в целом при таком истолковании получила некий практический смысл*).

Ясно, что круг статистических задач, понимаемых таким образом, очертить заранее вряд ли возможно: мало ли какие исследова -

*) Обычно такого рода истолкования в ту и другую сторону основываются на частотной интерпретации понятия вероятности. Обычно именно к такой интерпретации элементов класса Φ_{mn} привязываются практические смыслы используемых функционалов (например, функционал среднего риска или еще какой-нибудь в этом же роде). Сама эта частотная интерпретация с методологической точки зрения не вполне удовлетворительна (при ней статистические утверждения становятся принципиально нефальсифицируемыми); однако следует иметь в виду, что один и тот же вероятностный функционал может иметь несколько различных содержательных истолкований, среди которых могут найтись и истолкования, лишенные недостатков, связанных с частотной интерпретацией вероятности.

ния, мало ли каких свойств, мало ли каких вероятностных функционалов могут оказаться полезными в практической деятельности при соответствующем истолковании. Поэтому в настоящее время в литературе рассматриваются лишь некоторые исторически возникшие классы статистических постановок. Из них нам достаточно ограничиться следующим классом:

1. Задано m (задано пространство R^m возможных состояний изучаемого объекта).

2. Задано n (задано пространство R^n выборок фиксированной длины n).

3. Задан непустой подкласс $\mathcal{F}_{mn}$ класса Φ_{mn}. Выражение $F(x_1,\dots,x_{m+n}) \in \mathcal{F}_{mn}$ интерпретируется: $F(x_1,\dots,x_{m+n})$ заведомо не является *истинной* функцией распределения величин ξ и η, что бы ни означало при этом слово "истинная". Класс $\mathcal{F}_{mn}$ называется *априорной статистической информацией*. (Остальные элементы постановки – тоже, конечно, априорная информация, но она не называется статистической.)

4. Задано некоторое множество D, каждый элемент d которого интерпретируется как одно из всех предполагаемых в данной конкретной задаче *решений*, ради получения коих вообще затевается данная постановка. Природа элементов множества D может быть самой различной: d могут быть высказываниями, числами, действиями, функциями, операторами – чем угодно, лишь бы выполнялось то, о чем будет сказано чуть ниже.

5. Задан класс Δ функций вида δ: $R^n \to D$. Любая δ из Δ называется *решающим правилом*. Выражение $\delta(\bar{y}) = d$ интерпретируется как высказывание: если наблюдаемый результат есть $\bar{y}$, то решающее правило δ предписывает нам принять решение d.

6. На $\Phi_{mn} \times \Delta$ задан вероятностный функционал $\rho(F,\delta)$ вида $\rho: \Phi_{mn} \times \Delta \to \{и,л\}$. Выражение $\rho(F,\delta) = и$ интерпретируется как высказывание: если F – истинная функция распределения векторов ξ и η, то решающее правило δ приемлемо для эксплуатации с точки зрения подразумеваемых в данной конкретной постановке практических целей. Выражение $\rho(F,\delta) = л$ интерпретируется аналогичным образом с заменой "приемлемо" на "не приемлемо".

Упомянутое выше ограничение на D сводится к тому, чтобы природа элементов d из D не делала бессмысленной указанную интерпретацию ρ.

Требуется:

а) определить, истинно или ложно высказывание

$$(\exists \delta \in \Delta)(\forall F \in \Phi_{mn})(F \in \mathcal{F}_{mn} \Rightarrow \rho(F,\delta) = \text{и}); \qquad (1)$$

б) если высказывание (I) истинно, то найти хотя бы одно δ_0 такое, что

$$(\forall F \in \Phi_{mn})(F \in \mathcal{F}_{mn} \Rightarrow \rho(F,\delta_0) = \text{и}); \qquad (2)$$

в) если высказывание (I) ложно, то объявить, что данная постановка не имеет решения.

Мы видим, что любая постановка рассматриваемого класса полностью определяется набором своих исходных данных, а именно упорядоченной шестеркой $\langle m,n,\mathcal{F}_{mn},D,\Delta,\rho\rangle$, где все элементы имеют описанные выше смыслы. Поэтому каждую такую постановку мы просто отождествляем с соответствующей упорядоченной шестеркой.

Пусть $S' = \langle m',n',\mathcal{F}'_{m'n'},D',\Delta',\rho'\rangle$ и $S'' = \langle m'',n'',\mathcal{F}''_{m''n''},D'',\Delta'',\rho''\rangle$ – произвольные постановки из нашего класса. Будем говорить, что S'' имеет более полную априорную статистическую информацию, чем S' (символически $S' \prec S''$, если и только если $\mathcal{F}'_{m'n'} \supseteq \mathcal{F}''_{m''n''}$).

Отношение $\prec$ является, очевидно, частичным порядком на рассматриваемом классе постановок, а его название ("порядок по полноте априорной статистической информации") мотивируется следующим обстоятельством: если $\mathcal{F}'_{m'n'} \supseteq \mathcal{F}''_{m''n''}$ и высказывание (I) истинно для некоторого ρ при $\mathcal{F}_{mn} = \mathcal{F}'_{m'n'}$, то и подавно для этого же ρ истинно высказывание (I) при $\mathcal{F}_{mn} = \mathcal{F}''_{m''n''}$.

Минимальные элементы в $\prec$ – это задачи, в которых $\mathcal{F}_{mn} = \Phi_{mn}$; максимальные – те задачи, у которых $\mathcal{F}_{mn}$ суть единичные подклассы класса Φ_{mn}.

Упорядочив все наши постановки по полноте априорной статистической информации, мы можем, кроме того, еще и разбить их на два подкласса – подкласс параметрических постановок и подкласс непараметрических. Критерий этого разбиения таков: постановка является параметрической, если соответствующая ей априорная статистическая информация $\mathcal{F}_{mn}$ есть параметризуемый класс; постановка является непараметрической, если $\mathcal{F}_{mn}$ – непараметризуемый класс. Класс функций $\mathcal{F}_{mn}$ будем называть

п а р а м е т р и з у е м ы м, если и только если существует гомеоморфное вложение $\mathcal{F}_{mn}$ в R^1 для подходящего l, и н е п а р а м е т р и з у е м ы м – в противном случае. При этом топология на $\mathcal{F}_{mn}$ индуцируется метрикой

$$\mu(F_1,F_2)=\sup_{-\infty<x_1,\dots,x_{m+n}<\infty}|F_1(x_1,\dots,x_{m+n})-F_2(x_1,\dots,x_{m+n})|,\ F_1,F_2\in\mathcal{F}_{mn};$$

топология на R^1 – евклидовой метрикой*).

Как связано упорядочение по полноте априорной статистической информации с разбиением на параметрические и непараметрические постановки? В общем случае только так, что все минимальные в $\prec$ постановки принадлежат заведомо классу непараметрических, а все максимальные в $\prec$ – заведомо принадлежат классу параметрических. Промежуточные постановки могут принадлежать тому или другому классу. Каждый раз это надо определять отдельно, и, вообще говоря, это задача непростая, ибо это есть задача установления существования упомянутого гомеоморфизма.

Но если нет тесной связи между классификацией на параметрические и непараметрические постановки и упорядочением их по полноте априорной статистической информации, то зачем вообще нужна эта классификация? (Мы считаем ясным вопрос, зачем нам может пригодиться знание порядка $\prec$.) Конечно, было бы неплохо, если бы мы имели некоторый общий результат типа: если постановка $\langle n,\dots,\rho\rangle$ непараметрическая, то она имеет решение (соответствующее высказывание (I) истинно) тогда и только тогда, когда ρ принадлежит некоторому классу R_1 функционалов, а если $\langle n,\dots,\rho\rangle$ – параметрическая, то она имеет решение тогда и только тогда, когда ρ принадлежит классу R_2 функционалов. Было бы уже совсем хорошо, если бы вдобавок к этому мы умели решать любую постановку с ρ из R_1 и ρ из R_2. Понятно, что мы не имеем ни того, ни другого результата. Нет даже никаких сколь-нибудь обнадеживающих косвенных

*) Иногда говорят, что $\mathcal{F}_{mn}$ параметризуем, если существует просто вложение (инъекция) $\mathcal{F}_{mn}$ в R^1 для подходящего l. Так определять параметризуемость бесполезно, ибо при таком определении любой подкласс $\mathcal{F}_{mn}$ класса Φ_{mn} является параметризуемым. В самом деле, класс Φ_{mn} имеет мощность континуума, так как все F из Φ_{mn} непрерывны справа; следовательно, инъекция из $\mathcal{F}_{mn}\subseteq\Phi_{mn}$ в R^1 всегда существует при $l\geq 1$.

свидетельств в пользу получения подобных результатов в будущем. Все, что мы умеем делать с рассматриваемой классификаций, так это извлекать из нее в отдельных случаях подсказки насчет предполагаемых методов решения классифицируемых задач или, того меньше, констатировать, что вот такими-то конкретными приемами мы не можем решить рассматриваемую постановку задачи. Здесь уместны примеры.

ПРИМЕР I. $S_1 = \langle 1,1, \mathcal{F}^1_{11}, D_1, \Delta_1, \rho_1 \rangle$, $\mathcal{F}^1_{11} = \{F_1(x,y)\}$; F_1 такова, что величины

$$p_1(x) \doteq \int_{-\infty}^{+\infty} \frac{\partial^2 F_1(x,y)}{\partial x \cdot \partial y}\, dy$$

и

$$w_1(y/x) \doteq \frac{\partial^2 F_1(x,y)}{\partial x \cdot \partial y} \cdot \frac{1}{p_1(x)}$$

имеют смысл;

$$\rho_1(F,\delta) = \begin{cases} \text{и, если имеет смысл и верно равенство} \\ \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} dy \right) \left(\int_{-\infty}^{+\infty} L_1(x, \delta(y)) \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} \cdot \frac{dy}{\int_{-\infty}^{+\infty} \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} dy} \right) dx = \\ = \inf_{\delta^* \in \Delta_1} \int_{-\infty}^{+\infty} p_1(x) \left(\int_{-\infty}^{+\infty} L_1(x, \delta^*(y)) w_1(y/x) dy \right) dx; \\ \text{л – в любом другом случае.} \end{cases}$$

Здесь $L_1: R^1 \times D_1 \to R^1$. Функция $L_1(x,d) = r$ означает: ущерб, который несет статистик, приняв решение $d \in D_1$, когда истинное значение величины ξ есть x, измеряется числом r. Иными словами, L_1 – функция потерь.

С таким ρ_1 мы бы встретились, если бы имели дело с байесовой задачей.

Эта постановка является максимальной в $\prec$. Она параметрическая. Решение (при обычных ограничениях на D_1, Δ_1, L_1, которые мы здесь не уточняем) – байесово.

ПРИМЕР 2. $S_2 = \langle 1, 1, \mathcal{F}^2_{11}, D_2, \Delta_2, \rho_2 \rangle$. $\mathcal{F}^2_{11} = \{F_a(x,y)\}_{a \in R^1}$.

Каждая F_a из $\mathcal{F}^2_{11}$ такова, что величины

$$p_a(x) \Leftrightarrow \int_{-\infty}^{+\infty} \frac{\partial^2 F_a(x,y)}{\partial x \cdot \partial y}\, dy$$

и

$$w_a(y/x) \Leftrightarrow \frac{\partial^2 F_a(x\ y)}{\partial x \cdot \partial y} \cdot \frac{1}{p_a(x)}$$

имеют смысл и непрерывны по a;

$$\rho_2(F,\delta) = \begin{cases} \text{и, если имеет смысл и верно неравенство} \\ \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} dy \right) \left(\int_{-\infty}^{+\infty} L_2(x, \delta(y)) \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} \cdot \frac{dy}{\int_{-\infty}^{+\infty} \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} dy} \right) dx \leq \\ \leq \inf_{\delta^* \in \Delta_2} \max_{a \in R^1} \int_{-\infty}^{+\infty} p_a(x) \left(\int_{-\infty}^{+\infty} L_2(x, \delta^*(y)) w_a(y/x) dy \right) dx; \\ \Lambda \text{ – в любом другом случае.} \end{cases}$$

Здесь L_2: $R^1 \times D_2 \to R^1$ – функция потерь.

С таким ρ_2 мы бы встретились, если бы имели дело с минимаксной задачей.

Эта постановка параметрическая. При обычных (неуточняемых здесь) ограничениях на D_2, Δ_2, L_2 она имеет известное решение.

ПРИМЕР 3. $S_3 = \langle 1, 1, \mathcal{F}^3_{11}, D_3, \Delta_3, \rho_3 \rangle$. $F(x,y) \in \mathcal{F}^3_{11}$ тогда и только тогда, когда величины

$$p_F(x) \Leftrightarrow \int_{-\infty}^{+\infty} \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} \cdot dy$$

и

$$w_F(y/x) \Leftrightarrow \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} \cdot \frac{1}{p_F(x)}$$

имеют смысл;

$$\rho_3(F,\delta)=\begin{cases} \text{и, если имеет смысл и верно неравенство} \\ \int_{-\infty}^{+\infty} p_F(x)\left(\int_{-\infty}^{+\infty} L_3(x,\delta(y))w_F(y/x)dy\right)dx \le \\ \le \inf_{\delta^*\in\Delta_3}\max_{F^*\in\mathcal{F}_{11}^3}\int_{-\infty}^{+\infty} p_{F^*}(x)\left(\int_{-\infty}^{+\infty} L_3(x,\delta^*(y))w_{F^*}(y/x)dy\right)dx; \\ \Lambda \text{ - в любом другом случае.} \end{cases}$$

Бальд [4] в 1950 году показал, что при достаточно широких ограничениях на L_3, Δ_3 и D_3 высказывание (I) для ρ_3 будет истинным, т.е. что рассматриваемая постановка имеет решение. Но это решение мы не всегда умеем находить. Разумеется, если $\mathcal{F}_{11}^3$ мы могли бы гомеоморфно отобразить на R^1, то мы решили бы эту задачу, сведя ее к предыдущей. Но такого гомеоморфизма не существует. Поэтому нужны какие-то другие методы решения, отличающиеся от тех обычных, которыми мы решаем предыдущую постановку. (Мы, правда, можем, как уже было замечено, отобразить $\mathcal{F}_{11}^3$ взаимно-однозначно на R^1, но при этом обязательно $\iint$ не будет зависеть от a непрерывно; и поиски экстремума по a для $\iint$ становятся проблематичными.) Это как раз тот случай, когда из факта непараметричности постановки мы извлекаем указание на то, что ее решение не следует искать привычными методами.

Только что приведенные три примера постановок из класса, который мы рассматриваем, никоим образом, конечно, не исчерпывают всего содержания этого класса. Например, многие постановки из [I] можно было бы переизложить в наших терминах, не выходя при этом за пределы рассматриваемого класса или за пределы его очевидных расширений. Однако автор сознательно ограничился, так сказать, "классическими" случаями, следуя объявленной в начале статьи цели: п р о я с н и т ь, а не обогатить суть дела.

Л и т е р а т у р а

I. ТАРАСЕНКО Ф.П. Непараметрическая статистика. -Томск: 1976. - 294 с. (Томский ун-т).

2. ХЕННЕКЕН П.Л., ТОРТА А. Теория вероятностей и некоторые ее приложения. -М.: Наука, 1974. - 472 с.

3. КРАМЕР Г. Математические методы статистики.-М.: Мир,1975. - 648 с.

4. WALD A. Statistical Decision Functions.- N.Y.: John Wiley Inc.,1950.- 220 p.

Поступила в ред.-изд.отд.
13 августа 1985 года