Доклады Академии Наук СССР
1950. Том LXXV, № 2

МАТЕМАТИКА

Е. Б. ДЫНКИН

О ДОСТАТОЧНЫХ И НЕОБХОДИМЫХ СТАТИСТИКАХ ДЛЯ СЕМЕЙСТВА РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

(Представлено академиком А. Н. Колмогоровым 14 IX 1950)

В предлагаемой работе исследуется общая задача вычисления достаточных статистик (см. (1), стр. 530—531, а также (2)) для заданного семейства одномерных распределений вероятностей. Мы рассматриваем статистики, принимающие векторные значения; очевидно, при этом отпадает необходимость изучать отдельно достаточные системы статистик. Определение 2 вводит новое понятие необходимой статистики. Если знание любой достаточной статистики дает достаточный материал для того, чтобы оценивать неизвестный параметр, достаточный в том смысле, что знание полного результата наблюдений не добавляет к этому материалу ничего существенного, то знание всякой необходимой статистики необходимо для того, чтобы не происходила существенная потеря информации. Для каждого семейства распределений при фиксированном объеме выборки существует единственная с точностью до эквивалентности необходимая и достаточная статистика. Теорема 1 дает путь для ее вычисления. Вместе с тем теорема 1 вводит важное понятие ранга семейства распределений. Для семейств бесконечного ранга понятие достаточной статистики является бесплодным. Все семейства конечного ранга выделяются теоремой 2 (частным случаем последней является теорема Дармуа (3)). В математической статистике важную роль играют семейства распределений, получающиеся из некоторых распределений линейными преобразованиями прямой линии. Специальному исследованию таких семейств распределений посвящены теоремы 3 и 4*.

Мы рассматриваем семейство $\mathfrak{S}$ одномерных распределений $P_\theta(A)$ (параметр θ пробегает некоторое вспомогательное множество S). При этом мы предполагаем, что в некотором интервале Δ (конечном или бесконечном) каждое распределение $P_\theta(A)$ задается плотностью $p(x, \theta)$, являющейся положительной кусочно гладкой функцией ** (условие регулярности). Пусть производится серия из n независимых испытаний, причем результаты каждого испытания подчинены одному и тому же закону распределения из класса $\mathfrak{S}$. Возможные результаты образуют n-мерное пространство R^n.

Определение 1. Достаточной статистикой для семейства распределений $\mathfrak{S}$ в интервале по выборке объема n называется всякая функция $\chi(x_1, \ldots, x_n)$, принимающая значения из некоторого векторного пространства R^m, определенная для всех $x_1 \in \Delta, \ldots, x_n \in \Delta$ и удовлетворяющая в этой области соотношению

$$p(x_1, \theta)\, p(x_2, \theta) \ldots p(x_n, \theta) = \overline{p}(\chi(x_1, x_2, \ldots, x_n), \theta)\, q(x_1, x_2, \ldots, x_n).$$

* Эти теоремы дают, в частности, ответ на поставленный А. Н. Колмогоровым вопрос о том, какой вид могут иметь достаточные статистики для указанных семейств распределений.

** Функцию $p(x)$ мы называем кусочно гладкой в Δ, если существует область $G \subset \Delta$ такая, что $\overline{G} = \overline{\Delta}$ и $dp(x)/dx$ существует и непрерывна в G.

Положим сокращенно $(x_1, x_2, \ldots, x_n) = x$. Пусть $\chi_1(x)$ и $\chi_2(x)$ — две функции, определенные в некоторой области G пространства R^n. Условимся говорить, что χ_2 подчинена χ_1, если из $\chi_1(x') = \chi_1(x'')$ вытекает $\chi_2(x') = \chi_2(x'')$. Будем называть функции χ_1 и χ_2 эквивалентными, если каждая из них подчинена другой. Статистика $\varepsilon(x) = x$ достаточна для любой системы распределений. Мы будем говорить, что статистика $\chi(x)$ тривиальна в области G, если она эквивалентна $\varepsilon(x)$ в некоторой области $\tilde{G} \subset G$.

Определение 2. Необходимой статистикой для семейства распределений $\mathfrak{S}$ в интервале Δ по выборке объема n называется всякая функция, определенная для $x_1 \in \Delta$, $x_2 \in \Delta, \ldots, x_n \in \Delta$ и подчиненная в этой области любой достаточной статистике.

Теорема 1. *Пусть θ_0 — произвольный элемент S. Положим $g_x(\theta) = \ln p(x, \theta) - \ln p(x, \theta_0)$ и обозначим через L минимальное линейное пространство функций, определенных на Δ, содержащее константы и содержащее при любом $\theta \in S$ функцию $g_x(\theta)$. Пусть размерность L равна $r+1$ (не исключено, что $r = \infty$). Тогда:*

А. *Для всякого конечного $n \leqslant r$ произвольная достаточная статистика для семейства $\mathfrak{S}$ в интервале Δ по выборке объема n тривиальна.*

Б. *Если функции $1, \varphi_1(x), \varphi_2(x), \ldots, \varphi_r(x)$ образуют базис в L, то при любом $n \geqslant r$ система функций*

$$\chi_i(x_1, x_2, \ldots, x_n) = \varphi_i(x_1) + \varphi_i(x_2) + \ldots + \varphi_i(x_n)$$
$$(i = 1, 2, \ldots, r; \quad x_1, \ldots, x_n \in \Delta)$$

функционально независима и образует необходимую и достаточную статистику для семейства $\mathfrak{S}$ в интервале Δ по выборке объема n.

Число r, определяемое теоремой 1, мы назовем рангом системы распределений $\mathfrak{S}$ в интервале Δ.

Теорема 2. *Для того чтобы система распределений $\mathfrak{S}$ имела конечный ранг в интервале Δ, необходимо и достаточно, чтобы плотность $p(x, \theta)$ представилась в виде*

$$p(x, \theta) = \exp\left(\sum_{i=1}^{r} \varphi_i(x)\, c_i(\theta) + c_0(\theta) + \varphi_0(x)\right) \quad (x \in \Delta, \theta \in S),$$

где $\varphi_1(x), \ldots, \varphi_r(x)$ кусочно гладки в интервале Δ. Если при этом $1, \varphi_1(x), \ldots, \varphi_r(x)$ и $1, c_1(\theta), \ldots, c_r(\theta)$ линейно независимые системы функций, то ранг $\mathfrak{S}$ равен r и для $n \geqslant r$ система функций

$$\chi_i(x_1, \ldots, x_n) = \varphi_i(x_1) + \varphi_i(x_2) + \ldots + \varphi_i(x_n)$$
$$(i = 1, 2, \ldots, r; \qquad x_1, \ldots, x_n \in \Delta)$$

функционально независима и образует необходимую и достаточную статистику для $\mathfrak{S}$ по выборке объема n.

Теорема 3. *Пусть каждому $\theta \in S$ сопоставлена функция распределения $F(x, \theta)$. Пусть семейство $\mathfrak{S}$ этих распределений удовлетворяет в интервале Δ условию регулярности. Тогда:*

А. *Если для некоторого $\delta > 0$ семейство $\mathfrak{S}_1$ распределений $F(x - \alpha, \theta)$ $(|\alpha| < \delta, \theta \in S)$ имеет конечный ранг в Δ, то плотность $p(x, \theta)$ представляется в виде*

$$p(x, \theta) = \exp\left(\sum_{i=1}^{s} c_i(\theta)\, x^{n_i} e^{\mu_i x}\right) \quad (x \in \Delta, \theta \in S), \tag{1}$$

где μ_i — комплексные, n_i — целые постоянные, $c_i(\theta)$ — функции, принимающие комплексные значения.

Б. *Если Δ не содержит нуля и для некоторого $p > 1$ семейство $\mathfrak{S}_2$ распределений $F(x/\sigma, \theta)$ $(\theta \in S, 1/p < \sigma < p)$ имеет конечный ранг*

в Δ, *то плотность* $p(x, \theta)$ *представляется в виде*

$$p(x, \theta) = \exp\left(\sum_{i=1}^{s} c_i(\theta)(\ln|x|)^{n_i}|x|^{\mu_i}\right) \quad (x \in \Delta, \theta \in S) \tag{2}$$

(μ_i, n_i, $c_i(\theta)$ *как в формуле* (1)).

B. *Если для некоторых* $\delta > 0$ *и* $p > 1$ *семейство* $\mathfrak{S}_3$ *распределений* $F\left(\frac{x-\alpha}{\sigma}, \theta\right)$ $\left(\theta \in S, |\alpha| < \delta, \frac{1}{p} < \sigma < p\right)$ *имеет конечный ранг в* Δ, *то*

$$p(x, \theta) = \exp Q(x, \theta) \quad (x \in \Delta, \theta \in S), \tag{3}$$

где $Q(x, \theta)$ *— многочлен относительно* x *с комплексными коэффициентами, зависящими от* θ.

Необходимая и достаточная статистика для семейства $\mathfrak{S}_1$ *может быть образована из функций вида* $x_1^k e^{\lambda x_1} + x_2^k e^{\lambda x_2} + \ldots + x_n^k e^{\lambda x_n}$; *для семейства* $\mathfrak{S}_2$ *— из функций вида* $|x_1|^\lambda (\ln|x_1|)^k + |x_2|^\lambda (\ln|x_2|)^k + \ldots + |x_n|^\lambda (\ln|x_n|)^k$ *и для семейства* $\mathfrak{S}_3$ *— из функций вида* $x_1^k + x_2^k + \ldots + x_n^k$ (k *— целые,* λ *— комплексные постоянные*) **.

Для многих важных распределений задающая их функция плотности вероятностей равна нулю вне некоторого интервала. Если рассматривать в качестве интервала Δ всю прямую, то условие регулярности здесь не выполняется и теорема 3 непосредственно не применима. Вместо нее может быть использована следующая теорема.

Теорема 4. *Пусть функция* $p(x, \theta)$ *при каждом* θ *из* S *представляет плотность вероятностей некоторого одномерного распределения. Пусть при всех* $\theta \in S$ $p(x, \theta)$ *положительна и кусочно гладка относительно* x *в некотором интервале* Δ *и равна нулю вне* Δ. *Обозначим через* $\mathfrak{S}_1$ *семейство распределений* $p(x - \alpha, \theta)$ ($\theta \in S$, $-\infty < \alpha < +\infty$), *через* $\mathfrak{S}_2$ *— семейство распределений* $\frac{1}{\sigma} p\left(\frac{x}{\sigma}, \theta\right)$ ($\theta \in S, 0 < \sigma < \infty$) *и через* $\mathfrak{S}_3$ *— семейство распределений* $\frac{1}{\sigma} p\left(\frac{x-\alpha}{\sigma}, \theta\right)$, ($\theta \in S, -\infty < \alpha < +\infty, 0 < \sigma < +\infty$). *Для того чтобы семейства* $\mathfrak{S}_1$, $\mathfrak{S}_2$, $\mathfrak{S}_3$ *имели конечный ранг на всей прямой, необходимо и достаточно, чтобы для* $\theta \in S$, $x \in \Delta$ *плотность* $p(x, \theta)$ *представлялась, соответственно, в виде* (1), (2) *или* (3) (*в случае* $\mathfrak{S}_2$ *дополнительно предполагается, что интервал* Δ *не содержит нуля*).

Для семейств $\mathfrak{S}_1$ *и* $\mathfrak{S}_3$ *необходимая и достаточная статистика дается системой функций, указанных в теореме* 3, *если* $\Delta = (-\infty, +\infty)$. *Эта система функций должна быть пополнена функцией* $\min(x_1, x_2, \ldots, x_n)$ *в случае* $\Delta = (a, +\infty)$ (a *конечно*) *и функцией* $\max(x_1, x_2, \ldots, x_n)$ *в случае* $\Delta = (-\infty, b)$ (b *конечно*). *Наконец, в случае* $\Delta = (a, b)$, *где* a *и* b *конечны, к системе функций теоремы* 3 *должны быть присоединены обе функции* $\min(x_1, x_2, \ldots, x_n)$ *и* $\max(x_1, x_2, \ldots, x_n)$. *Аналогично, для семейства* $\mathfrak{S}_2$ *необходимая и достаточная статистика совпадает с указанной в теореме* 3, *если* $\Delta = (0, +\infty)$ *или* $\Delta = (-\infty, 0)$; *получается из указанной в теореме* 3 *статистики присоединением* $\min(x_1, x_2, \ldots, x_n)$, *если* $\Delta = (a, +\infty)$ $(a > 0)$ *или* $\Delta = (-\infty, b)$ $(b < 0)$; *присоединением* $\max(x_1, x_2, \ldots, x_n)$, *если* $\Delta = (0, a)$ $(0 < a < +\infty)$ *или* $\Delta = (b, 0)$ $(-\infty < b < 0)$, *и присоединением* $\max(x_1, \ldots, x_n)$ *и* $\min(x_1, \ldots, x_n)$, *если* $\Delta = (a, b)$, *где* a *и* b *конечны и отличны от нуля.*

Примеры.

1. Гауссовская плотность $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ имеет вид (3). В со-

* Нетрудно указать тот набор пар (k, λ), которому отвечает система функций, дающих необходимую и достаточную статистику. Из-за недостатка мест мы опускаем формулировку соответствующего правила.

ответствии с теоремой 3 пара функций $\left(\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} x_i^2\right)$ является необходимой и достаточной статистикой для семейства распределений $\frac{1}{\sigma} p\left(\frac{x-\alpha}{\sigma}\right)$.

2. Исследуем семейство распределений, задаваемое на положительной полуоси плотностями

$$p(x, \beta, \gamma, \mu) = C x^{\beta} e^{-\gamma x^{\mu}}. \tag{4}$$

Если μ известно, то необходимая и достаточная статистика задается парой функций $\left(\sum_{i=1}^{n} x_i^{\mu},\ \sum_{i=1}^{n} \ln x_i\right)$ при неизвестных β и γ *, функцией $\sum_{i=1}^{n} \ln x_i$ при неизвестном β и известном γ и функцией $\sum_{i=1}^{n} x_i^{\mu}$ при неизвестном γ и известном β. Если μ неизвестно, то ранг семейства (4) в любом интервале Δ бесконечен и, значит, нетривиальные достаточные статистики отсутствуют (последнее было впервые доказано Пинскером).

3. Для семейства $p(x-\alpha, \beta, \gamma, \mu)$, где $p(x, \beta, \gamma, \mu)$ определено формулой (4), при известном μ необходимая и достаточная статистика получается, согласно теореме 4, присоединением к функциям, указанным в примере 2, функции $\min(x_1, x_2, \ldots, x_n)$. При неизвестном μ ранг рассматриваемого семейства бесконечен.

4. Плотность

$$p(x, \theta, \alpha_1, \sigma_1, \alpha_2, \sigma_2) = \frac{\theta}{\sqrt{2\pi}\sigma_1} e^{-(x-\alpha_1)^2/2\sigma_1^2} + \frac{(1-\theta)}{\sqrt{2\pi}\sigma_2} e^{-(x-\alpha_2)^2/2\sigma_2^2}$$

$(0 < \theta < 1)$ получается смешивание двух гауссовских плотностей. Если хотя бы один из параметров $\theta, \alpha_1, \sigma_1, \alpha_2, \sigma_2$ неизвестен, то соответствующее семейство распределений имеет бесконечный ранг.

5. Для семейства распределений Лапласа $p(x, \alpha, \sigma) = \frac{1}{2\sigma} e^{-|x-\alpha|/\sigma}$ при известном α необходимая и достаточная статистика равна $\sum_{i=1}^{n} |x_i - \alpha|$. При неизвестном α семейство имеет ранг ∞.

6. Плотность $p(x) = e^{-(x+e^{-x})}$ встречается при исследовании предельного поведения максимума m независимых случайных величин при $m \to \infty$. Для семейства распределений $p(x-\alpha)$ необходимая и достаточная статистика равна $\sum_{i=1}^{n} e^{-x_i}$ (ср. теорему 3).

7. Рассмотрим произвольный интервал (a, b) и равномерное распределение в этом интервале. Семейство всех таких распределений получается из любого одного из них линейными преобразованиями. Необходимая и достаточная статистика дается парой функций $\min(x_1, x_2, \ldots, x_n)$ и $\max(x_1, x_2, \ldots, x_n)$ (этот пример был впервые рассмотрен А. Н. Колмогоровым [4]).

Поступило
28 VI 1950

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] H. Cramer, Math. Methods of Statistics, 1946; русск. пер. Г. Крамер, Математические методы статистики, М., 1948. [2] А. Н. Колмогоров, Изв. АН СССР, сер. матем., **14**, № 4 (1950). [3] G. Darmois, C. R., **200**, 1265 (1935). [4] А. Н. Колмогоров, Изв. АН СССР, сер. матем. (1942).

* Выражение «μ известно, β и γ неизвестны» означает, что рассматривается семейство распределений $p(x, \beta, \gamma, \mu)$, где μ фиксировано, а β и γ изменяются в некоторых интервалах.