

1985

Кибернетика

ОТДЕЛЬНЫЙ ОТТИСК

СТАТИСТИЧЕСКИЕ МОДЕЛИ В БИОЛОГИИ

В настоящей статье рассматривается один класс случайных функций, описывающих статистическую зависимость между количеством произведенных испытаний и количеством наступивших при этом разных событий. Такого рода зависимости имеют место не только в биологии, но также в информатике, лингвистике, библиотечном деле, технике. В качестве примеров можно привести статистические зависимости между следующими величинами:

- длиной текста в словоупотреблениях и количеством разных слов;
- количеством книговыдач и количеством разных наименований выданных книг;
- количеством пойманных особей мотыльков и количеством разных их видов (из числа попавших в ловушку);
- количеством отказов элементов некоторой системы (отказавший элемент сразу заменяется исправным) и количеством разных отказавших элементов.

Отметим, что в опыте можно наблюдать лишь некоторую реализацию (траекторию) случайной функции. Ниже будет рассматриваться математическое ожидание случайной функции, графически представляющее собой среднюю кривую, около которой располагаются возможные реализации. Эту кривую можно назвать кривой роста разных событий.

При известном математическом ожидании случайной функции (кривой роста) нетрудно найти количество разных событий, наступающих ровно 0, 1, 2, ..., m раз при X испытаниях, т. е. частотный спектр или, другими словами, статистическую структуру выборки (эта задача решена в математической лингвистике).

В настоящей статье ставится задача: построить систему кривых роста разных событий и на ее основе — систему дискретных распределений, а также рассмотреть практические приложения построенных моделей в биологии.

ЗАВИСИМОСТЬ МЕЖДУ КРИВОЙ РОСТА И ЗАКОНАМИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ НОВЫХ СОБЫТИЙ

Пусть имеется n несовместных событий A_1, A_2, \dots, A_n , составляющих полную группу, вероятности которых соответственно равны p_1, p_2, \dots, p_n ($\sum_{k=1}^n p_k = 1$).

Пусть, далее, производятся независимые испытания, в каждом из которых может наступить любое из n разных событий.

Введем понятие «нового» события. Под новым будем понимать любое из n разных событий при первом его появлении от начала испытаний. Таким образом, число наступивших разных событий равно числу новых событий.

Найдем математическое ожидание числа разных (новых) событий, наступающих при X испытаниях. Оно задается формулой [1]

$$M[Y(X)] = \sum_{k=1}^n [1 - (1 - p_k)^X]. \quad (1)$$

Математическое ожидание числа разных событий, наступивших ровно m раз при X испытаниях (частотный спектр), определится по формуле

$$M[Y_m(X)] = \sum_{k=1}^n C_X^m p_k^m (1 - p_k)^{X-m}. \quad (2)$$

Важной величиной, характеризующей скорость роста новых событий, является вероятность появления какого-нибудь нового события $A^h = \sum_{k=1}^n A_k^h$ в $(X+1)$ -м испытании (равная математическому ожиданию числа новых событий, наступающих при одном $(X+1)$ -м испытании):

$$P(A^h, X+1) = \sum_{k=1}^n (1 - p_k)^X p_k = M[Y(X+1)] - M[Y(X)]. \quad (3)$$

Введем еще одну величину — среднее значение вероятностей новых событий, которые могут наступить при одном X -м испытании (обозначим его \bar{p}_X):

$$\bar{p}_X = \sum_{k=1}^n p_k^2 (1 - p_k)^{X-1}. \quad (4)$$

Тогда накопленная вероятность новых событий, наступивших при X испытаниях, будет равна сумме средних вероятностей \bar{p}_X :

$$\begin{aligned} \bar{F}(X) &= \sum_{i=1}^X \bar{p}_i = 1 - \sum_{k=1}^n p_k (1 - p_k)^X = \\ &= 1 - P(A^h, X+1). \end{aligned} \quad (5)$$

Если вероятности отдельных событий малы, а число испытаний X достаточно большое, то вероятности p_k можно аппроксимировать непрерывной

плотностью $p(t)$, удовлетворяющей условию $\int_{k-1}^k p(t) dt = p_k$, а формулы (1) — (5) представить в виде

$$y = \int_0^n (1 - e^{-xp(t)}) dt, \quad (6)$$

$$y_m = C_x^m \int_0^n [p(t)]^m e^{-(x-m)p(t)} dt, \quad (7)$$

$$P(A^n, x) = \int_0^n p(t) e^{-xp(t)} dt = \frac{dy}{dx}, \quad (8)$$

$$\bar{p}(x) = \int_0^n [p(t)]^2 e^{-xp(t)} dt = -\frac{d^2y}{dx^2}, \quad (9)$$

$$\bar{F}(x) = \int_0^x \bar{p}(x) dx = 1 - \frac{dy}{dx}. \quad (10)$$

Как видно из формул (8), (9), вероятность появления нового события при x произведенных испытаниях равна значению первой производной в точке (x, y) кривой роста новых событий $y = f(x)$ (6), а средняя плотность $\bar{p}(x)$ — второй производной, взятой со знаком «минус». Здесь x — число произведенных испытаний, y — среднее значение числа наступивших новых событий.

Обозначим далее \bar{p}_j среднее значение вероятностей новых событий, которые могут наступить j -ми от начала испытаний (j — порядковый номер нового события), а $\bar{p}(y)$ — среднюю плотность распределения вероятностей новых событий, аппроксимирующую вероятности \bar{p}_j . Тогда

$$\bar{p}(y) = \bar{p}(x) \frac{dx}{dy} = -\frac{d^2y}{dx^2} \frac{dx}{dy} = -\frac{d}{dy} \left(\frac{dy}{dx} \right), \quad (11)$$

или, с учетом (6),

$$\bar{p}(y) = \int_0^n \frac{[p(t)]^2}{e^{xp(t)}} dt / \int_0^n \frac{p(t)}{e^{xp(t)}} dt; \quad (11')$$

$$\bar{F}(y) = \int_0^y \bar{p}(y) dy = 1 - \frac{dy}{dx}. \quad (12)$$

Формулы (10) и (12) позволяют находить кривую роста новых событий по заданной функции распределения $\bar{F}(x)$ или $\bar{F}(y)$:

$$y = \int [1 - \bar{F}(x)] dx + C \quad (13)$$

$$x = \int \frac{dy}{1 - \bar{F}(y)} + C, \quad (14)$$

где постоянная интегрирования C находится из условия $y = 0$ при $x = 0$

Если известны обе функции распределения, то кривая роста находится непосредственно из равенства $\bar{F}(x) = \bar{F}(y)$.

Отметим, что при больших x и малых m между зависимостями (6) и (7) существует связь, установленная В. М. Калининым [1]:

$$y_m = (-1)^{m+1} \frac{x^m}{m!} \frac{d^m y}{dx^m}. \quad (15)$$

По формуле (15) можно рассчитать частотный спектр, т. е. количество событий с частотой появления 0, 1, ..., m раз, если задана кривая роста новых событий $y = f(x)$.

Если известна средняя плотность $\bar{p}(x)$, то частотный спектр рассчитывается по формуле

$$y_m = (-1)^m \frac{x^m}{m!} \frac{d^{m-2} \bar{p}(x)}{dx^{m-2}},$$

которая следует из (15) и (13). При этом плотность $\bar{p}(x)$ должна быть убывающей функцией, не имеющей точек перегиба.

Выведенные формулы позволяют по одному известному закону распределения вероятностей новых событий построить системы непрерывных распределений и кривых роста новых событий, а также систему дискретных распределений [2].

ПОСТРОЕНИЕ СИСТЕМ НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ И КРИВЫХ РОСТА НОВЫХ СОБЫТИЙ

Пусть все n событий, составляющих полную группу, имеют равные вероятности $p_k = 1/n = \alpha$. Следовательно, плотность $p(t)$, аппроксимирующая вероятности p_k , также постоянна: $p(t) = \alpha$. В этом случае формула (11') дает

$$\bar{p}(y) = \alpha. \quad (16)$$

Итак, один закон распределения вероятностей новых событий задан. Восстановим по этому закону кривую роста новых событий $y = f(x)$ и среднюю плотность $\bar{p}(x)$, для чего используем формулы (12), (14), (9). Функция распределения здесь равна $\bar{F}(y) = \alpha y = 1 - (1 - \alpha y)$. Тогда согласно (14)

имеем $x = \frac{1}{\alpha} \ln(1 - \alpha y)$, откуда

$$y = \frac{1}{\alpha} \left(1 - \frac{1}{e^{\alpha x}} \right). \quad (17)$$

Далее по формуле (9) находим

$$\bar{p}(x) = \frac{\alpha}{e^{\alpha x}}, \quad (18)$$

т. е. второй закон распределения.

Пусть теперь средняя плотность $\bar{p}(y)$ задается формулой

$$\bar{p}(y) = \frac{\alpha}{e^{\alpha y}}. \quad (18')$$

Прделав ту же последовательность операций

что и в первом случае, найдем

$$y = \frac{1}{\alpha} \ln(1 + \alpha x), \quad (19)$$

$$\bar{p}(x) = \frac{\alpha}{(1 + \alpha x)^2}. \quad (20)$$

На следующем этапе средняя плотность $\bar{p}(y)$ будет задаваться формулой

$$\bar{p}(y) = \frac{\alpha}{(1 + \alpha y)^2}, \quad (20')$$

и т. д.

Этих результатов достаточно, чтобы сделать обобщение. Оно достигается путем введения нового параметра u , при определенных значениях которого из общих формул будут следовать рассмотренные выше частные случаи. Итак, обобщая (16), (18'), (20'), получаем

$$\bar{p}(y) = \alpha (1 - \alpha u y)^{\frac{1}{u}-1}. \quad (21)$$

Далее, на основании (18), (20) можем записать

$$\bar{p}(x) = \alpha [1 - \alpha (u - 1)x]^{\frac{1}{u}-1}. \quad (22)$$

По обобщенной средней плотности $\bar{p}(y)$ легко найти обобщенную кривую роста новых событий. Здесь $\bar{F}(y) = 1 - (1 - \alpha u y)^{1/u}$ и формула (14) дает

$$y = \frac{1}{\alpha u} \left[1 - \frac{1}{(1 + \alpha (1 - u)x)^{\frac{1}{1-u}}} \right] \quad (23)$$

Рассмотрим обобщенные плотности (21), (22). Будем относить распределения, а также кривые роста к I типу при $0 < u < \infty$, ко II типу — при $u \rightarrow 0$, к III типу — при $-\infty < u < 0$. При $u = 1$ из формулы (21) следует (16), т. е. равномерное распределение. Кривая роста (23) при $u \rightarrow 1$ имеет вид (17). При $u \rightarrow 0$ из (21) имеем показательный закон (18'), а кривая роста задается формулой (19), и т. д. При переходе от плотности $p(y)$ к плотности $p(x)$ параметр u уменьшается на единицу.

Кривая роста новых событий (23) позволяет оценить вероятность появления нового события (8) в точке с координатами (x, y)

$$P(A^u, x) = \frac{dy}{dx} = [1 - \alpha (u - 1)x]^{\frac{1}{u}-1}, \quad (24)$$

а также рассчитать объем выборки x при заданной вероятности $P(A^u, x)$.

Вероятность появления нового события приближенно можно найти по количеству событий, наступивших при x испытаниях один раз. Соответствующая расчетная формула получается из (8) и (15):

$$P(A^u, x) = \frac{dy}{dx} = \frac{y_{m=1}}{x}. \quad (25)$$

Осталось найти частотные спектры в случае распределений I—III типов, заданных обобщенной плотностью (21). Для этого используем формулы (15) и (23).

ПОСТРОЕНИЕ СИСТЕМЫ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ

Распределения I типа ($0 < u < \infty$). Из (21) следует, что в данном случае число разных (новых) событий, наступающих при x испытаниях, ограничено: $0 < y < 1/\alpha u = n$. Кривая роста новых событий задается формулой (23). Дифференцируя m раз по x выражение (23) и подставляя m -ю производную в (15), найдем

$$y_m = \left(\frac{\alpha u x}{1 + \alpha (1 - u)x} \right)^m \frac{\prod_{i=0}^{m-1} \left[1 + i \left(\frac{1}{u} - 1 \right) \right]}{m!} y_{m=0}, \quad m = 1, 2, \dots, \quad (26)$$

где

$$y_{m=0} = \frac{1}{\alpha u [1 + \alpha (1 - u)x]^{\frac{1}{1-u}}}. \quad (27)$$

Разделив, далее, (26) на $n = 1/\alpha u$, получим выражение для вероятности наступления событий ровно m раз при x испытаниях: $p_m = y_m/n$ (при этом удобно разделить на n величину $y_{m=0}$).

Исследования показывают, что частными случаями распределения I типа (26) являются: биномиальное — при $u > 1$; Пуассона — при $u \rightarrow 1$; отрицательное биномиальное — при $0 < u < 1$ (в том числе геометрическое распределение — при $u = 1/2$).

Распределение II типа ($u \rightarrow 0$). В данном случае кривая роста новых событий задается формулой (19), на основании которой имеем

$$y_m = \left(\frac{\alpha x}{1 + \alpha x} \right)^m \frac{1}{\alpha m}, \quad m = 1, 2, \dots \quad (28)$$

Разделив (28) на (19), получим

$$\frac{y_m}{y} = p_m = \left(\frac{\alpha x}{1 + \alpha x} \right)^m \frac{1}{m \ln(1 + \alpha x)}. \quad (28')$$

Последнее распределение известно как распределение Фишера по логарифмическому ряду и находит широкое применение в биологии [3].

Распределения III типа ($-\infty < u < 0$). Кривая роста новых событий задается формулой (23). Из (23) и (15) имеем

$$y_m = \left(\frac{-\alpha u x}{1 + \alpha (1 - u)x} \right)^{m-1} \frac{\prod_{i=1}^{m-1} \left[i \left(1 - \frac{1}{u} \right) - 1 \right]}{m!} y_{m=1}, \quad m = 2, 3, \dots, \quad (29)$$

где

$$y_{m=1} = \frac{x}{[1 + \alpha(1-u)x]^{1-u}} \quad (30)$$

Разделив (29) на (23), получим выражение для вероятности $p_m = y_m/y$.

Можно показать, что распределения III типа, заданные общей формулой (29), приводятся к форме (26), т. е. к распределениям I типа.

ОЦЕНИВАНИЕ ПАРАМЕТРОВ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ

Для установления типа выравнивающего распределения необходимо вычислить эмпирическое отношение $(x/y)^*$ и сравнить его с расчетным

$$\frac{x}{y} = \frac{\frac{x}{y_{m=1}} - 1}{\ln \frac{x}{y_{m=1}}} \quad (31)$$

При $(x/y)^* = x/y$ выравнивающее распределение относится ко II типу, при $(x/y)^* < x/y$ — к I типу, при $(x/y)^* > x/y$ — к III типу. Далее находятся оценки параметров α, u .

В случае распределений I типа

$$\alpha = \frac{1}{x^2} \sum_{m \geq 1} m^2 y_m - \frac{1}{x}, \quad (32)$$

$$u = \frac{1}{\alpha n}, \quad (33)$$

где $x = \sum_{m \geq 1} m y_m, \quad n = \sum_{m \geq 0} y_m$.

Таблица 1. Распределение по видам числа особей липидоптеры, пойманных световой ловушкой на Ротамстедской экспериментальной биостанции в 1935 г. (3)

Число особей, представляющих вид m	Число видов, представленных особями y_m		
	зафиксированное при наблюдении	рассчитанное при $x=6814$	рассчитанное при $x=100\,000$
1	2	3	4
1	37	37,69	37,89
2	22	18,74	18,94
3	12	12,43	12,62
4	12	9,27	9,46
5	11	7,37	7,57
6	11	6,11	6,30
7	6	5,21	5,40
8	4	4,53	4,72
9	3	4,01	4,20
10	5	3,59	3,78
11	2	3,24	3,43
12	4	2,96	3,14
13	2	2,71	2,90
14	3	2,51	2,69
15	2	2,33	2,51
16 и >	61	74,31	173,06

В случае распределений II типа оценка параметра α находится методом итераций по формуле, которая следует из (19):

$$\alpha_{i+1} = \frac{1}{y} \ln(1 + \alpha_i x), \quad (34)$$

где $y = \sum_{m \geq 1} y_m$; α_i — значение параметра α на предыдущем шаге итерации (в качестве первого приближения можно принять $\alpha_1 = 1/y_{m=1} - 1/x$, что вытекает из (19) и (25)).

В случае распределений III типа (а также I типа) оценка параметра u может быть найдена методом итераций по формуле

$$u_{i+1} = (1 - u_i) \frac{x}{y} \frac{\left[1 - \left(\frac{y_{m=1}}{x}\right)^{u_i}\right]}{\left[\left(\frac{x}{y_{m=1}}\right)^{1-u_i} - 1\right]}. \quad (35)$$

Тогда оценка параметра α равна

$$\alpha = \frac{1}{uy} \left[1 - \left(\frac{y_{m=1}}{x}\right)^u\right] = \frac{1}{(1-u)x} \left[\left(\frac{x}{y_{m=1}}\right)^{1-u} - 1\right]. \quad (36)$$

Таким образом, для оценивания параметров α, u достаточно знать три величины: $x, y, y_{m=1}$.

Отметим, что формулы (32), (36) справедливы для распределений трех типов.

При известных оценках параметров α, u необходимо вычислить теоретические значения y_m и сравнить их со статистическими. Расчет осуществляется по рекуррентной формуле

$$y_{m+1} = y_m \frac{\alpha x [u + m(1-u)]}{[1 + \alpha(1-u)x](m+1)}, \quad (37)$$

которая справедлива для распределений всех трех типов. Вначале по формуле (30) вычисляется количество событий с частотой $m=1$, т. е. $y_{m=1}$; далее по формуле (37) последовательно находятся: $y_{m=2}$ (при $m=1$), $y_{m=3}$ ($m=2$) и т. д. В случае распределений I типа дополнительно вычисляется величина $y_{m=0}$ по формуле (27).

РАСПРЕДЕЛЕНИЕ ЧИСЛА ОСОБЕЙ ПО ВИДАМ

В [3, с. 126] дан пример статистического распределения по видам особей мотыльков, пойманных световой ловушкой (табл. 1), а также выравнивающего распределения Фишера по логарифмическому ряду.

Общее число пойманных особей (объем выборки $x = \sum_{m \geq 1} m y_m$) равно 6814, число разных видов $y = \sum_{m \geq 1} y_m$ составило 197, в том числе с частотой один раз — $y_{m=1} = 37$.

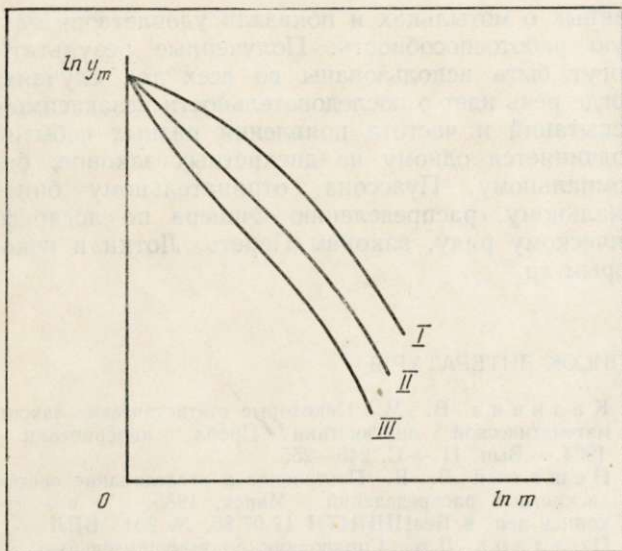


Рис. 1

По данным табл. 1 требуется установить закон распределения по видам числа особей в данной выборке и в выборке объемом $x = 100\,000$ особей, для обоих случаев вычислить вероятность появления нового вида.

Установим тип выравнивающего распределения, для чего вычислим отношение $(x/y)^* = 6814/197 = 34,6$. То же отношение, вычисленное по формуле (31), равно $x/y = 35,1$. Поскольку в обоих случаях получились близкие результаты, то в качестве выравнивающего может быть принято распределение II типа. Оценка параметра α , вычисленная по формуле (34), равна $\alpha = 0,026382$. Кривая роста новых видов задается формулой (19).

В табл. 1 приведены расчетные значения y_m , найденные по формулам (37), (30) при $a \rightarrow 0$, $\alpha = 0,026382$ (столбец 3). Те же результаты получаются непосредственно по формуле (28). Как и следовало ожидать, они близки к данным, рассчитанным по распределению Фишера [3].

Вычислим по формуле (24) вероятность появления нового вида при $x = 6814$, $u \rightarrow 0$, $\alpha = 0,026382$: $P(A^u, x) = 1/(1 + \alpha x) = 0,005532$. При $x = 100\,000$ эта вероятность равна $0,000379$, а среднее значение числа разных видов согласно (19) равно $y = 298,6$. В табл. 1 (столбец 4) дано распределение числа особей по видам при $x = 100\,000$.

Для $x \rightarrow \infty$ формула (28) принимает вид

$$y_m = \frac{1}{\alpha m}, \quad (38)$$

где $\alpha = 1/y_{m=1}$. В логарифмических координатах распределение (38) представляет собой прямую. Распределения I и III типов, а также II типа при $x < \infty$ в тех же координатах изображаются кривыми (см. рис. 1). Распределения III типа имеют точку перегиба и в средней части их можно аппрок-

симировать прямой $\ln y_m = \ln a - b \ln m$, откуда

$$y_m = \frac{a}{m^b} \quad (m = 1, 2, \dots; b > 1), \quad (39)$$

т. е. имеем так называемый гиперболический закон распределения, установленный эмпирическим путем В. Парето [4, с. 7]. При $b = 2$ он известен как закон Лотки.

Таким образом, закон Парето является эмпирическим аналогом распределений III типа, заданных формулой (29). При $b = 1$ он переходит в распределение II типа (38).

КРИВАЯ РОСТА НОВЫХ ВИДОВ

Рассчитаем по формуле (19) при найденной оценке α значения y (число видов) при некоторых значениях x (число особей). Результаты расчетов приведены в табл. 2, столбец 3. Для проверки справедливости формулы (19) необходимо иметь эмпирическую зависимость $y = f(x)$. В столбце 2 приведены значения y , восстановленные по эмпирическим данным табл. 1 по формуле В. М. Калинина [1],

$$y = y_0 - \sum_{m \geq 1} y_m \left(1 - \frac{x}{x_0}\right)^m, \quad (40)$$

где $y_0 = 197$, $x_0 = 6814$.

Данные в столбцах 2 и 3 табл. 2 различаются не более чем на 5 %

Отметим, что формула (40) дает возможность восстановить кривую роста новых событий по частотному спектру при $0 < x < x_0$, а формула (19) позволяет также делать прогноз (см. табл. 2, столбец 3).

Таблица 2. Зависимость между количеством особей мотыльков и количеством разных их видов

Количество особей x	Количество видов y (восстановлено по частотному спектру)	$y = \frac{1}{\alpha} \ln(1 + \alpha x)$
1	2	3
300	84,2	82,9
500	96,5	100,5
1000	119,8	125,5
1500	136,1	140,4
2000	148,1	151,0
2500	157,5	159,4
3000	165,0	166,2
4000	176,6	177,0
5000	185,3	185,3
6000	192,2	192,2
6814	197,0	197,0
10000	—	211,5
20000	—	237,7
30000	—	253,0
40000	—	263,9
50000	—	272,4
100000	—	298,6

Сделаем некоторые выводы.

Поскольку система дискретных распределений взаимосвязана с системой кривых роста новых событий, а также с законами распределения вероятностей новых событий, она позволяет решать широкий круг задач с единой точки зрения. Например, всего лишь по трем заданным величинам — x , y , $y_{m=1}$ — эта система распределений позволяет рассчитывать:

- оценки параметров α , u выравнивающего распределения;
- кривую роста новых событий $y = f(x)$;
- частотный спектр, т. е. количество разных событий, наступающих ровно $0, 1, 2, \dots, m$ раз при x испытаниях;
- средние плотности распределения вероятностей новых событий $\bar{p}(x)$, $\bar{p}(y)$;
- вероятности появления и не появления нового события в точке (x, y) кривой роста новых событий, а также значения величин x, y при заданных вероятностях появления и не появления нового события.

Система кривых роста новых событий и система дискретных распределений были проверены на

данных о мотыльках и показали удовлетворительную работоспособность. Полученные результаты могут быть использованы во всех тех случаях, когда речь идет о последовательности независимых испытаний и частота появления разных событий подчиняется одному из дискретных законов: биномиальному, Пуассона, отрицательному биномиальному, распределению Фишера по логарифмическому ряду, законам Парето, Лотки и некоторым др.

СПИСОК ЛИТЕРАТУРЫ

1. Калинин В. М. Некоторые статистические законы математической лингвистики // Пробл. кибернетики.— 1964.— Вып. 11.— С. 246—255.
2. Нешиной В. В. Построение и исследование систем дискретных распределений.— Минск, 1985.— 71 с.— Рукопись деп. в БелНИИТИ 17.07.85, № 931—БЕЛ.
3. Поллард Дж. Справочник по вычислительным методам статистики.— М.: Финансы и статистика, 1982.— 344 с.
4. Петров В. М., Яблонский А. И. Математика и социальные процессы.— М.: Знание, 1980.— 60 с.

Поступила 27.03.85