УДК 519.213.1

МОДЕЛИРОВАНИЕ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ ДАННЫХ
Modeling the laws of data distribution

Л. А. Козинцева, магистр 1 курса
Н. И. Шингарева, кандидат сельскохозяйственных наук
Уральский государственный аграрный университет
(Екатеринбург, ул. Карла Либкнехта, 42)

Аннотация

Моделирование законов распределения данных применяется при разработке или исследовании готовых алгоритмов, когда свойства данных из реальных источников неизвестны и нельзя спрогнозировать результат их выполнения.

Для моделирования данных по одному из хорошо известных законов распределения можно использовать, например, нормальный закон. Применяя исследуемый алгоритм к модельным данным, можно заранее предположить, каким окажется результат его выполнения. Если он окажется удовлетворительным, можно применить его и к реальным данным.

Ключевые слова: моделирование, данные, процесс, планирование, элементы, система, развитие.

Summary

Data distribution laws modeling is used in the development or research of ready-made algorithms, when the properties of data from real sources are unknown and it is impossible to predict the result of their implementation.

For example, the normal law can be used to model data according to one of the well-known distribution laws. By applying the algorithm under study to the model data, it is possible to predict in advance what the result of its execution will be. If it turns out to be satisfactory, you can try to apply it to real data.

Keywords: modeling, data, process, planning, elements, system, development.

Большей информативностью, по сравнению с такими статистическими характеристиками как математическое ожидание, дисперсия, обладает закон распределения вероятности случайной величины X. Представим, что X принимает случайные значения из некоторого диапазона. Например, X – диаметр вытачиваемой детали. Диаметр может отклоняться от запланированного идеального значения под влиянием различных факторов, которые нельзя учесть, поэтому он является случайной слабо предсказуемой величиной. Но в результате длительного наблюдения за выпускаемыми деталями можно отметить, сколько деталей из 1000 имели диаметр X_1 (обозначим N_{X1}), сколько деталей имели диаметр X_2 (обозначим N_{X2}) и так далее. В итоге можно построить гистограмму частости диаметров, откладывая для X_1 величину $N_{X1}/1000$, для X_2 величину $N_{X2}/1000$ и так далее. (Обратите внимание, если быть точным, N_{X1} – это число деталей, диаметр которых не просто равен X_1, а находится в диапазоне от $X_1 - \Delta/2$ до $X_1 + \Delta/2$, где $\Delta = X_1 - X_2$). Важно, что сумма всех частностей будет равна 1 (суммарная площадь гистограммы неизменна). Если X меняется непрерывно, опытов проведено очень много, то в пределе $N \to \infty$ гистограмма превращается в график распределения вероятности случайной величины [1]. На рис. 1.1, а показан пример гистограммы дискретного распределения, а на рис. 1.1, б показан вариант непрерывного распределения случайной величины.

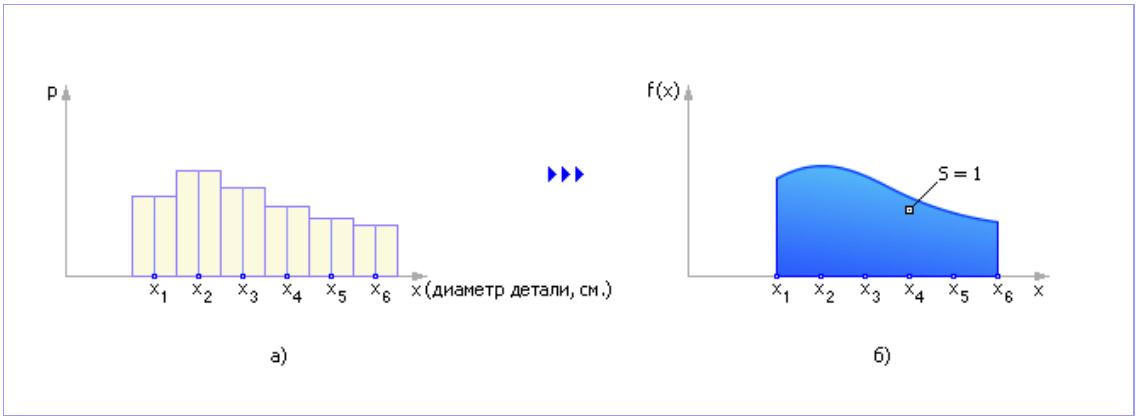

Рис. 1.1. Сравнение дискретного и непрерывного законов распределения случайной величины

Обозначим: h_i – высота i-го столбца, $f(x)$ – распределение вероятности (показывает насколько вероятно некоторое событие x). Значение h_i операцией нормировки необходимо перевести в единицы вероятности появления значений x из интервала $x_i < x \le x_{i+1}$: $P_i = h_i/(h_1 + h_2 + \ldots + h_i + \ldots + h_n)$.

Операция нормировки обеспечивает сумму вероятностей всех n событий равную 1:

$$\sum_{i=1}^{n} P_i = 1$$

На рис. 1.2 показаны графически переход от произвольного непрерывного закона распределения к дискретному (рис. 1.2, а), отображение получаемых вероятностей на интервал $r_{pp}[0; 1]$ и генерация случайных событий с использованием эталонного равномерно распределенного ГСЧ (рис. 1.2, б).

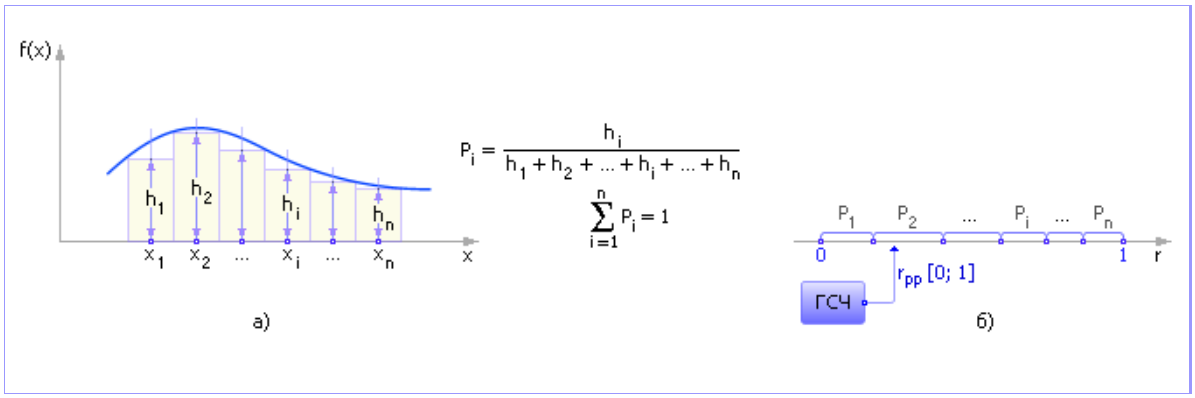

Рис. 1.2. Иллюстрация метода ступенчатой аппроксимации

Заметим, что внутри интервала $x_i < x \le x_{i+1}$ значение x теперь не различимо, одинаково. Метод огрубляет изначальную постановку задачи, переходя от непрерывного закона распределения к дискретному. Поэтому следует учитывать количество разбиений n из условий точности представления [2].

На рис. 1.3 показан фрагмент алгоритма, реализующего описанный метод. Алгоритм генерирует случайное число, равномерно распределенное от 0 до 1. Затем, нужно сначала сгенерировать случайное число r от 0 до 1. Затем в цикле будем сравнивать границы отрезков с этим числом.

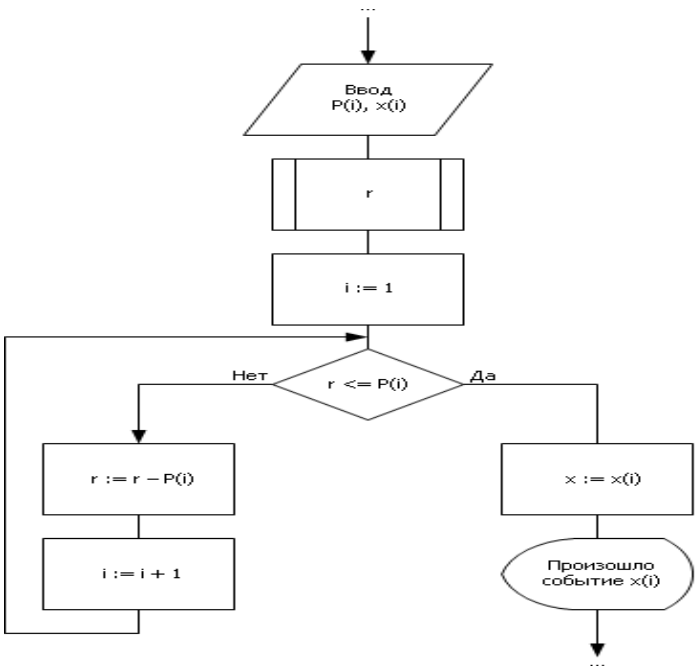

Рис. 1.3. Блок-схема алгоритма, реализующего метод ступенчатой аппроксимации

Заметим, что внутри интервала $x_i < x \le x_{i+1}$ значение x теперь не различимо, одинаково. Метод огрубляет изначальную постановку задачи, переходя от непрерывного закона распределения к дискретному. Поэтому следует учитывать количество разбиений n из условий точности представления [3].

Метод усечения используется в случае, когда функция задана аналитически (в виде формулы). График функции вписывают в прямоугольник (рис. 1.4). На ось Y подают случайное равномерно распределенное число из ГСЧ. На ось X подают случайное равномерно распределенное число из ГСЧ. Если точка в пересечении этих двух координат лежит ниже кривой плотности вероятности, то событие X произошло, иначе нет.

Недостатком метода является то, что те точки, которые оказались выше кривой распределения плотности вероятности, отбрасываются как ненужные, и время, затраченное на их вычисление, оказывается напрасным. Метод применим только для аналитических функций плотности вероятности [4].

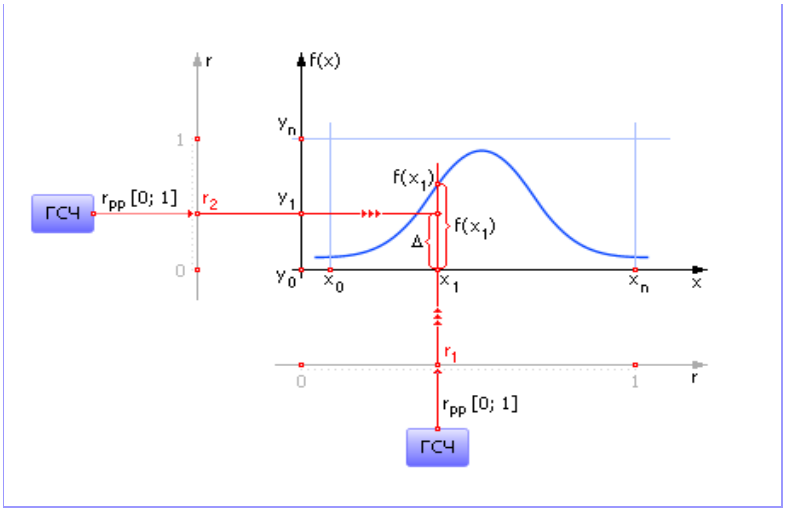

Рис. 1.4. Иллюстрация метода усечения

Метод взятия обратной функции

Допустим, что нам задан интегральный закон распределения вероятности $F(x)$, где $f(x)$ – функция плотности вероятности и

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Тогда достаточно разыграть случайное число, равномерно распределенное в интервале от 0 до 1. Поскольку функция F тоже изменяется в данном интервале, то случайное событие x можно определить взятием обратной функции по графику или аналитически: $x = F^{-1}(r)$. Здесь r – число, генерируемое эталонным ГСЧ в интервале от 0 до 1, x_1 – сгенерированная в итоге случайная величина. Графически суть метода изображена на рис. 1.5.

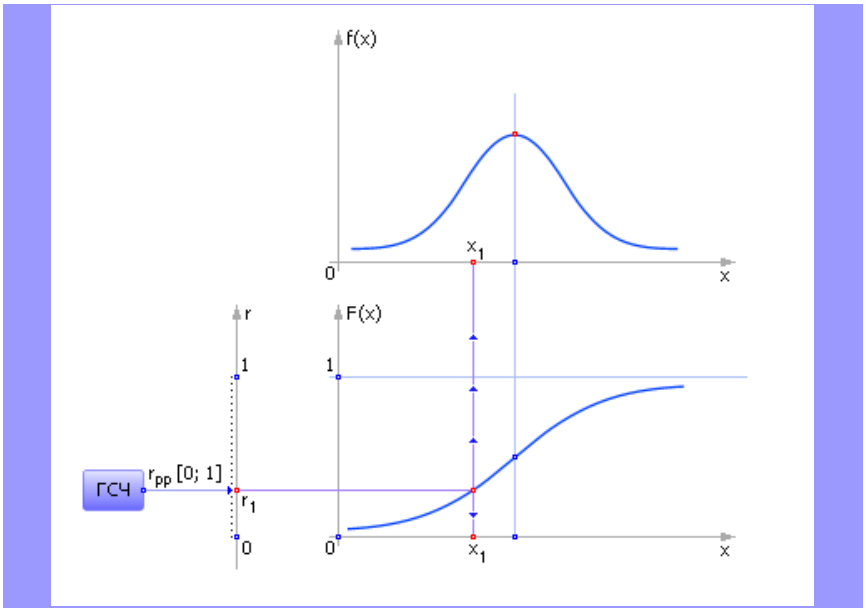

Рис. 1.5. Иллюстрация метода обратной функции для генерации случайных событий x, значения которых распределены непрерывно

На рисунке показаны графики плотности вероятности и интегральной плотности вероятности от x.Данным методом особенно удобно пользоваться в случае, когда интегральный закон распределения вероятности задан аналитически и возможно аналитическое взятие обратной функции от него, как это и показано на следующем примере [3].

Рассмотрим пример моделирования законов распределения: К существующим эмпирическим частотам получим теоретические частоты и определим критерий согласия различными способами.

С помощью формул определим стандартное отклонение и теоретические частоты по эмпирическому распределению.

$$t_i = \frac{(Xi - \bar{X})}{\sigma}; \quad \tilde{n}_i = \frac{N \cdot C_x}{\sigma}\varphi(t_i)$$

Таблица 1

Вычисление выравнивающих (теоретических) частот нормального распределения

Центральные значения классов (X_i)	Частоты, (n_i)	Отклонения, $(Xi - \overline{X})$	Стандартноеотклонение, (t_i)	Относительные ординаты нормальной кривой, $\varphi(t_i)$	Теоретические частоты, $(\tilde{n}_i)$
27,0	5	-40,39	1,38	0,1539	2,8
42,2	7	-25,19	0,86	0,2756	5,0
57,4	7	-9,99	0,34	0,3765	6,8
72,6	4	5,21	0,18	0,3825	7,0
87,8	3	20,41	0,70	0,3123	5,7
103,0	6	35,61	1,22	0,1895	3,4
118,2	3	50,81	1,73	0,0893	1,6
ИТОГО	Σ=35				Σ=32,3

Далее рассчитаем критерий согласия

Таблица 2

Расчет критерия согласия (способ 1)

Центральные значения классов, (X_i)	Частоты		$(n_i - \tilde{n}_i)$	$(n_i - \tilde{n}_i)^2$	$\chi^2 = \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}$
	эмпири-ческие (n_i)	теорети-ческие $(\tilde{n}_i)$			
27,0	5	2,8	+2,2	4,84	1,729
42,2	7	5,0	+2,0	4,00	0,800
57,4	7	6,8	+0,8	0,04	0,006
72,6	4	7,0	-3,0	9,00	1,286
87,8	3	5,7	-2,7	7,29	1,279
103,0	6	3,4	+2,6	6,76	1,988
118,2	3	1,6	+1,4	1,96	1,225
ИТОГО	Σ=35	Σ=32,3			$\chi^2_{выч}$=Σ8,313

Делаем вывод: $df=k-l-1$=7-2-1=4 число степеней свободы

На уровне значимости α=10%

значит $\chi^2_{(st)10\,\%}$=9,49

=> выражение $\chi^2_{\phi} < \chi^2_{(st)10\%}$ или 8,313 < 9,49 выполняется

В случае, если выражение выполняется, то гипотезу принимаем и с вероятностью можно утверждать, что расхождение между теоретическими и эмпирическими частотами случайно. Следовательно, есть основания утверждать, что эмпирическое распределение подчиняется нормальному распределению.

Так как объемы выборки не равны, возьмем еще один пример расчета критерия согласия для неравных выборок.

Таблица 3

Расчет критерия согласия (способ 2)

$N_1 = 35$

$N_2 = 32,3$

Эмпирии-ческие (n_i)	Теорети-ческие $(\widetilde{n_i})$	$N2 * ni - N1 * \widetilde{ni})^2$	$(N2 * ni - N1 * \widetilde{ni})^2$	$ni + \widetilde{ni}$	$x^2 = \frac{(N2 * ni - N1 * \widetilde{ni})^2}{ni + \widetilde{ni}}$
5	2,8	+63,5	4032,25	7,8	516,955
7	5,0	+51,1	2611,21	12,0	217,601
7	6,8	-126	158,76	13,8	11,504
4	7,0	-115,8	13409,64	11,0	1219,058
3	5,7	-102,6	10526,76	8,7	1209,972
6	3,4	74,8	5595,04	9,4	595,217
3	1,6	40,9	1672,81	4,6	363,654
					$\chi^2_{выч}$=Σ4133,961

$\chi 2 = \frac{1}{N1*N2} * \sum \frac{(N2*ni-N1*\widetilde{ni})^2}{ni+\widetilde{ni}} = \frac{1}{35*32,3} * 4133,961 = 3,657$

Делаем вывод: df=k-l-1=7-2-1=4 число степеней свободы

На уровне значимости α=10%

значит $\chi^2_{(st)10\,\%}$=9,49

=> выражение $\chi^2_{ф} < \chi^2_{(st)10\%}$ или 3,657 < 9,49 выполняется

Рассмотри еще один пример расчета критерия согласия

Таблица 4

Расчет критерия согласия по Колмогорову – Смирнову (способ 3)

Центральные значения клас-сов, (X_i)	Частоты		Накопительная частота по $\sum(n_i)$	Накопительная частота по $\sum(\widetilde{n_i})$	[Д] $=\sum(n_i)-\sum(\widetilde{n_i})$
	эмпирии-ческие (n_i)	теорети-ческие $(\widetilde{n_i})$			
27,0	5	2,8	5	2,8	2,2
42,2	7	5,0	12	7,8	4,2
57,4	7	6,8	19	14,6	4,4 *max*
72,6	4	7,0	23	21,6	1,4
87,8	3	5,7	26	27,3	1,3
103,0	6	3,4	32	30,7	1,3
118,2	3	1,6	35	32,3	2,7
ИТОГО	Σ=35	Σ=32,3			

$\chi^2_{выч}=\frac{[Д]}{\sqrt{N}} = \frac{[4,4]}{\sqrt{35}} = 0,7$

$\chi_{2ф}$ = 1,36 для уровня значимости α=10%

=>выражение $\chi^2_ф<\chi^2_{(st)10\%}$ или 0,7<1,36 выполняется

Отклонения фактических (эмпирических) частот от теоретических являются случайными. Следовательно, нулевая гипотеза принимается и есть основания утверждать, что эмпирическое распределение подчиняется нормальному распределению.

В заключении, можно сделать вывод что моделирование законов распределения данных позволяет имитировать появление конкретных случайных событий согласно вероятностям заданного распределения. Также для моделирования многомерного массива данных, распределённых по нормальному закону, можно использовать преобразование Бокса-Мюллера: при помощи двух случайных чисел, распределённых на интервале (0;1], получаются одновременно два числа, распределённых по нормальному закону.

Библиографический список

1. Моделирование случайной величины с заданным законом распределения. URL: https://stratum.ac.ru/education/textbooks/modelir/lection24.html.
2. Моделирование законов распределения. URL: https://studopedia.net/15_62938_modelirovanie-zakonov-raspredeleniya.html.
3. Ландшафт – Википедия. URL: https://ru.wikipedia.org/wiki/Ландшафт.
4. Моделирование законов распределения случайных величин. URL https://studfile.net/preview/5270816/page:5/.