ЭКОНОМИКА
И МАТЕМАТИЧЕСКИЕ МЕТОДЫ

1969
ТОМ V, ВЫП. 5

ОБЗОР ПРОГРАММ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

И. П. ФРЕНКИНА

(Москва)

1. ВВЕДЕНИЕ

В настоящее время составлено большое количество программ по статистической обработке различного рода информации: технической, экономической, медицинской и т. п.

К 1969 г. в лаборатории типовых алгоритмов и программ ЦЭМИ АН СССР имелись сведения примерно о 130 программах по математической статистике: около 40% для ЭВМ типа БЭСМ-3М (М-20); 24% для машин типа Минск-2 (22); 13% для машин типа Урал; 23% для других типов машин.

В основном программы составлены в машинном коде (80%), но появляется все больше программ, составленных на АЛГОЛЕ (20% по имеющимся сведениям).

В программах по математической статистике обычно производятся следующие расчеты: 1) вычисление числовых характеристик случайной величины; 2) определение теоретической функции распределения и степени ее согласованности с истинной функцией распределения; 3) корреляционный анализ; 4) регрессионный анализ [1—20].

2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

В программах, как правило, предусматривается вычисление различных центральных моментов случайной величины в качестве ее числовых характеристик.

Первый центральный момент дает математическое ожидание случайной величины

$$m_x = M[x] = \sum_{i=1}^{n} x_i p_i,$$

где x_i — возможные значения случайной величины; p_i — вероятности этих значений.

Обычно рассматриваются центрированные случайные величины, т. е. отклонения случайных величин от их математических ожиданий $\overset{0}{x} = x_i - m_x$.

Центральным моментом порядка S случайной величины x_i называется математическое ожидание S-й степени соответствующей центрированной случайной величины

$$\mu_s[x] = M[\overset{0}{x}{}^s] = [M(x - m_x)^s].$$

Второй центральный момент называется дисперсией случайной величины

$$\mu_2 = D[x] = M(\overset{0}{x^2}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2 = \alpha_2 - m_x^2,$$

где

$$\alpha_2 = \sum_{i=1}^{n} x_i^2 p_i.$$

Корень квадратный из дисперсии называется средним квадратическим отклонением

$$\sigma = \sigma[x] = \sqrt{D[x]} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2}.$$

Третий центральный момент служит для характеристики асимметрии или «скошенности» распределения. Коэффициент асимметрии (или асимметрия) S_k вычисляется по формуле

$$S_k = \frac{\mu_3}{\sigma^3} = \frac{n}{(n-1)(n-2)\sigma^3}\sum_{j=1}^{n}(x_i{}^j - M_i)^3 = j_{3i},$$

где

$$\mu_3 = M[x^3] = \sum_{i=1}^{n}(x_i - m_x)^3 p_i = \alpha_3 - 3\alpha_2 m_x + 2m_x{}^3$$

и

$$\alpha_3 = \sum_{i=1}^{n} x_i{}^3 p_i.$$

Эксцесс Ex записывается через четвертый центральный момент $Ex = (\mu_4/\sigma^4) - 3$ (для нормального закона распределения $\mu_4/\sigma^4 = 3$ и эксцесс $Ex = 0$). Здесь

$$\mu_4 = M[x^4] = \sum_{i=1}^{n}(x_i - m_x)^4 p_i.$$

Эксцесс характеризует «крутость», т. е. островершинность ($Ex > 0$) или плосковершинность ($Ex < 0$) кривых.

3. ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Выравнивание статистических рядов производится путем подбора теоретических функций распределения. Определение параметров теоретической функции распределения производится в программах методом наименьших квадратов из системы нормальных уравнений. Система нормальных уравнений составляется из условия, что $Q = \sum_{i=1}^{n}\eta_i = \min$ ($\eta_i = y_i - y_{Ti}$), т. е. $\partial Q/\partial a_j = 0$, где j — число параметров a_j-й теоретической функции распределения $y_T = y(x, a_j)$.

Наиболее часто аппроксимация производится линейной, параболической, логарифмической функциями.

При определении степени согласованности теоретического и статистического распределения используются различные критерии согласия (меры

расхождения), в том числе λ-критерий Колмогорова, χ^2-критерий Пирсона и др.

Согласно критерию Пирсона в качестве меры расхождения между теоретическим и статистическим распределениями берут сумму квадратов отклонений $(p_i^* - p_i)$ с весами c_i

$$U = \sum_{i=1}^{k} c_i (p_i^* - p_i)^2.$$

Здесь p_i — теоретические вероятности, p_i^* — наблюденные частоты.

Результаты опытов сводим в k разрядов и обозначаем через m_i число значений в i-м разряде. Тогда $p_i^* = m_i / n$ и, беря $c_i = n / p_i$ (n — число опытов), получаем асимптотическую формулу для функции распределения случайной величины

$$\chi^2 = U = n \sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}$$

Распределение χ^2 зависит от числа степеней свободы r, т. е. от разности числа разрядов и числа независимых условий, наложенных на частоты p_i^*.

Для χ^2 составлены таблицы доверительных границ при различном числе степеней свободы, по которым по r и χ^2 определяется вероятность того, что величина, имеющая распределение χ^2 с r степенями свободы, превзойдет данное значение χ^2. Если эта вероятность мала, то гипотеза отбрасывается как неправдоподобная. Если эта вероятность относительно велика, то гипотеза не противоречит опытным данным.

По критерию Колмогорова рассматривается максимальное значение модуля разности между статистической функцией распределения $F^*(x)$ и соответствующей теоретической функцией распределения $F(x)$: $D = \max |F^*(x) - F(x)|$.

Определяется $\lambda = D\sqrt{n}$ и по λ по таблицам находят вероятность $p(\lambda)$ того, что максимальное расхождение между $F^*(x)$ и $F(x)$ будет не меньше чем фактически наблюденное. При малом $p(\lambda)$ гипотеза неправдоподобна; при больших $p(\lambda)$ гипотеза совместима с опытными данными.

Меру согласованности можно также определить по критерию ω^2. В этом случае вычисляется мера расхождения между теоретической $F(x)$ и эмпирической $F^*(x)$ функциями распределения как средний квадрат отклонений по всем возможным значениям аргумента

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n} \sum_{k=1}^{n} \left[F(x_k) - \frac{2k-1}{2n} \right]^2.$$

4. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Если имеем систему двух случайных величин x и y, то для характеристики их рассеивания и связи между ними вводится корреляционный момент (или ковариация) K_{xy}, равный математическому ожиданию произведения центрированных величин $K_{xy} = M[\overset{0}{x}\overset{0}{y}] = M(x - m_x)(y - m_y)$, где m_x, m_y — математическое ожидание x и y.

Характеристикой связи в чистом виде для величин x и y служит коэффициент корреляции $r_{xy} = K_{xy} / \sigma_x \sigma_y$, где σ_x и σ_y — средние квадратичные отклонения. Случайные величины, для которых корреляционный момент равен нулю, называются некоррелированными, хотя они могут быть

зависимыми. Коэффициент корреляции характеризует степень тесноты линейной зависимости между случайными величинами.

При $y = ax + b$ имеем $r_{xy} = \pm 1$. В общем случае имеем $-1 < r_{xy} < +1$.

При положительной корреляции между случайными величинами ($r_{xy} > 0$) при возрастании одной величины другая имеет тенденцию в среднем возрастать. При отрицательной корреляции ($r_{xy} < 0$) при возрастании одной случайной величины другая имеет тенденцию в среднем убывать.

Если имеем систему нескольких случайных величин, то составляется корреляционная матрица $\|K_{ij}\|$ системы случайных величин; причем

$$K_{ij} = M[\overset{0}{x}_i \overset{0}{x}_j] \text{ при } i \neq j \text{ и}$$

$$K_{ii} = M[\overset{0}{x}_i{}^2] = D_i \text{ при } i = j \ (D_i - \text{дисперсия}).$$

Нормированная корреляционная матрица $\|r_{ij}\|$ составляется из коэффициентов корреляции $r_{ij} = K_{ij} / \sigma_i \sigma_j$, где $\sigma_i = \sqrt{D_i}$, $\sigma_j = \sqrt{D_j}$, D — дисперсия.

Если случайные величины характеризуются качественными признаками, то вводится коэффициент ранговой корреляции (по Спирмену): отдельным случайным величинам приписывают порядковые номера в соответствии с убыванием качества. При двух качественных признаках каждой случайной величине приписывают два порядковых номера ξ_i и η_i. Коэффициент ранговой корреляции R (по Спирмену) определяется формулой

$$R = \sum_{i=1}^{n} \xi_i \eta_i / Q,$$

где $Q = \Sigma \xi_i{}^2 = \Sigma \eta_i{}^2 = [n(n-1)(n+1)]/3$.

После преобразований для R можно получить формулу

$$R = 1 - \frac{6 \Sigma d^2}{n(n-1)(n+1)} = 1 - \frac{6SR}{n^3 - n},$$

где d — разность порядковых номеров по обоим признакам. $R = +1$ при совпадении обоих порядковых номеров ($\xi - \eta = 0$). $R = -1$ — когда оба ряда полностью противоположны ($\xi + \eta = 0$). Коэффициент ранговой корреляции T (по Кендаллу) определяется для случайных величин, упорядоченных по двум качественным признакам. Обозначим через x_i, y_i случайные величины, характеризующие качественные признаки.

Вводим новые случайные величины $x_{ik}(y_{ik})$, определяемые равенствами

$$x_{ik} = \begin{cases} +1, \text{ если } x_i < x_k, \\ 0, \text{ если } x_i = x_k, \\ -1, \text{ если } x_i > x_k. \end{cases}$$

Тогда коэффициент корреляции T по Кендаллу равен $T = 2S/[n(n-1)]$, где

$$S = \sum x_{ik} y_{ik} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sign}(y_j - y_i).$$

Значения T находятся между $(+1)$ и (-1) $(-1 \leqslant T \leqslant +1)$, причем $T = 1$ при совпадении последовательности порядковых номеров и $T = -1$ при противоположных последовательностях.

Корреляционные функции рассматриваются также для стационарных эргодических случайных функций, т. е. таких, для которых каждая отдельная реализация случайной функции может заменить при обработке множество реализаций (т. е. среднее по времени равно среднему по множеству наблюдений).

Для эргодических стационарных случайных функций вычисляются автокорреляционные функции по одной реализации

$$K_x\left(\frac{mT}{n}\right)=\frac{1}{n-m}\sum_{i=1}^{n-m}\dot{x}(t_i)\,x(t_i+m),$$

где n — количество значений случайной функции на $[0, T]$, а

$$\dot{x}(t_i)=x(t_i)-\frac{1}{n}\sum_{j=1}^{n}x(t_j),$$

т. е. центрированное значение случайной функции, и по множеству реализаций

$$\rho_x(t)=\frac{\dfrac{1}{n(m-\mu)}\sum_{i=1}^{n}\sum_{j=1}^{m-\mu}\left(x_{ij}-\left(\sum_{i,j=1}^{nm}x_{ij}/nm\right)\right)\left(x_{i,j+\mu}-\left(\sum_{i,j=1}^{nm}x_{ij}/nm\right)\right)}{\dfrac{1}{nm}\sum_{i,j=1}^{nm}\left(x_{ij}-\left(\sum_{i,j=1}^{nm}x_{ij}/nm\right)\right)^2},$$

где n — количество реализаций случайной функции; m — количество сечений случайной функции; x_{ij} — текущее значение случайной функции; $\mu=\tau/T_n$ — нормированный параметр (T_n — временной интервал между двумя последовательными изменениями случайной функции).

Вычисляется также взаимная корреляционная функция для двух эргодических случайных функций $x(\tau)$, $y(\tau)$

$$\rho_{xy}(\tau)=\frac{1}{N-\mu}\sum_{i=1}^{N-\mu}\left(x_i-\frac{1}{N}\sum_{i=1}^{N}x_i\right)\left(y_{i+\mu}-\frac{1}{N}\sum_{i=1}^{N}y_i\right)\times$$

$$\times\left[\frac{1}{N-1}\left(\sum_{i=1}^{N}\left(x_i-\left[\sum_{i=1}^{N}x_i/N\right]\right)^2\sum_{i=1}^{N}y_i-\left[\sum_{i=1}^{N}y_i/N\right]\right)^2\right]^{-1/2},$$

где N — множество дискретных значений каждого случайного процесса; x_i, y_i — текущие значения случайных функций; $\mu=\tau/T_n$ — нормированный параметр.

5. РЕГРЕССИОННЫЙ АНАЛИЗ

Коэффициент регрессии γ определяется из условия, чтобы при $y=\gamma x+z$ дисперсия z была наименьшей.

Коэффициент регрессии связан с коэффициентом корреляции r соотношением $\gamma=r\sigma_y/\sigma_x$.

Выборочный коэффициент регрессии γ_1 определяется как угловой коэффициент в эмпирической линии регрессии $y-m_y=\gamma_1(x-m_x)$, выбираемый таким образом, чтобы сумма квадратов отклонений точек (x_i, y_i) от прямой была наименьшей. Если переменная x является временем, то линия регрессии называется трендом.

Для различных заданных зависимостй между y и x (линейной, квадратичной, k-го порядка, тригонометрической) определяют коэффициенты в формулах из условия, чтобы отклонение истинного графика функции от соответствующей линии регрессии было наименьшим.

Можно производить также авторегрессивное преобразование по формулам $x_i' = x_i - rx_{i-1}$; $y_i' = y_i - ry_{i-1}$, где r — коэффициент автокорреляции.

6. ХАРАКТЕРИСТИКИ НЕКОТОРЫХ ПРОГРАММ

В печати опубликован ряд программ по математической статистике. В двух сборниках [22, 23], изданных ЦЭМИ АН СССР, опубликована система программ по математической статистике, написанных на АЛГОЛЕ для α-транслятора, трансляторов ТА-1 и ТА-1М. Программы оформлены в виде процедур. По этим процедурам могут быть вычислены следующие величины: оценки числовых характеристик случайного вектора, коэффициент корреляции, корреляционная матрица, математическое ожидание, оценки параметров линейного или квадратичного классификатора, автокорреляционная функция эргодической стационарной случайной функции по одной или множеству реализаций, взаимная корреляционная функция для двух стационарных эргодических случайных функций по одной и множеству реализаций, оценки статистического распределения случайной величины, оценка степени согласованности теоретического и статистического распределения с помощью χ^2-критерия Пирсона, λ-критерия Колмогорова, ω-критерия и т. п.

В программе статистической обработки динамических рядов [21], составленной в машинных кодах БЭСМ-3М, производится одновременное выравнивание нескольких динамических рядов с числом членов каждого динамического ряда $n < 30$. Параметры кривой, по которой производится выравнивание, определяются из системы нормальных уравнений по методу наименьших квадратов.

По программе вычисляются: средняя квадратическая ошибка, оценка дисперсии параметров кривой, выборочный коэффициент вариации, достоверность коэффициента регрессии, коэффициент точности прогноза, соотношение фон Неймана, отсутствие систематического сдвига в наблюдениях по критерию Аббе, показатель для оценки параметров кривой. По коэффициенту автокорреляции производится авторегрессивное преобразование.

В [24] приведено 7 стандартных программ по статистическому анализу. В них произведено вычисление математического ожидания и дисперсии (в целых числах), коэффициента корреляции, ранговых коэффициентов корреляции по Спирмену и Кендаллу, коэффициентов линии регрессии (для ортогональных полиномов Чебышева), доверительных интервалов, проверяется однородность ряда дисперсий, вычисляются числовые характеристики статистического ряда (эмпирические плотности и вероятности распределения, математические ожидания, дисперсии, коэффициент вариации, эксцесс, плотность вероятности нормального закона распределения, проверка с помощью критерия однородности ряда дисперсий постоянства дисперсии зависимой переменной при изменении значений аргумента, т. е. χ^2-распределение с k степенями свободы).

Значительное число программ посвящено вычислению корреляционных зависимостей в медицине, металлургии, статистической обработке всевозможных технологических параметров, статистической обработке анкетных данных и т. п.

ЛИТЕРАТУРА

1. Т. Андерсон. Введение в многомерный статистический анализ. М., Физматгиз, 1963.
2. Б. Л. ван дер Варден. Математическая статистика. М., Изд-во иностр. лит., 1960.
3. Е. С. Вентцель. Теория вероятностей. М., Физматгиз, 1962.
4. В. Е. Гмурман. Введение в теорию вероятностей и математическую статистику. М., «Высшая школа», 1966.
5. Б. В. Гнеденко. Курс теории вероятностей. М., «Наука», 1965.
6. А. И. Карасев. Основы математической статистики. М., Росвузиздат, 1962.
7. Г. Крамер. Математические методы статистики. М., Изд-во иностр. лит., 1948.
8. Н. А. Лившиц, В. Н. Пугачев. Вероятностный анализ систем автоматического управления. Т. 1. М., «Сов. радио», 1963.
9. Ю. В. Линник. Метод наименьших квадратов и основы теории обработки наблюдений. М., Физматгиз, 1958.
10. Д. А. Райков. О разложении законов Гаусса и Пуассона. Изв. АН СССР. Сер. матем., 1938, стр. 91—124.
11. Сборник задач по теории вероятностей, математической статистике и теории случайных функций. Под общ. ред. А. А. Свешникова, М., «Наука», 1965.
12. Н. В. Смирнов, Н. В. Дунин-Барковский. Краткий курс математической статистики для технических приложений. М., Физматгиз, 1959.
13. Г. Тинтнер. Введение в эконометрию. М., «Статистика», 1965.
14. С. Уилкс. Математическая статистика. М., «Наука», 1967.
15. В. А. Унковский. Теория вероятностей. М., Воен.-мор. изд-во, 1953.
16. В. Феллер. Введение в теорию вероятностей и ее приложения. Т. 1. М., «Мир», 1967.
17. А. Хальд. Математическая статистика с техническими приложениями. М., Изд. иностр. лит., 1956.
18. Э. Хеди, Д. Диллон. Производственные функции в сельском хозяйстве. М., «Прогресс», 1965.
19. Б. М. Щиголев. Математическая обработка наблюдений. М., Физматгиз, 1960.
20. H. Theil. Economic Forecasts and Policy. Amsterdam, 1961.
21. И. П. Френкина. Программа статистической обработки динамических рядов. В сб. Стандартные программы, серия 4, № 1. М., 1967 (ЦЭМИ АН СССР).
22. В. И. Тихомиров и др. Алгоритмы вычисления статистических характеристик случайной величины. В сб. Программы и алгоритмы, вып. 16. М., 1968 (ЦЭМИ АН СССР).
23. В. И. Тихомиров и др. Алгоритмы вычисления статистических характеристик системы случайных величин, случайных функций, системы случайных функций. В сб. Программы и алгоритмы, вып. 17, М., 1968 (ЦЭМИ АН СССР).
24. Математическое обеспечение ЭВМ «Минск-2». Вып. 2. Минск, 1968 (Институт математики АН БССР).

Поступила в редакцию
10 IV 1969