ЛИНЕЙНЫЙ ВАРИАНТ АНАЛИЗА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ

Войнов Кирилл Николаевич
Д.т.н., профессор, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
forstar@mail.ru
Наср Тарек Мохаммед Абдулджаббар
Аспирант, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
tarek01021988@mail.ru
Афанасьев Максим Яковлевич
К.т.н., доцент, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
amax@niuitmo.ru
Хилдаяти Анниса
Аспирант, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
hildayati.annisa@mail.ru

Аннотация. В статье обращается внимание на формирование банка статистических данных, накапливаемых исследователем, при наблюдении за каким-либо физическом явлением, например, износом деталей в работе механизма/машины, давлением, температурой, перемещением, скоростью, объёмы выбросов газов с предприятий и из выхлопных труб автомобилей и др. При этом первоначально рассматривается наиболее простой случай, соответствующий линейной зависимости, при ограниченном числе точек наблюдений за развитием процесса во времени. Показаны компьютерные расчёты с использованием математической оболочки MathCad. Кроме того, объяснён алгоритм поиска уравнения, максимально быстро определяющего наилучший вариант приближения теоретической аппроксимации данных. Кроме того, приводится прикладной пример с использованием критерия, с помощью которого можно установить следующее: а оставлять ли в общей статистической сводке наблюдений неожиданно появившееся в экспериментах резко выделяющееся значение по отношению к остальным, которое является, по всей видимости, нехарактерным, что позволит его тогда не принимать в расчёт.

Ключевые слова: статистические данные, износ, банк данных, компьютерная обработка, информация.

LINEAR VARIANTS OF STATISTICAL DATA ANALYSIS

K. Voinov
T. Nasr
M. Afanasev
A. Hildayati

Summary. This paper is concerned with the formation of a bank of statistical data accumulated by the researcher when observing any physical phenomenon, for example, the wear and tear of parts in the operation of machinery, pressure, temperature, displacement, speed, the volume of gases emissions from the enterprises and from the exhaust pipes of cars and others. In this case, the simplest case corresponding to the linear dependence is initially considered, with a limited number of observation points of process development in time. Computer calculations using the mathematical shell of MathCad are shown. In addition, an algorithm for finding the equation determining the best approximating variant of the theoretical approximation of data. Moreover, an applied example is given with the use of a criterion by which it is possible to establish the following: Is it possible to leave in the general statistical summary of observations an unexpectedly appearing in the experiments sharply distinguishable value with respect to the others, which is apparently uncharacteristic, which will allow it then not to take into account.

Keywords: statistical data, wears, data bank, computer processing, information.

Хорошо известно следующее. Через одну точку можно провести неограниченное число прямых линий или кривых; через две точки можно провести прямую линию; через три и большее число разбросанных точек наблюдения за конкретным физическим явлением можно провести несколько ломаных отрезков соединяющихся между собой прямых, либо одну прямую, построенную, например, методом наименьших квадратов или иным способом.

Для условного примера, который будет далее приведён, воспользуемся линейным износом детали на первом этапе её приработки в паре трения. Второй этап стабильного и обычно более медленного процесса развития износа, а также третьего (катастрофического по В. Ф. Лоренцу) этапа, не рассматриваем. Так как начало развития износа связано с исходным размером детали, имеющей номинальное значение и симметричный

Таблица 1

x_i	y^1_i	y^2_i	y^3_i	y^4_i	y^5_i
0	0	0	0	0	0
150	26.449	54.595	86.601	104.718	178.758

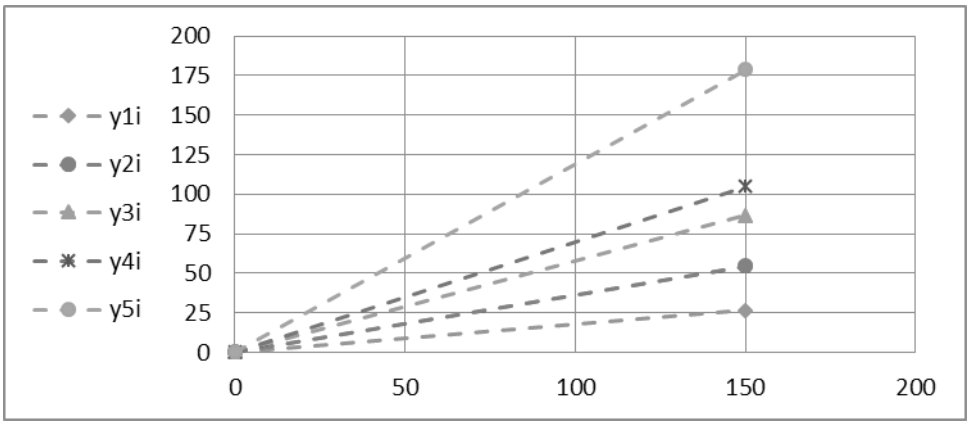

Рис. 1. Пять возможных вариантов изнашивания деталей

либо несимметричный допуск на изготовление, то соответственно прямые могут идти как из начала прямоугольной системы координат, так и могут быть смещены в основном по вертикальной оси y. Вместе с тем, если начало развития износа детали происходит не мгновенно с началом её эксплуатации, а с некоторой задержкой, то тогда будет иметь место начало развития износа со сдвигом по оси абсцисс также с учётом значения и знака указанного поля допуска на изготовление.

Ввиду того, что материалов, из которых могут быть изготовлены детали, образующие пары трения, может быть великое множество, то и темп развития изнашивания может быть на первом этапе (приработки) весьма различным. Сформировав исходный банк данных о возможных характерах изнашивания, можно с помощью вычислительной техники (компьютера) и программирования заставить вычислительную машину быстро найти наиболее близкое уравнение, описывающее развитие износа во времени (или в течение длительности пути трения), подставив лишь исходные числовые значения [1] — [8].

Линеаризация процесса

В наиболее общем случае линейная функция может быть записана в следующем виде:

$$y = ax + b \tag{1.1}$$

где a, b — параметры.

Причём функция монотонно возрастает при a>0, монотонно убывает при a<0 и постоянна при $a = 0$. Если $b = 0$, то имеет место прямая пропорциональность, при которой $y = ax$, a прямая проходит через начало координат. Другая возможная запись прямой линии такая:

$$y = kx \pm b \tag{1.2}$$

где y угловой коэффициент прямой $k = tg\alpha$; α — угол между положительным направлением оси абсцисс Ox и прямой; b — отрезок, отсекаемый прямой на ординатной оси Oy с учётом знака.

Используя данную информацию, в компьютерной оболочке MathCad приведём несколько прямых с разным наклоном, то есть с разным пространственным расположением в прямоугольной системе координат (рис. 1).

$$i = 1..2;\ k1 = 10;\ k2 = 20;\ k3 = 30;\ k4 = 40;\ k5 = 50;$$
$$h = 0.017453$$

$$t1_i = \tan(k1 \cdot h) \qquad y1_i = x_i \cdot t1_i$$
$$t2_i = \tan(k2 \cdot h) \qquad y2_i = x_i \cdot t2_i$$
$$t3_i = \tan(k3 \cdot h) \qquad y3_i = x_i \cdot t3_i$$
$$t4_i = \tan(k4 \cdot h) \qquad y4_i = x_i \cdot t4_i$$
$$t5_i = \tan(k5 \cdot h) \qquad y5_i = x_i \cdot t5_i$$

На рис. 1 показан пример построения пяти прямых, выходящих из начала координат и имеющих разный тангенс угла наклона ($k1$,…, $k5$); величина h — переводит градусы угла в радианы.

Если зафиксированные при проведении эксплуатации или в лаборатории значения наблюдаемой физиче-

Итог расчёта Таб 2

n_j	t_j	m_j
1	0	-8.8
2	28	43.4
3	100	95.6
4	150	147.8

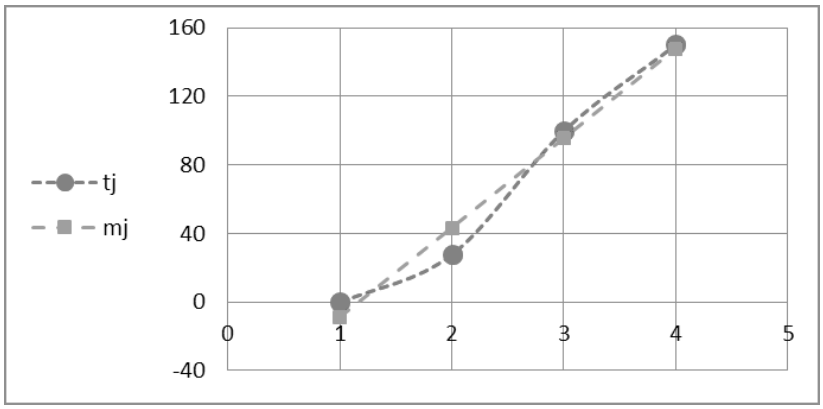

Рис. 2. Линейная аппроксимация данных

ской величины не ложатся на прямую, то можно использовать метод наименьших квадратов, чтобы получить оптимально проходящую прямую. Покажем решение на двух примерах.

Пример 1.

Часы наработки, h, ч: 0 50 62 90 Σ = 202
Выбросы газа, t, см^3: 0 28 100 150 Σ = 278
Номера контроля, n: 1 2 3 4 Σ = 10
n^2: 1 4 9 16 Σ = 30
tn: 0 56 300 600 Σ = 956

Для получения расчётного уравнения следует записать следующую систему уравнений для последующего вычисления параметров a_0 и a_1:

$$a_0 h + a_1 \sum n = \sum t$$

$$a_0 \sum n + a_1 \sum n^2 = \sum t \cdot n$$

Подставив числовые значения, получаем:

$4a_0 + 10a_1 = 278$
$10a_0 + 30a_1 = 956.$

Уравняем значения параметра a_0 путём умножения на –2,5 все числа первой строки, то есть:

–10–25a_1 = — 695. Складывая, получим: 5a_1 = 261. Откуда a_1 = 52,2. Теперь из первого уравнения вычисляем a_0.

В частности, имеем: 4a_0 + 522 = 278, откуда a_0 = — 61.

Тогда окончательно получаем: t_n = –61 + 52,2n. Меняя значения n = 1, 2, 3, 4, находим числовые величины для построения прямой линии:

–8,8; 43,4; 105,0; 147,8.

Составим программу MathCad для компьютера, чтобы графически посмотреть полученное решение (рис. 2).

$$j = 1....4;\ m_j = -61 + 52.2 \cdot n_j$$

См. таблицу 2.

В рисунке 2 показан вариант компьютерного построения: ломаная линия t_j — данные эксперимента, линия m_j из точек — построенная аппроксимирующая прямая по методу наименьших квадратов.

Изменяющаяся условная величина выбросов газов t_j в атмосферу дана в возрастающем порядке, так как по мере начала работы котельной/печи или двигателя автомобиля системы постепенно выходят на свой устойчивый режим работы.

Пример 2.

На практике при проведении опытов (научных исследований) могут встречаться случаи, когда по тем или иным причинам появляется одно значение (а иногда не-

Таблица 3

n	Уровень значимости, α				n	Уровень значимости, α			
	0.05	0.02	0.01	0.001		0.05	0.02	0.01	0.001
2	15.56	38.97	77.96	779.69	16	2.20	2.68	3.04	4.20
3	4.97	8.04	11.46	36.48	17	2.18	2.65	3.00	4.13
4	3.56	5.08	6.53	14.47	18	2.16	2.64	2.99	4.07
5	3.04	4.11	5.04	9.43	19	2.15	2.62	2.95	4.02
6	2.78	3.64	4.36	7.41	20	2.14	2.60	2.93	3.98
7	2.61	3.36	3.96	6.37	22	2.12	2.58	2.89	3.91
8	2.51	3.18	3.71	5.73	24	2.11	2.55	2.86	3.84
9	2.43	3.05	3.53	5.31	26	2.10	2.53	2.84	3.79
10	2.37	2.96	3.41	5.01	28	2.09	2.52	2.82	3.76
11	2.32	2.88	3.31	4.79	30	2.08	2.50	2.80	3.72
12	2.29	2.83	3.23	4.62	40	2.04	2.45	2.74	3.60
13	2.26	2.78	3.17	4.48	60	2.02	2.41	2.68	3.49
14	2.23	2.74	3.12	4.37	120	1.99	2.37	2.63	3.39
15	2.22	2.71	3.08	4.27	∞	1.96	2.32	2.57	3.29

Примечание. Округления сделаны до сотых долей.

сколько), которые резко выделяются из общей сводки статистических данных. Тогда возникает естественный вопрос: а оставлять эти данные в расчёте или их можно исключить (как нехарактерные)? Учёные разработали несколько методов для анализа подобных ситуаций. Здесь изложим метод, известный как критерий Романовского В. И. Суть состоит в следующем.

Первоначально по выборке вычисляют среднее $x_{ср}$ и среднее квадратичное отклонение s без учёта спорного члена ряда распределения x_1 или x_n. Далее вводят коэффициент t_α, зависящий от α и членов ряда наблюдений n, причём обеспечиваемая вероятность принятия решения будет $P = 1 - \alpha$ (табл. 3). Таблица 3 значений для анализы резко выделяющихся значений приведена далее.

Тогда, если $(x_{ср} - x_1 / s) \gg t_\alpha$ или $|x_{ср} - x_n| \gg t_\alpha$ то с выбранной вероятностью значение x_1 или x_n можно исключить из общей сводки наблюдений.

Наконец, если имеется несколько грубо выделяющихся значений, то величины $x_{ср}$ и s определяются без них, после чего каждое в отдельности проверяется по приведённой схеме. Использование компьютера позволяет по специальной программе осуществлять такие операции практически мгновенно

Как было объяснено ранее, развитие физического процесса, за которым ведётся наблюдение, может начинаться с некоторой сдвижкой по осям координат, что обусловлено, например, наличием допусков или несколько замедленной реакцией системы/объекта. Отмеченное представлено на рис. 3 для положительного допуска (пример с отрицательным значением не приводится в силу тривиальности решения, когда свободный член берётся в уравнениях со знаком минус). При этом базовые выражения использованы из п. 1).

Где $i = 1..2;\ k1 = 10;\ k2 = 20;\ k3 = 30;\ k4 = 40;\ k5 = 50;\ h = 0.017453$

$$t1_i = \tan(k1 \cdot h) \qquad y6_i = x_i \cdot t1_i + 15$$
$$t2_i = \tan(k2 \cdot h) \qquad y7_i = x_i \cdot t2_i + 15$$
$$t3_i = \tan(k3 \cdot h) \qquad y8_i = x_i \cdot t3_i + 15$$
$$t4_i = \tan(k4 \cdot h) \qquad y9_i = x_i \cdot t4_i + 15$$
$$t5_i = \tan(k5 \cdot h) \qquad y10_i = x_i \cdot t5_i + 15$$

На рис. 3 показан пример линейного описания наблюдаемого процесса, когда исходное значение исследуемой физической величины имеет положительный допуск

Таким образом, имея большой банк статистических данных, по соответствующей компьютерной программе можно не только описать физический процесс математически и графически, но и проверить возможное наличие в собранной информации нехарактерных точек (показаний приборов), которые можно исключать из расчёта.

При этом не будет представлять какой-то особой проблемы выполнить подобные расчёты при нелинейных процессах с их последующей аппроксимацией.

Лучшее приближение к математическому описанию конкретного процесса вычислительная машина по заложенной в неё программе также делает в считанные секунды.

Таблица 4

$y6_i$	$y7_i$	$y8_i$	$y9_i$	$y10_i$
15	15	15	15	15
41.449	69.595	101.601	119.718	193.758

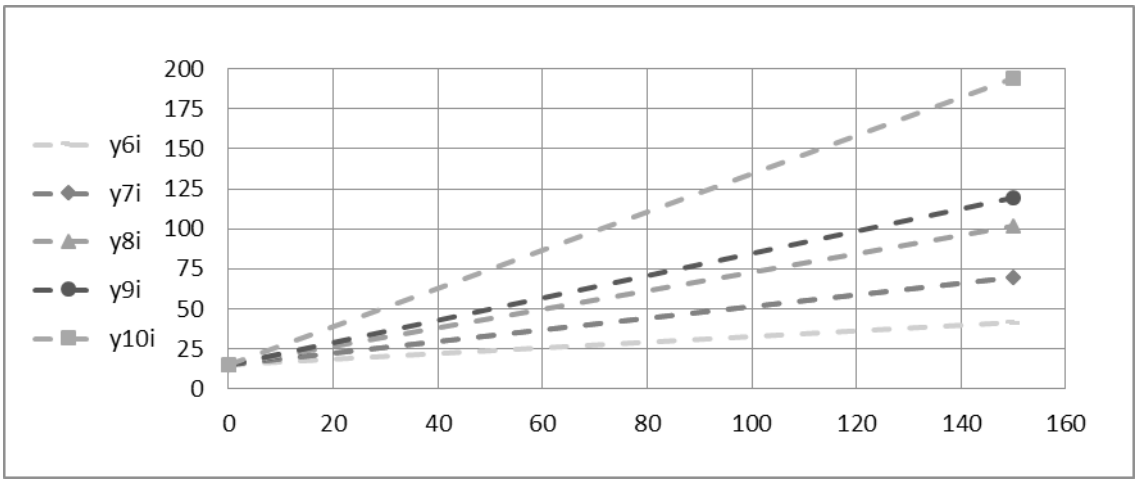

Рис. 3. Пример линейного описания наблюдаемого процесса, когда исходное значение исследуемой физической величины имеет положительный допуск

Заключение

Рассмотрены варианты обработки базы данных с линеаризацией неких физических процессов, с приведением формул, аппроксимаций с учётом начальных допусков. Представлен алгоритм анализа статистической информации, когда в ней возможно появление нехарактерных по отношению к остальным значениям данных. Все расчёты и программы составлены в компьютерном варианте с использованием оболочки MathCad.

ЛИТЕРАТУРА

1. Романовский В. И. Применения математической статистики в опытном деле. М.-Л. Государственное издательство технико-теоретической литературы: ОГИЗ-ГОСТЕХИЗДАТ, 1947. — 248 с.
2. Венецкий И. Г., Кильдишев Г. С. Основы математической статистики. М.: Госстатиздат, 1963. — 308 с.
3. Четыркин Е. М., Калихман И. Л. Вероятность и статистика. М.: Финансы и статистика,1982. — 319 с.
4. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высшая школа, 1979. — 400 с.
5. Кудрявцев Е. М. Mathcad 2000 Pro. М.: ДМК Пресс, 2001. — 576 с.
6. Черняк А. А. и др. Математика для экономистов на базе Mathcad. Санкт-Петербург: БХВ-Петербург, 2003. —496 с.
7. Гутер Р. С., Овчинский Б. В. Элементы численного анализа и математической обработки результатов опыта. М.: Наука, 1970. — 432 с.
8. Колмогоров А. Н., Журбенко И. Г., Прохоров А. В. Введение в теорию вероятностей. М.: Наука, 1982. — 160 с.

© Войнов Кирилл Николаевич (forstar@mail.ru), Наср Тарек Мохаммед Абдулджаббар (tarek01021988@mail.ru),
Афанасьев Максим Яковлевич (amax@niuitmo.ru), Хилдаяти Анниса (hildayati.annisa@mail.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»