Лекция 5

ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ ДИАГНОСТИКИ

Многих проблем диагностики, как и измерительной техники вообще, не существовало бы, если бы измерения диагностических параметров не были неразрывно связаны с погрешностями измерения физических величин. Кроме того, данные, полученные в результате диагностики, представляют собой, как правило, случайные величины. Неизбежность случайных погрешностей и случайный характер диагностических данных приводит к необходимости применения соответствующего аппарата математической статистики и теории вероятностей.

Современное программное обеспечение позволяет быстро провести анализ данных с оценкой достоверности заключений. Однако необходимо понимание алгоритмов обработки информации, заложенных в основу работы вычислительных программ, что позволяет правильно интерпретировать полученные результаты. Рассмотрим некоторые общие принципы статистической обработки измерительной информации и иллюстрации их применения. Принципы статистического анализа едины, независимо от сферы их приложения — науки, техники, медицины, экономики, социологии. Исходные данные могут быть получены посредством счета, измерений, социологических опросов и т. д.

Множество данных, полученных по результатам измерений одной или нескольких характеристик объекта, составляет основу статистического анализа. Полученные при измерениях данные образуются *ряды измерений*. Это — наборы чисел, являющихся случайными величинами как из-за неизбежных погрешностей измерений, так и вследствие вариации характеристик объектов контроля. Будем обозначать характеристики — названия измеряемых величин, заглавными латинскими буквами X, Y, Z и т. д. Если ряд содержит N членов, говорят о *выборке* размером N. Наблюдаемые значения случайной величины X обозначим малыми латинскими буквами с индексами, означающими номер измерения, например, x_i — результат i-го измерения параметра X.

При безграничном увеличении N выборка переходит в *генеральную совокупность*. Выборка — часть генеральной совокупности, реально получаемая на практике.

Расположив члены выборки в порядке возрастания, получим *вариационный ряд*, определяемый соотношениями $x_i \le x_{i+1}\ \left(i = 1, 2, ..., N\right)$.

При большом объеме выборки прибегают к группировке данных, то есть объединяют несколько соседних результатов в *группы* (*интервалы*). Их ширину выбирают по возможности одинаковой для всех интервалов, число интервалов k рекомендуется выбирать, руководствуясь, например, соотношением $k \approx 5\lg N$. Для выборки объемом $N = 50$ измерений получим $k = 8$, при $N = 100$ будем иметь $k = 10$ и т. д. Найденное количество групп позволяет назначить *величину интервала* изменения переменной в группе:

$$h = \left(x_{\max} - x_{\min}\right)/k\ .$$

Этот интервал и число групп округляются до целых или рациональных, то есть удобных для восприятия величин.

Наиболее полная характеристика статистических свойств генеральной совокупности — *функция распределения* $F\left(x\right)$, вероятность появления значения величины X, меньшего, чем x:

$$F\left(x\right) = P\left(X < x\right).$$

Производная от функции распределения носит название плотности вероятностей:

$$w\left(x\right) = dF\left(x\right)/dx\ ,$$

следовательно,

$$F\left(x\right) = \int\limits_{-\infty}^{x} w\left(x\right)dx\ .$$

Вероятность попадания значения x случайной величины X в интервал $x_1 \le x \le x_2$ выражается как

$$P\left(x_1 \le X \le x_2\right) = \int\limits_{x_1}^{x_2} w\left(x\right)dx = F\left(x_2\right) - F\left(x_1\right).$$

Очевидны соотношения:

$$\int\limits_{-\infty}^{\infty} w\left(x\right)dx = 1;\qquad F\left(-\infty\right) = 0;\qquad F\left(\infty\right) = 1.$$

На основании сделанной группировки может быть получена эмпирическая функция распределения. Для ее построения используют формулу

$$F^*(x)=\begin{cases} 0 & \text{при} \quad x<x_1; \\ i/N & \text{"} \quad x_i<x_{i+1}; \\ 1 & \text{"} \quad x\ge x_N, \end{cases}$$

где i — количество значений попавших в интервал $\left[x_1;x_{i+1}\right]$. Эмпирическая функция распределения имеет скачки величиной в точках $x=x_1$, $x=x_2$, … $x=x_N$.

Пример Л5.1. Собраны данные о значениях наружного диаметра оболочек твэлов реактора ВВЭР-1000, изготовленных в течение 25 рабочих дней. Ежедневно контролировали по 5 случайно отобранных оболочек. Результаты приведены в прилагаемой таблице (значения диаметров d_i приведены в сотых долях миллиметра после вычитания 9 мм). В соответствии с техническими требованиями значение параметра должно составлять $9{,}13^{+0{,}06}_{-0{,}05}$ мм.

День	1	2	3	4	5	6	7	8	9	10	11	12	13
d_1	12	13	15	11	13	13	12	13	12	12	13	14	13
d_2	15	13	13	14	15	13	11	13	14	14	15	13	15
d_3	14	16	16	14	17	14	15	13	12	13	14	15	11
d_4	11	12	14	13	12	13	13	14	15	12	11	14	13
d_5	13	14	14	15	12	15	16	15	13	16	13	11	13

День	14	15	16	17	18	19	20	21	22	23	24	25
d_1	12	15	16	12	11	13	12	14	15	13	13	14
d_2	14	11	11	13	12	15	14	15	14	12	12	15
d_3	15	14	14	14	16	12	13	11	13	15	12	17
d_4	15	13	17	14	13	13	13	13	14	11	13	11
d_5	13	15	14	15	14	12	14	13	15	14	13	14

Рассмотрим весь набор данных как единую выборку, характеризующую качество технологического процесса, и применим группировку данных. Объем выборки $N=25\times5=125$, рекомендуемое количество интервалов группировки $k\approx10$. Величины $x_{\min}$ и $x_{\max}$ соответственно равны 11 и 17, тогда $h=0{,}6$. Но длину интервала нельзя выбрать меньшей единицы ввиду ограниченной этой величиной точности данных. Далее, чтобы границы интервалов не совпали с результатами измерений, сместим их на половину единицы измерения соответственно в меньшую и большую сторону от $x_{\min}$ и $x_{\max}$. Тогда количество интервалов с длиной равной 1 будет равно $k=6+2=8$. В результате получим таблицу частот появления результатов в группах:

Границы интервала	Абсолютная частота	Относительная частота	Относительная накопленная частота
10,5 ... 11,5	12	0,096	0,096
11,5 ... 12,5	18	0,144	0,240
12,5 ... 13,5	36	0,288	0,528
13,5 ... 14,5	28	0,224	0,752
14,5 ... 15,5	22	0,176	0,928
15,5 ... 16,5	6	0,048	0,976
16,5 ... 17,5	3	0,024	1,000
Всего	125	1,000	

Сгруппированные данные дают возможность их графического представления в виде столбчатой диаграммы (*полигона*) накопленных относительных частот, показанных в последнем столбце таблицы. Этот полигон является эмпирическим приближением к функции распределения $F^*(x)$, рис. Л5.1, *б*.

Предпоследний столбец позволяет построить *гистограмму* распределения. Гистограмма представляет собой столбчатую диаграмму, в основании которой отложены интервалы группирования данных, а по вертикальной оси — абсолютные или относительные частоты для данного интервала (рис. Л5.1, *а*). Для получения нормированного распределения следует число результатов, отложенное по вертикальной оси, поделить на общее число измерений N. Для аппроксимации может быть выбран и другой закон распределения.

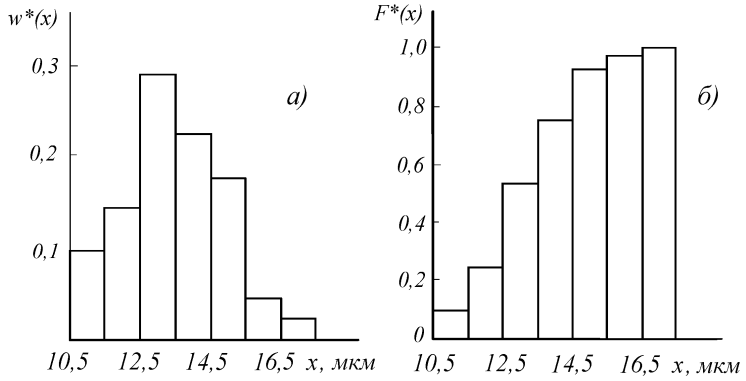

Рис. Л5.1. Гистограммы распределения (*а*) и накопленных частот (*б*) для приведенного примера

Пример Л5.2. Исследовалась стойкость труб с покрытием и без покрытия против коррозии. С этой целью 12 пар образцов, вырезанных из указанных двух групп труб, поместили попарно в среды с разной коррозионной активностью и после выдержки в течение года измерили среднюю глубину коррозионного поражения металла.

Для сопоставлявшихся пар получены следующие результаты:

Номер пары	1	2	3	4	5	6	7	8	9	10	11	12
Средняя глубина коррозии, мкм												
С покрытием	39	43	43	52	52	59	40	45	47	62	40	27
Без покрытия	42	37	61	74	55	57	44	55	37	70	52	55

Необходимо проверить, подтверждают ли данные результаты предположение о том, что трубы с покрытием менее подвержены коррозии.

Данный пример удобен тем, что его анализ позволяет обойтись простейшими вычислительными средствами и вместе с тем затрагивает ряд характерных для такого анализа проблем, которые мы последовательно рассмотрим в дальнейших лекциях.

В приведенном примере имеются две выборки: первая — для образцов с покрытием, вторая — для образцов без покрытия. Обозначим случайные величины (значения глубины коррозии) через X и X_1 для первой и второй выборок соответственно. Для простоты временно будем рассматривать только первую выборку. Расположив члены выборки в порядке возрастания, получим *вариационный ряд*, определяемый соотношениями $x_i \le x_{i+1}$ $\left(i = 1, 2, ..., N\right)$. Для рассматриваемой выборки вариационный ряд имеет вид: 27, 39, 40, 40, 43, 43, 45, 47, 52, 52, 59, 62. Эмпирическая функция распределения имеет вид, показанный на рис. Л5.2.

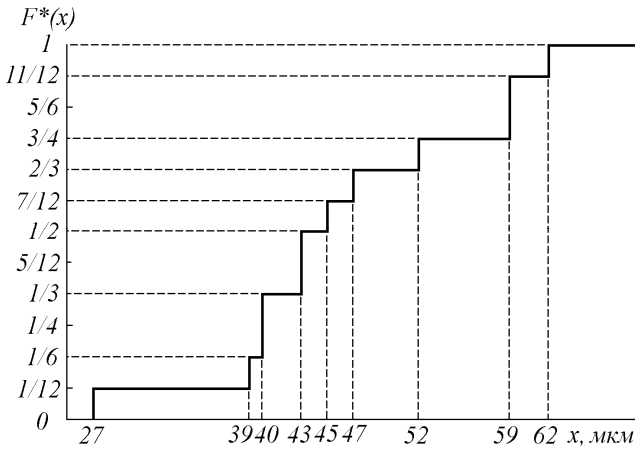

Рис. Л5.2. Эмпирическая функция распределения для примера с коррозией труб

При малом числе измерений эмпирическая плотность вероятностей может быть получена дифференцированием сглаженной эмпирической функции распределения, однако погрешности, возникающие при таком построении, оказываются значительными. Это является платой за малый объем выборки.

С функцией (плотностью) распределения связаны *статистики* — числовые характеристики функции распределения, являющиеся «кратким описанием» результатов измерений. Важнейшими из статистик являются следующие (в скобках приведены выборочные статистики).

1). Математическое ожидание (среднее значение):

$$a = \mu = M\{X\} = \int_{-\infty}^{\infty} x w(x)\, dx; \qquad \left(\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \right).$$

2). Мода $\mathrm{Mo}\{X\}$, определяемая соотношением $w\left(\mathrm{Mo}\{x\}\right) = \max$.

Иными словами, мода — значение случайной величины, при котором плотность вероятностей имеет максимальное значение. Для симметричных распределений мода и среднее значение совпадают.

3). Медиана $\mathrm{Me}\{X\}$, соответствующая условию

$$P\left(x < \mathrm{Me}\{X\}\right) = P\left(x > \mathrm{Me}\{X\}\right) = 0{,}5.$$

4). Квантиль распределения уровня p — величина x_p ,определяемая уравнением

$$P\left(x < x_p\right) = F\left(x_p\right) = p,$$

или

$$x_p = F^{-1}\left(p\right),$$

где функция $F^{-1}\left(p\right)$ — обратная функции $F\left(x\right)$. Квантиль $x_{0,5}$ соответствует медиане распределения.

5). Дисперсия:

$$D\{X\} = \sigma_x^2 = \int_{-\infty}^{\infty} \left(x - a\right)^2 w\left(x\right) dx; \qquad \left(s^2 = 1/\left(N-1\right) \sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2 \right).$$

6). Среднеквадратическое (стандартное) отклонение:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{D}; \qquad \left(s = \sqrt{s^2} \right)$$

и связанный с ним *коэффициент вариации*:

$$\gamma = \sigma/\mu; \qquad \left(\gamma_{выб} = s/\overline{x} \right).$$

При малых объемах выборки $\left(N<10...15\right)$ значения среднеквадратического отклонения, вычисленные по экспериментальным данным, следует скорректировать, умножив на величину, указанную в таблице:

Значения поправочного множителя *k* к расчетной величине среднеквадратического отклонения при малом объеме выборки размером *N*

N	2	3	4	5	6	7	8	9	10
k	1,25	1,13	1,09	1,06	1,05	1,04	1,04	1,03	1,03

Генеральные совокупности чаще всего могут быть описаны нормальным (гауссовым) распределением, которое характеризуется плотностью вероятностей

$$w(x)=\left(1/\sqrt{2\pi\sigma}\right)\exp\left\{-(x-\mu)^2/2\sigma^2\right\}$$

и обозначается для краткости $N\left(\mu,\sigma^2\right)$, где μ и σ^2 — параметры распределения. Заменой переменных $z=\left(x-\mu\right)/\sigma$ можно перейти к нормированному нормальному распределению $N\left(0,1\right)$:

$$\varphi(z)=\left(1/\sqrt{2\pi}\right)\exp\left\{-z^2/2\right\},$$

функция распределения которого (функция Лапласа):

$$\Phi\left(z\right)=\left(1/\sqrt{2\pi}\right)\int_{-\infty}^{z}\exp\left(-z^2/2\right)dz\,.$$

Очевидно, что

$$\Phi\left(-z\right)=1-\Phi\left(z\right);\qquad\qquad \varphi(x)=\varphi(z)/\sigma\,.$$

Можно показать, что распределение среднего значения $\overline{x}$ нормального распределения при известном генеральном среднем μ и известной генеральной дисперсии σ является нормальным распределением $N\left(\mu,\sigma^2/N\right)$:

$$w(\overline{x})=\left(1/\sqrt{2\pi\sigma/N}\right)\exp\left\{-N(\overline{x}-\mu)^2/2\sigma^2\right\}.$$

Последнее соотношение описывает хорошо известный факт уменьшения случайной погрешности измерения величины при повторении измерений пропорционально квадратному корню из числа измерений.

Кроме названных статистических характеристик используют начальные моменты k-го порядка

Начальный момент k-го порядка:

$$\nu_k = \int_{-\infty}^{\infty} x^k w(x)\,dx \qquad \left(h_k = (1/N)\sum_{i=1}^{N} x_i^k \right).$$

Центральный момент k–го порядка:

$$\mu_k = \int_{-\infty}^{\infty} (x-\mu)^k\, w(x)\,dx \qquad \left(m_k = \left(1/(N-1)\right) \sum_{i=1}^{N} (x_i - \overline{x})^k \right).$$

Обычно используют моменты первых четырех порядков. Для них существуют соотношения:

$$\mu_2 = \nu_2 - \nu_1^2; \qquad \left(m_2 = h_2 - h_1^2 \right);$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3; \qquad \left(m_3 = h_3 - 3h_1h_2 + 2h_1^3 \right);$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4; \qquad \left(m_4 = h_4 - 4h_1h_3 + 6h_1^2h_2 - 3h_1^4 \right).$$

Очевидно, что $\nu_1 = \mu$; $\nu_2 = \sigma^2$; $h_1 = \overline{x}$; $m_2 = s^2$.

Третий центральный момент используют для вычисления показателя асимметрии распределения

$$S_k = \mu_3 / \sigma^3; \qquad \left(S_{k\ выб} = m_3 / s^3 \right).$$

Четвертый центральный момент применяют для вычисления показателя эксцесса, являющегося характеристикой «островершинности» распределения

$$E_k = \mu_4 / \sigma^4 - 3; \qquad \left(E_{k\ выб} = m_4 / s^4 - 3 \right).$$

Заметное отличие от нуля величин S_k и E_k указывают на отклонение распределения от нормального.

Приведенные формулы дают численные, так называемые *точечные* оценки статистик. Оценки должны быть состоятельными, несмещенными и эффективными. Оценка характеристики называется:

состоятельной, если по мере увеличения объема выборки стремится к генеральному значению характеристики;

несмещенной, если ее математическое ожидание равно генеральному значению;

эффективной, если обладает минимальной по сравнению с другими оценками дисперсией.

Точечные оценки недостаточно полно характеризует результаты измерений из-за их естественного рассеяния. В каждой конкретной серии измерений значение оценки в той или иной степени отличается от полученных ранее и от истинного значения параметра. Поэтому необходимо знание величины погрешности, с которой оценен параметр, или, как говорят, получить *интервальную оценку* параметра. Ее обычно характеризуют *доверительным интервалом* для заданной вероятности p попадания предлагаемой оценки в этот интервал. Величину p выбирают, исходя из допустимой вероятности α того, что истинное значение измеряемой величины окажется вне указанного интервала.

Смысл понятия доверительные интервалы состоит в том, что, для любой малой вероятности α можно указать такое положительное значение ε, при котором разность между истинным значением параметра ϑ и его оценкой $\hat{\vartheta}$, полученной в результате измерений, по модулю не превышает ε с вероятность $p = 1 - \alpha$, то есть, если $\left|\vartheta - \hat{\vartheta}\right| < \varepsilon$, то

$$P\left(\hat{\vartheta} - \varepsilon < \vartheta < \hat{\vartheta} + \varepsilon\right) = 1 - \alpha .$$

Другими словами, если многократно, например N раз, воспроизводить выборки, и каждый раз находить доверительные интервалы для параметра ϑ, то в $p = (1 - \alpha) \cdot N$ случаях эти доверительные интервалы будут содержать истинное значение ϑ.