Глава 36

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ И МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ

36.1. Основные понятия математической статистики

Выборочным методом называют метод исследования общих свойств совокупности каких-либо объектов на основе изучения свойств лишь части этих объектов, взятых на выборку. Генеральной совокупностью называется множество однородных объектов, из которого выделяется некоторое подмножество, называемое выборочной совокупностью или выборкой. Объемом совокупности (генеральной или выборочной) называется число ее объектов. При изучении некоторого признака выборочной совокупности проводят испытания (наблюдения). Пусть посредством независимых испытаний, проведенных в одинаковых условиях, получены числовые значения $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, где n – объем выборки. Располагают эти значения в порядке возрастания:

$$x_1, x_2, \ldots, x_n \quad (x_1 \le x_2 \le \ldots \le x_n)$$

и называют полученную последовательность дискретным вариационным рядом, а сами значения x_i – вариантами. Среди вариант могут оказаться равные, тогда дискретный вариационный ряд можно записать так:

$$\begin{array}{cccc} x_1 & x_2 & \ldots & x_k, \\ n_1 & n_2 & \ldots & n_k, \end{array} \tag{36.1}$$

где n_i – частота появления значения x_i, причем

$$\sum_{i=1}^{k} n_i = n. \tag{36.2}$$

Относительной частотой w_i варианты x_i называется отношение ее частоты к объему выборки:

$$w_i = \frac{n_i}{n}, \quad \sum_{i=1}^{k} \frac{n_i}{n} = 1, \quad \sum_{i=1}^{k} w_i = 1.$$

Статистическим распределением выборки называется соответствие между вариантами и их частотами (или относительными частотами). Статистическое распределение может быть задано, например, с помощью таблицы.

Эмпирической функцией распределения (функцией распределения выборки) называется функция, определяющая для каждого значения x относительную частоту события $X < x$:

$$F^*(x) = n_x/n,$$

где n_x – число вариант, меньших x; n – объем выборки. Функция $F^*(x)$ обладает следующими свойствами: 1) $0 \le F^*(x) \le 1$; 2) $F^*(x)$ – неубывающая функция; 3) если a – наименьшая, b – наибольшая варианты, то $F^*(x) = 0$ при $x \le a$; $F^*(x) = 1$ при $x \ge b$.

Пусть случайная величина X имеет распределение $F(x, \alpha)$, содержащее неизвестный параметр α. Оценить параметр α – значит приближенно определить его значение по некоторой выборке $x_1, x_2, \ldots, x_n$. Оценку параметра α обозначим через $\tilde{\alpha} : \tilde{\alpha} = \tilde{\alpha}(x_1, x_2, \ldots, x_n)$. Оценка $\tilde{\alpha}$ параметра α называется несмещенной, если $M(\tilde{\alpha}) = \alpha$, и смещенной, если $M(\tilde{\alpha}) \ne \alpha$. Оценка $\tilde{\alpha}$ параметра α называется состоятельной, если $\lim\limits_{n \to \infty} P(|\tilde{\alpha} - \alpha| < \varepsilon) = 1$ при любом $\varepsilon > 0$. Оценка $\tilde{\alpha}$ называется эффективной, если при заданном n она имеет наименьшую дисперсию, т. е. $D(\tilde{\alpha}) = D_{\min}$.

Генеральной средней $x_г$ называется среднее арифметическое значений $x_1, x_2, \ldots, x_N$ генеральной совокупности объема N: $x_г = \dfrac{1}{N}\sum\limits_{i=1}^{N} x_i$.

Выборочной средней $x_в$ называется среднее арифметическое выборки $x_1, x_2, \ldots, x_n$ объема n: $x_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$, или

$$x_в = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \tag{36.3}$$

если выборка имеет вид (36.1).

Выборочную среднюю принимают в качестве оценки генеральной средней. Эта оценка является несмещенной и состоятельной, так как

$$M(X_в) = x_г, \quad \lim_{n \to \infty} P(|X_в - x_г| < \varepsilon) = 1.$$

Генеральной дисперсией $D_г$ называется среднее арифметическое квадратов отклонения значений генеральной совокупности $x_1, x_2, \ldots, x_N$ от их среднего значения $x_г$:

$$D_г = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_г)^2.$$

Генеральным средним квадратическим отклонением $\sigma_г$ называется корень квадратный из генеральной дисперсии: $\sigma_г = \sqrt{D_г}$.

Выборочной дисперсией $D_в$ называется среднее арифметическое квадратов отклонения значений выборки $x_1, x_2, \ldots, x_n$ от их среднего значения $x_в$:

$$D_в = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_в)^2 \text{ или } D_в = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - x_в)^2, \tag{36.4}$$

если выборка имеет вид (36.1).

Выборочное среднее квадратическое отклонение $\sigma_в$ определяется формулой

$$\sigma_в = \sqrt{D_в}. \tag{36.5}$$

Для вычисления выборочной дисперсии можно пользоваться формулой $D_в = \overline{x^2} - (x_в)^2$, где

$$x_в = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2$$

(n_i – частота x_i, $n_1 + n_2 + \cdots + n_k = n$). Аналогичная формула верна и для генеральной дисперсии.

Так как $M(D_в) = (n-1)/n)D_г$, т.е. $M(D_в) \neq D_г$, то выборочная дисперсия является смещенной оценкой генеральной дисперсии. Чтобы получить несмещенную оценку, генеральной дисперсии $D_г$, вводят понятие эмпирической (или исправленной) дисперсии s^2:

$$s^2 = \frac{n}{n-1}D_в, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - x_в)^2.$$

Для оценки генерального среднего квадратического отклонения служит исправленное среднее квадратическое отклонение, или эмпирический стандарт s:

$$s = \sqrt{s^2}, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - x_в)^2}. \tag{36.6}$$

В случае, когда все значения выборки $x_1, x_2, \ldots, x_n$ различны, т. е. $n_i = 1$, $k = n$, формулы для s^2 и s принимают вид

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_в)^2, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_в)^2}.$$

Если выборка задана в виде распределения равноотстоящих вариант, то выборочное среднее $x_в$ и выборочную дисперсию $D_в$ удобно находить методом произведений по формулам

$$x_в = M_1^* h + C, \quad D_в = (M_2^* - (M_1^*)^2)h^2, \tag{36.7}$$

где C – варианта, имеющая наибольшую частоту (ложный нуль), h – шаг, M_1^* – условный момент первого порядка, M_2^* – условный момент второго порядка;

h, u_i, M_1^*, M_2^* определяются соответственно формулами:

$$h = x_{i+1} - x_i \ (i = 1, 2, \ldots, n-1), \ u_i = (x_i - C)/h; \tag{36.8}$$

$$M_1^* = \frac{1}{n}\sum_{i=1}^{k} n_i u_i, \ M_2^* = \frac{1}{n}\sum_{i=1}^{k} n_i u_i^2; \tag{36.9}$$

u_i – условная варианта, n – объем выборки, n_i – частота варианты x_i $\left(\sum_{i=1}^{k} n_i = n\right)$.

Пример 36.1. Методом произведений найти выборочную среднюю, выборочную дисперсию и выборочное среднее квадратическое отклонение по данному статистическому распределению выборки:

x_i	10,2	10,9	11,6	12,3	13,0	13,7	14,4
n_i	8	10	60	12	5	3	2

Данная выборка является равноотстоящей, так как разности между двумя последующими вариантами постоянны: $x_{i+1} - x_i = h$ при $i = 1, 2, \ldots, 6$, причем $h = 0{,}7$. По формуле (36.2) находим $n = 100$. Наибольшую частоту имеет варианта $x_3 = 11{,}6$, т. е. $C = 11{,}6$. С помощью второй из формул (36.8) находим условные варианты u_i и составляем таблицу (табл. 36.1) значений величин, входящих в формулы (36.7). По формулам (36.9) находим $M_1^* = 13/100 = 0{,}13$, $M_2^* = 133/100 = 1{,}33$. С помощью формул (36.7) получаем $x_в = 0{,}13 \cdot 0{,}7 + 11{,}6 = 11{,}691 \approx 11{,}7$, $D_в = (1{,}33 - 0{,}13^2) \cdot 0{,}7^2 = 0{,}643419 \approx 0{,}64$. В соответствии с формулой (36.5) находим $\sigma_в = \sqrt{0{,}64} = 0{,}8$.

Таблица 36.1

i	x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$
1	10,2	8	–2	–16	32
2	10,9	10	–1	–10	10
3	11,6	60	0	0	0
4	12,3	12	1	12	12
5	13,0	5	2	10	20
6	13,7	3	3	9	27
7	14,4	2	4	8	32
Σ		100		13	133

36.2. Доверительный интервал. Доверительная вероятность

Оценка, определяемая одним числом, называется точечной. Оценка, определяемая двумя числами – концами интервалов, называется интервальной.

Доверительной вероятностью (надежностью) оценки $\tilde{\alpha}$ параметра α называется вероятность γ, с которой осуществляется неравенство $|\alpha-\tilde{\alpha}|<\delta$, т. е.

$$P(|\alpha-\tilde{\alpha}|<\delta)=\gamma \text{ или } P(\tilde{\alpha}-\delta<\alpha<\tilde{\alpha}+\delta)=\gamma.$$

Эта формула означает следующее: вероятность того, что интервал $(\tilde{\alpha}-\delta,\tilde{\alpha}+\delta)$ заключает в себе (покрывает) неизвестный параметр α, равна γ. Интервал $(\tilde{\alpha}-\delta,\tilde{\alpha}+\delta)$, который покрывает неизвестный параметр α с заданной надежностью γ, называется доверительным интервалом. Концы доверительного интервала называют доверительными границами.

Если случайная величина X имеет нормальное распределение с заданным средним квадратическим отклонением σ и неизвестным математическим ожиданием a, то

$$P\left(x_в-\frac{\sigma t}{\sqrt{n}}<a<x_в+\frac{\sigma t}{\sqrt{n}}\right)=\gamma, \tag{36.10}$$

где

$$\delta=\frac{\sigma t}{\sqrt{n}},\ 2\Phi(t)=\gamma, \tag{36.11}$$

т. е. доверительный интервал

$$I=(x_в-\sigma t/\sqrt{n},x_в+\sigma t/\sqrt{n}) \tag{36.12}$$

покрывает неизвестный параметр a с надежностью γ. Значение γ задано заранее; число $\Phi(t)$ определяется второй из формул (36.11); значение t находится с помощью таблиц значений функции Лапласа; точность оценки δ выражается первой из формул (36.11).

Пример 36.2. Найти доверительный интервал для оценки математического ожидания a нормальной случайной величины с надежностью $\gamma=0{,}95$, зная выборочную среднюю $x_в=75{,}15$, объем выборки $n=64$, среднее квадратическое отклонение $\sigma=8$.

Доверительный интервал определяется формулой (36.12). Чтобы найти концы доверительного интервала, необходимо знать значение t (значения $x_в, n, \sigma$ заданы). Второе из равенств (36.11) примет вид $2\Phi(t)=0{,}95$, откуда $\Phi(t)=0{,}475$. По таблице значений функции Лапласа находим $t=1{,}96$. Подставляя значения $x_в, \sigma, t, n$ в выражения для концов доверительного интервала, получаем

$$x_в-\frac{\sigma t}{\sqrt{n}}=75{,}15-\frac{8\cdot 1{,}96}{\sqrt{64}}=75{,}15-1{,}96=73{,}19;\ x_в+\frac{\sigma t}{\sqrt{n}}=77{,}11.$$

Следовательно, $73{,}19<a<77{,}11$, т.е. $(73{,}19;77{,}11)$ – искомый доверительный интервал.

36.3. Оценка точного значения измеряемой величины

Пусть в итоге n независимых измерений некоторой величины X получены следующие результаты:

$$x_1, x_2, \ldots, x_n. \tag{36.13}$$

Будем предполагать, что эти результаты свободны от грубых и систематических ошибок (неверные результаты отброшены, на систематические ошибки введены поправки).

Оценить точное значение a измеряемой величины – значит:

а) определить функцию $\alpha = \alpha(x_1, x_2, \ldots, x_n)$, которая обеспечивает достаточно близкое приближение к значению a;

б) указать границы интервала $(\alpha - \delta_1, \alpha + \delta_2)$, который с заданной вероятностью γ покрывает истинное значение a.

Среднее арифметическое значение (среднее значение) x результатов (36.13), среднее квадратическое отклонение s^* этих результатов от их среднего значения x и эмпирический стандарт s определяются соответственно формулами:

$$x = \frac{1}{n}\sum_{i=1}^{n} x_i; \tag{36.14}$$

$$s^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}; \tag{36.15}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2}. \tag{36.16}$$

Если все измерения проведены с одинаковой точностью, то в качестве оценки точного значения a измеряемой величины принимают среднее арифметическое значений результатов (36.13):

$$a \approx \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{36.17}$$

Эта оценка является несмещенной и состоятельной. Введенная оценка оказывается и эффективной при дополнительном предположении о том, что случайные ошибки измерений подчинены нормальному закону распределения. Это предположение имеется в виду и в дальнейшем. Оценка (36.17) относится к числу точечных оценок.

Симметрические доверительные оценки имеют вид

$$|a - x| < \delta, \text{ или } x - \delta < a < x + \delta \ (\delta > 0), \tag{36.18}$$

где x – среднее значение, определяемое формулой (36.14). Величина δ (точность оценки) определяется по заданной доверительной вероятности γ (надежности оценки).

Если известно среднее квадратическое отклонение σ, то доверительная оценка (36.18) имеет вид

$$|a - x| < \sigma t/\sqrt{n}, \tag{36.19}$$

где n – число измерений, а значение $t = t(\gamma)$ определяется по заданной доверительной вероятности γ из условия $2\Phi(t) = \gamma$ и находится с помощью таблиц. Точность оценки δ в этом случае выражается формулой

$$\delta = \sigma t/\sqrt{n}. \tag{36.20}$$

Если средняя квадратическая погрешность σ заранее неизвестна, то вместо нее применяют эмпирический стандарт s, который служит оценкой параметра σ. Доверительная оценка (36.18) принимает вид

$$|a - x| < st/\sqrt{n} \tag{36.21}$$

или

$$|a - x| < s^* t/\sqrt{k} \quad (k = n - 1), \tag{36.22}$$

где s^* и s определяются соответственно формулами (36.15) и (36.16), а множитель $t = t(\gamma, k)$ зависит не только от доверительной вероятности γ, но и от числа измерений n $(k = n - 1)$. Значения этого множителя определяются по таблицам.

Правило трех сигм представляет собой доверительную оценку

$$|a - x| < 3\sigma/\sqrt{n} \tag{36.23}$$

при известной величине σ или доверительную оценку

$$|a - x| < 3s/\sqrt{n} \tag{36.24}$$

при неизвестной величине σ. Оценка (36.23) имеет надежность $2\Phi(3) = 0{,}9973$ независимо от числа измерений. Оценка (36.24) зависит от числа измерений n (зависимость эта устанавливается с помощью соответствующих таблиц).

36.4. Оценки точности измерений

Предполагается, что измерения являются независимыми и равноточными (с одной и той же дисперсией), а их погрешности – случайными, причем распределены они по нормальному закону. В качестве показателя точности измерений оценивается дисперсия этого закона σ^2 или средняя квадратическая погрешность $\sigma = \sqrt{\sigma^2}$.

Точечные оценки дисперсии. 1. Если измеряют известную величину a, то в качестве эффективной оценки дисперсии σ^2 применяют квадрат среднего квадратического отклонения s^* результатов измерений (36.13) от значения a:

$$\sigma^2 \approx s^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2. \tag{36.25}$$

2. При измерениях неизвестной величины в качестве оценки дисперсии σ^2 применяют эмпирическую дисперсию s^2:

$$\sigma^2 \approx s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2, \tag{36.26}$$

где x – среднее арифметическое значений $x_1, x_2, \ldots, x_n$. Оценка (36.26) является несмещенной и состоятельной, но не является эффективной (она асимптотически

эффективна, т. е. ее дисперсия стремится к наименьшему значению при неограниченном увеличении числа измерений n).

3. Если проводится m серий измерений некоторой величины и известны количества измерений $n_1, n_2, \ldots, n_m$, а также средние арифметические результаты $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$ в каждой серии, то в качестве оценки дисперсии применяют эмпирическую дисперсию $\bar{s}^2$ из средних:

$$\sigma^2 \approx \bar{s}^2 = \frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{x}_i - x)^2, \tag{36.27}$$

где

$$x = \frac{1}{N}\sum_{i=1}^{m} n_i\bar{x}_i, \quad N = n_1 + \cdots n_m. \tag{36.28}$$

Эта оценка является несмещенной, состоятельной (и асимптотически эффективной при $m \to \infty$).

Доверительные оценки средней квадратической погрешности. При большом числе измерений доверительную оценку средней квадратической погрешности σ записывают в виде оценки относительного отклонения оцениваемого значения σ от эмпирического стандарта s (или s^*, или $\bar{s}$). Эта оценка имеет вид $|(\sigma - s)/s| < q$, или

$$s(1-q) < \sigma < s(1+q), \tag{36.29}$$

коэффициент $q = q(\gamma, k)$ находится с помощью соответствующих таблиц в зависимости от доверительной вероятности γ (надежности оценки) и от числа степеней свободы k ($k = 1$ в случае 1, $k = n-1$ в случае 2, $k = m-1$ в случае 3).

При малом числе измерений симметричная оценка (36.29) приводит к неоправданно большим доверительным интервалам; в этом случае применяют асимметричные доверительные оценки вида $sz_1 < \sigma < sz_2$, где s – эмпирический стандарт; значения коэффициентов $z_1 = z_1(\gamma, k)$, $z_2 = z_2(\gamma, k)$ находятся по таблицам.

36.5. Эмпирические формулы

Во многих науках (физика, химия, технические науки и др.) приходится пользоваться эмпирическими формулами, составленными на основании результатов наблюдений. Параметры эмпирических формул определяются по способу наименьших квадратов. Сначала устанавливается вид зависимости между двумя величинами. Это можно выполнить разными способами, например графически. Пусть результаты измерений представлены схемой

x	x_1	x_2	x_3	$\ldots$	x_n
y	y_1	y_2	y_3	$\cdots$	y_n

Упорядоченные пары чисел (x_i, y_i), $i = 1, 2, \ldots, n$, рассматриваются как прямоугольные декартовы координаты точек на плоскости $M_1(x_1, y_1)$, $M_2(x_2, y_2), \ldots, M_n(x_n, y_n)$. В выбранной системе координат строят точки $M_i(x_i, y_i)$, $i = 1, 2, \ldots, n$.

Если построенные точки $M_i(x_i, y_i)$ незначительно уклоняются от некоторой прямой, то полагают, что между величинами x и y существует линейная зависимость, т.е.

$$y = ax + b. \tag{36.30}$$

Параметры a и b эмпирической формулы (36.30) определяются из системы уравнений

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \quad a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i. \tag{36.31}$$

Если точки M_i $(i = 1, 2, \ldots, n)$ незначительно уклоняются от дуги некоторой параболы, то естественно предположить, что между величинами x и y существует квадратичная зависимость, т. е.

$$y = ax^2 + bx + c. \tag{36.32}$$

Параметры a, b, c эмпирической формулы (36.32) определяются из системы уравнений

$$\begin{aligned} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i^2, \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn &= \sum_{i=1}^{n} y_i. \end{aligned} \tag{36.33}$$

Пример 36.3. Экспериментально получены пять значений искомой функции $y = f(x)$ при пяти значениях аргумента:

x	1	2	3	4	5
y	4,7	5,7	4,2	2,2	2,7

Методом наименьших квадратов найти функцию $y = f(x)$ в виде $y = ax + b$.

Результаты измерений и их обработки запишем в табл. 36.2.

Таблица 36.2.

i	x_i	y_i	$x_i y_i$	x_i^2
1	1	4,7	4,7	1
2	2	5,7	11,4	4
3	3	4,2	12,6	9
4	4	2,2	8,8	16
5	5	2,7	13,5	25
Σ	15	19,5	51,0	55

Система уравнений (36.31) принимает вид

$$55a + 15b = 51{,}0, \quad 15a + 5b = 19{,}5.$$

Решая эту систему, находим $a = -0{,}75$, $b = 6{,}15$. Следовательно, получена эмпирическая формула $y = -0{,}75x + 6{,}15$.