Раздел 2. Математическая статистика.
9 Основные понятия и элементы выборочной теории.

Генеральной совокупностью называется множество всех мыслимых измерений некоторой случайной величины.

Выборочной совокупностью или **выборкой** называется некоторой множество значений генеральной совокупности, предназначенное для непосредственного исследования.

Количество элементов выборки n – называется **объемом выборки.**

Суть выборочного метода заключается в том, что по выборке делается вывод о генеральной совокупности в целом.

Ранжированным рядом называется выборка, упорядоченная по возрастанию.

Если выборка сделана из множества значений дискретной случайной величины, то она может быть сгруппирована в дискретный **вариационный ряд.**

Дискретный вариационный ряд или просто вариационный ряд – это соответствие между вариантами их частотами

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

или вариантами и их относительными частотами

x_i	x_1	x_2	…	x_k
w_i	w_1	w_2	…	w_n

Варианты x_i – это неповторяющиеся выборочные значения. **Частота варианты** n_i – это число, показывающее, сколько раз варианта x_i встречается в выборке.

Относительная частота варианты $w_i = \frac{n_i}{n}$.

Если выборка сделана из множества значений непрерывной случайной величины, то она может быть сгруппирована в **интервальный вариационный ряд.**

Интервальный вариационный ряд или просто интервальный ряд – это соответствие между частичными интервалами (интервалами группировки) их частотами (или относительными частотами).

$a_i - a_{i+1}$	$a_1 - a_2$	$a_2 - a_3$	…	$a_k - a_{k+1}$

n_i	n_1	n_2	…	n_k

Частота интервала n_i - это число, показывающее, сколько выборочных данных попало в интервал $[a_i; a_{i+1})$.

Накопленной частотой действительного числа x – называется количество выборочных данных, лежащих левее x на числовой оси. Обозначается – n_x.

Накопленной частотой i-ого интервала называется количество выборочных данных, лежащих от начала выборки до конца этого интервала. Обозначается – $n_i^{\text{накопл}}$.

Полигон частот – это ломаная линия с узлами в точках (x_i, n_i)

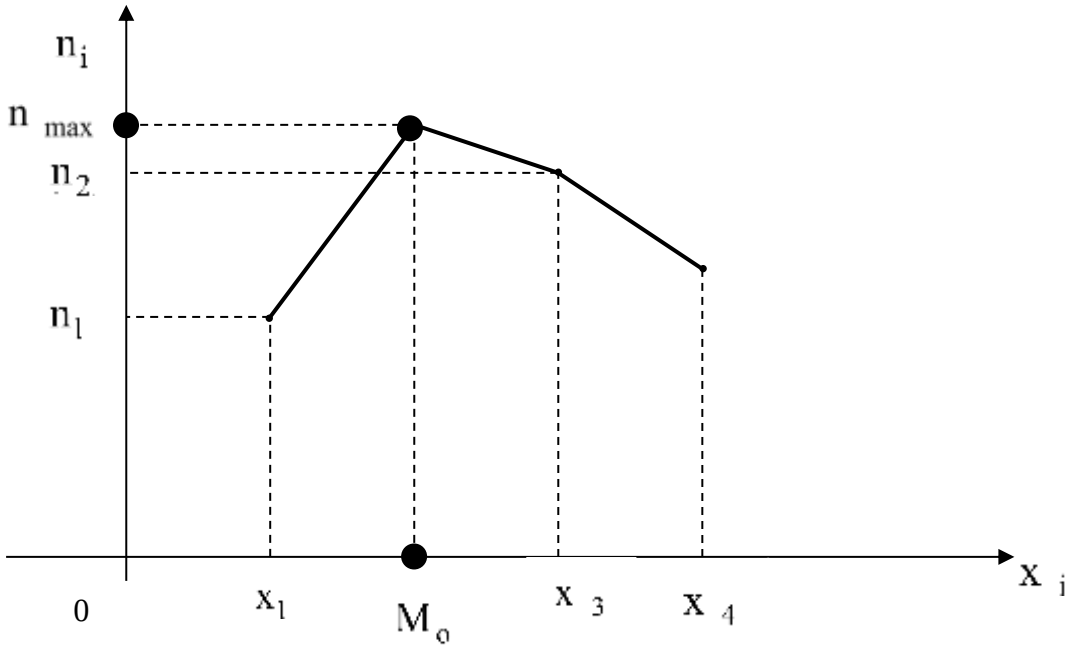

Рис. 10.1. Полигон частот

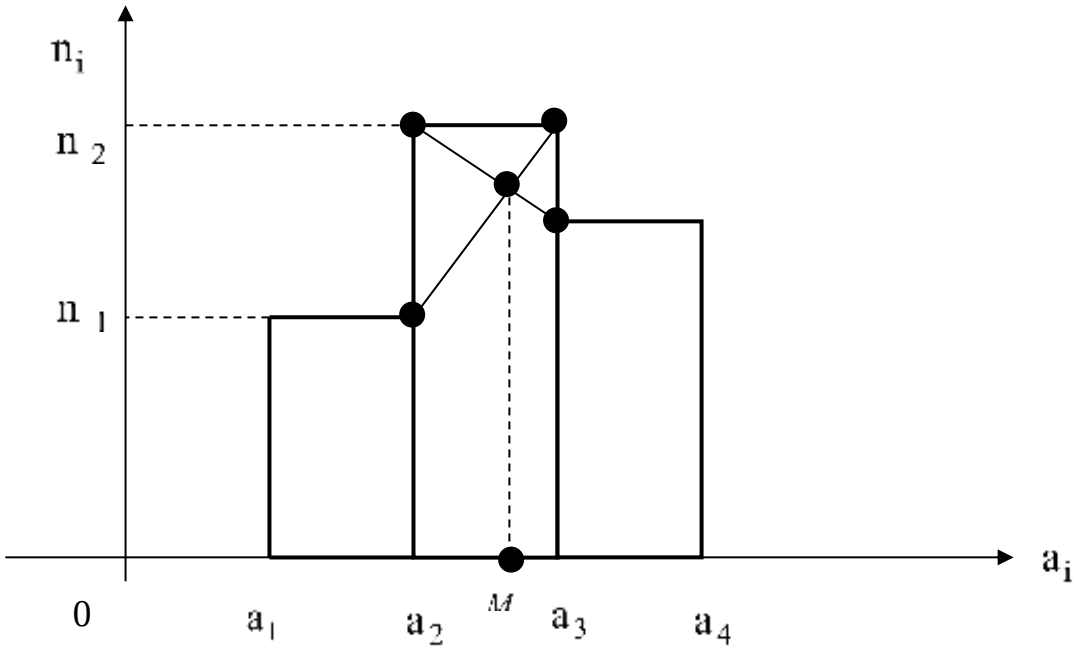

Рис. 10.2. Гистограмма частот

По **полигону** можно найти **моду** дискретного вариационного ряда.

Гистограмма – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых являются частичные интервалы, а высоты соответствуют частоте. По **гистограмме** можно найти **моду** интервального ряда. Полигон частот и гистограмма частот приведены на рис. 10.1 и 10.2.

Кумулята – это ломаная линия, с узлами в точке (x_i, n_{x_i}) для дискретного вариационного ряда и с узлами в точках (a_i, n_{a_i}) для интервального ряда.

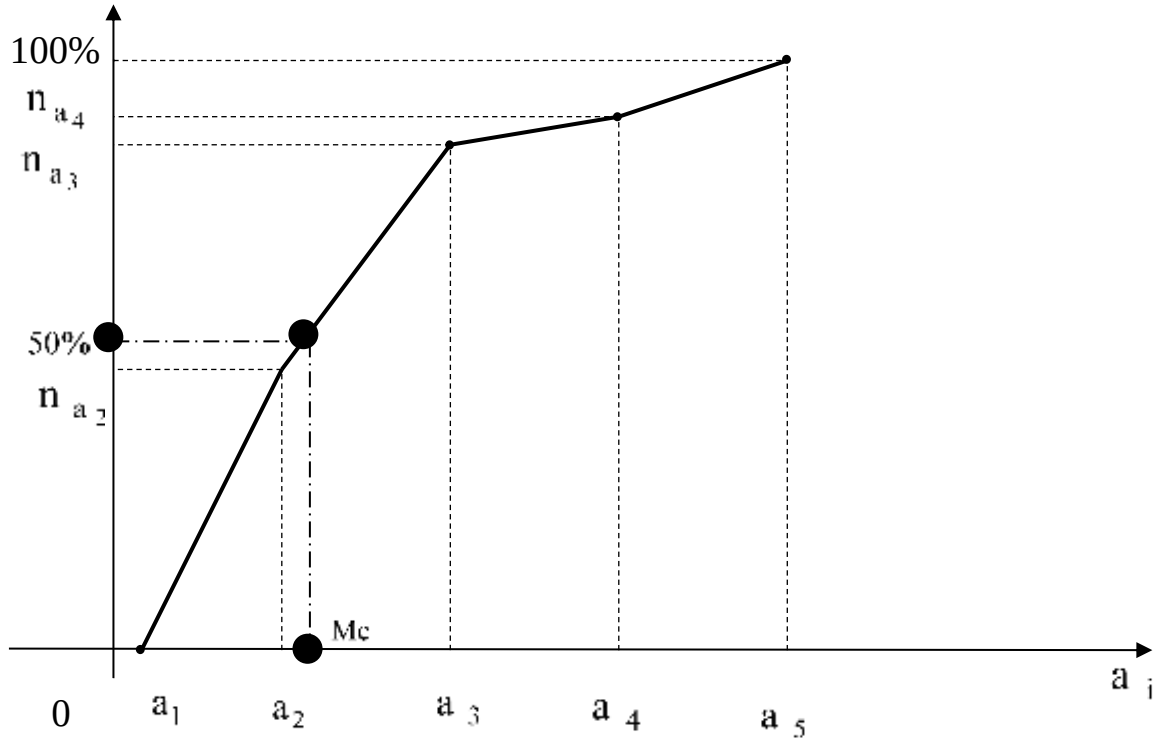

Рис. 10.3. Кумулята интервального ряда

По **кумуляте** можно найти **медиану** интервального ряда.

Эмпирическая функция распределения находится по формуле:

$$F_n(x) = \frac{n_x}{n}.$$

Здесь n – это объем выборки; n_x – это накопленная частота числа **x**, то есть число выборочных данных, строго меньших **x**.

Свойства функции $F_n(x)$:

1. $0 \le F_n(x) \le 1$.
2. $F_n(x)$ – неубывающая функция, т. е. $F_n(x_2) > F_n(x_1)$, если $x_2 \ge x_1$.
3. Если x_1 – наименьшая варианта, то $F_n(x) = 0$, при $x \le x_1$.
4. Если x_2 – наибольшая варианта, то $F_n(x) = 1$, при $x > x_2$.

Эмпирическая функция распределения – ступенчатая. Необходимо разбить ось на интервалы точками x_i, и воспользоваться формулой для каждого интервала в отдельности.

Пример 10.1. Найти эмпирическую функцию распределения $F_n(x)$ по распределению выборки.

x_i	15	20	25	30	35
n_i	10	15	30	20	25

Решение.

Воспользуемся формулой:

$$F_n(x) = \frac{n_x}{n},$$

где n – объем выборки (n=10+15+30+20+25=100),

n_x – число вариант, меньших аргумента x. Так как $F_n(x)$ является кусочно-постоянной (ступенчатой), разобьем область определения R на интервалы постоянства функции (см. рис. 10.4).

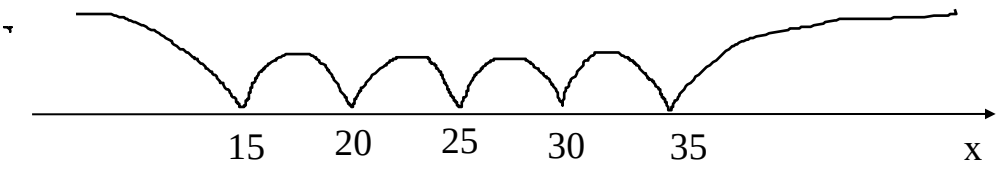

Рис. 10.4. Интервалы постоянства эмпирической функции распределения

1. При $x \le 15$ вариант, меньших x в выборке нет, то есть $n_x = 0$.

$$F_n(x) = \frac{n_x}{n} = \frac{0}{100} = 0.$$

2. При $15 < x \le 20$ варианты, меньшие x – это 10 вариант, каждая из которых равна 15, то есть n_x=10.

$$F_n(x) = \frac{10}{100} = 0{,}1.$$

3. При $20 < x \le 25$ вариант, меньших x – двадцать пять: 10 – равных 15 и 15 – равных 20, то есть n_x=10+15=25.

$$F_n(x) = \frac{25}{100} = 0{,}25.$$

4. При $25 < x \le 30$ вариант, меньших x – пятьдесят пять: 10 – равных 15, 15 – равных 20, 30 – равных 25, то есть n_x =10+15+25=55.

$$F_n(x) = \frac{55}{100} = 0{,}55.$$

5. При $30 < x \le 35$ вариант, меньших x – семьдесят пять: 10 – равных 15, 15 – равных 20, 30 – равных 25, 20 – равных 30, то есть n_x=10+15+25+30=75.

$$F_n(x) = \frac{75}{100} = 0{,}75.$$

5. При $x > 35$ все 100 вариант меньше x.

$$F_n(x) = \frac{100}{100} = 1.$$

Таким образом, эмпирическая функция распределения имеет

вид:

$$F_n(x)=\begin{cases}0, & x\le 15\\ 0{,}1, & 15<x\le 20\\ 0{,}25, & 20<x\le 25\\ 0{,}55, & 25<x\le 30\\ 0{,}75, & 30<x\le 35\\ 1, & x>35\end{cases}.$$

Точечные оценки параметров распределения

Точечной называется статистическая оценка, которая определяется одним числом $\Theta^* = f(x_1, x_2, \dots, x_n)$, где $x_1, x_2, \dots, x_n$ - результаты n наблюдений над количественным признаком X.

Точечная оценка Θ^* называется **несмещенной**, если

$$M(\Theta^*) = \Theta,$$

где Θ – оцениваемый параметр теоретического распределения. Если $M(\Theta^*) \neq \Theta$, статистическая оценка называется **смещенной.**

Точечная оценка Θ^* называется **состоятельной**, если

$$\lim_{n\to\infty} P\left(\left|\Theta^*(x_{1,} x_{2,\dots,} x_n) - \Theta\right| < \varepsilon \right) = 1,$$

где ε - сколь угодно малое положительное число.

Точечная оценка Θ^* называется **эффективной**, если данная оценка имеет наименьшую возможную дисперсию среди оценок данного параметра, сделанным по выборкам одинакового объема.

Оценки меры центральной тенденции

Мода выборки – это **наиболее часто встречающееся** выборочное значение.

Для непрерывной выборки, заданной в виде интервального ряда, мода находится по формуле:

$$Mo = x_0 + h \cdot \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})},$$

где x_0 – начало модального интервала; h – его длина; n_i – его частота; n_{i-1} – частота интервала, предшествующего модальному; n_{i+1} – частота интервала, следующего за модальным. Моду можно найти по полигону или гистограмме (см. рис. 10.1 и 10.2).

Медиана выборки – это середина ранжированного ряда. Иначе говоря – это точка числовой оси, левее и правее которой лежит по 50 % выборочных данных.

Для дискретного вариационного ряда медиана находится по формуле:

$$Me = x_{\frac{n+1}{2}},$$

если n – нечетное число;

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2},$$

если n – четное.

Для непрерывной выборки, заданной в виде интервального ряда, медиана находится по формуле:

$$Me = x_0 + h \cdot \frac{n/2 - n_{i-1}^{накопл}}{n_i},$$

где n – объем выборки; x_0 – начало медианного интервала; h – его длина; n_i – его частота; $n_{i-1}^{накопл}$ – накопленная частота интервала, предшествующего медианному. Медиану интервального ряда можно найти по кумуляте (см. рис. 10.3).

Выборочное среднее – это точечная оценка математического ожидания генеральной совокупности.

Для несгруппированной выборки формула для нахождения выборочного среднего имеет вид:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Здесь **n** – объем выборки, x_i – выборочные значения.

Для сгруппированной выборки формула для нахождения выборочного среднего имеет вид:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i .$$

Здесь **n** – объем выборки, **k** – количество групп выборки, x_i – варианты, n_i – соответствующие им частоты.

Для интервального ряда в последней формуле вместо x_i берут середины интервалов: $x_i = \frac{a_i + a_{i+1}}{2}$; n_i – частоты интервалов.

Оценки меры изменчивости

Выборочная дисперсия – это точечная оценка дисперсии генеральной совокупности.

Для несгруппированной выборки формула для нахождения выборочной дисперсии имеет вид:

$$D = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 .$$

Здесь n – объем выборки, x_i – выборочные значения.

Для сгруппированной выборки формула для нахождения выборочной дисперсии имеет вид:

$$D = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2 n_i .$$

Здесь n – объем выборки, k – количество групп выборки, x_i – варианты, n_i – соответствующие им частоты.

Для интервального ряда в последней формуле вместо x_i берут середины интервалов: $x_i = \frac{a_i + a_{i+1}}{2}$; n_i – частоты интервалов.

Выборочное среднеквадратическое отклонение – это точечная оценка среднеквадратического отклонения генеральной совокупности:

$$\sigma = \sqrt{D} .$$

Исправленная дисперсия – это наилучшая оценка генеральной дисперсии.

Исправленная дисперсия находится по формуле:

$$S^2 = \frac{n}{n-1} D .$$

Исправленной среднеквадратическое отклонение или **стандартное отклонение** – это наилучшая оценка среднеквадратического отклонения генеральной совокупности. Его формула имеет вид:

$$S = \sqrt{S^2} .$$

Исправленная дисперсия и стандартное отклонение являются **несмещенными оценками**, т. е. оценками, которые не дают систематической ошибки.

Характеристики формы кривой

Асимметрия эмпирического распределения

$$as = \frac{1}{n\sigma^3} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3$$

или

$$as = \frac{1}{n\sigma^3} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^3 \cdot n_i .$$

Асимметрия характеризует скошенность кривой распределения генеральной совокупности относительно математического ожидания.

Эксцесс эмпирического распределения

$$ex = \frac{1}{n\sigma^4} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3 - 3,$$

или

$$ex = \frac{1}{n\sigma^4} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^4 \cdot n_i - 3.$$

Эксцесс характеризует островершинность (плосковершинность) кривой распределения генеральной совокупности относительно нормального распределения.

Для интервального ряда в формулах асимметрии и эксцесса для сгруппированной выборки вместо x_i берут середины интервалов: $x_i = \frac{a_i + a_{i+1}}{2}$; n_i – частоты интервалов.

Пример 10.2.

Найти выборочное среднее и выборочную дисперсию по распределению выборки объема n=10:

x_i	26,1	26,3	26,7	27,4
n_i	2	3	4	1

Решение.

Если варианты являются дробными числами (особенно дробями много меньшими единицы), то целесообразно перейти при расчете к условным вариантам, умножив каждое x_i на коэффициент масштабирования k. Если взять k=10, то получим числа без десятичной запятой: 261, 263, 267, 274.

Если полученные условные варианты являются большими числами, имеет смысл прибавить к каждому из них условный ноль C, с целью преобразовать их в варианты небольшие по абсолютному значению. В данном случае удобно положить C = - 267.

Таким образом, преобразование x_i к условным вариантам u_i в данном случае заключается в следующем:

$u_i = kx_i + C = 10x_i - 267$.

Распределение условных вариант имеет вид:

u_i	-6	-4	0	7
n_i	2	3	4	1

Для нахождения выборочного среднего и выборочной дисперсии с помощью условных вариант используем соответствующие формулы, заменив в них x_i на u_i. Тогда условное среднее примет вид:

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{k} u_i n_i = \frac{1}{10}\cdot(-6\cdot 2 - 4\cdot 3 + 0\cdot 4 + 7\cdot 1) = \frac{-12-12+0+7}{10} = -1{,}7$$

.

Нетрудно показать, что для того, чтобы выразить из полученного результата $\bar{x}$, достаточно воспользоваться формулой:

$$\bar{x} = \frac{\bar{u} - C}{k}.$$

Откуда $\bar{x} = \frac{-1{,}7-(-267)}{10} = \frac{265{,}3}{10} = 26{,}53.$

Найдем далее дисперсию в условных вариантах:

$$D_u = \frac{1}{n}\sum_{i=1}^{k} u_i^{\,2}\cdot n_i - \bar{u}^2 = \frac{1}{10}\left((-6)^2\cdot 2 + (-4)^2\cdot 3 + 0^2\cdot 4 + 7^2\cdot 1\right) - (-1{,}7)^2 =$$

$$= \frac{72+48+0+49}{10} - 2{,}89 = 16{,}9 - 2{,}89 = 14{,}01.$$

Искомую выборочную дисперсию (для первоначальных вариант) найдем по формуле:

$$D_x = \frac{D_u}{k^2} = \frac{14{,}01}{100} = 0{,}1401.$$

10Оценивание неизвестных параметров распределений Аделений.

Метод максимального правдоподобия

Пусть дана выборка $x_1, x_2, \ldots, x_n$ объема n из генеральной совокупности X и задан закон распределения $p(x, \Theta)$ с точностью до неизвестного параметра Θ. Для того, чтобы получить точечную оценку Θ^* неизвестного параметра Θ с помощью метода максимального правдоподобия необходимо:

1. Построить функцию правдоподобия
$$L(x_1, x_2, \ldots x_n, \Theta) = p(x_1, \Theta) \cdot p(x_2, \Theta) \cdot \ldots \cdot p(x_n, \Theta);$$
2. Найти логарифмическую функцию правдоподобия $\ln L$;
3. Найти точку максимума $\ln L$, для чего решить уравнение:
$$\frac{\partial \ln L}{\partial \Theta} = 0.$$

Точка Θ^*, доставляющая максимум функции $\ln L$, является **оценкой максимального правдоподобия неизвестного параметра** Θ.

Замечание. Все переменные функции правдоподобия, кроме Θ, считаются фиксированными.

Метод моментов

Метод моментов нахождения точечных оценок неизвестных параметров распределения, состоит в приравнивании теоретических моментов эмпирическим моментам того же порядка.

Эмпирический начальный момент находится по выборке и имеет вид:
$$\tilde{\nu}_k = \frac{1}{n} \sum x_i^{\,k} n_i,$$
и эмпирический центральный момент k-ого порядка:
$$\tilde{\mu}_k = \frac{1}{n} \sum (x_i - \bar{x})^k n_i.$$

Если распределение имеет один неизвестный параметр, то для его отыскания приравнивают один теоретический момент одному эмпирическому моменту того же порядка. Например $\nu_1(X) = \tilde{\nu}_1$. Но так как $\nu_1(X) = \sum x_i p_i$ – математическое ожидание, а $\tilde{\nu}_1 = \frac{1}{n} \sum x_i n_i$ –

выборочное среднее, то приходим к равенству: $M(X) = \bar{x}$, из которого и находится неизвестный параметр распределения.

Метод наименьших квадратов

Метод наименьших квадратов (МНК) – это метод нахождения точечных оценок неизвестных параметров распределения. Часто он используется для нахождения оценок параметров зависимости между случайными величинами X и Y.

Пусть X и Y связаны зависимостью вида $y = f(x)$. Пусть даны результаты измерений (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Чтобы найти неизвестные параметры зависимости вычисляют рассогласования $\varepsilon_i = y_i - f(x_i)$, возводят их в квадрат, чтобы исключить их взаимное уничтожение из-за разных знаков, затем складывают. Полученную сумму минимизируют, находя, тем самым, оценки неизвестных параметров зависимости.

$$Q = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min .$$

Пусть зависимость между X и Y имеет линейный вид, т. е. $y = ax + b$. Чтобы найти оценки неизвестных параметров **a** и **b** необходимо:

1. С помощью расчетной таблицы (табл. 11.1) рассчитать коэффициенты системы линейных алгебраических уравнений (СЛАУ).

Таблица 11.1.

i	x_i	y_i	x_i^2	$x_i y_i$
1	x_1	y_1	x_1^2	$x_1 y_1$
2	x_2	y_2	x_2^2	$x_2 y_2$
…	…	…	…	…
n	x_n	y_n	x_n^2	$x_n y_n$
$\sum$	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i y_i$

2. Подставить полученные коэффициенты в СЛАУ:

$$a\sum x_i^2 + b\sum x_i = \sum x_i y_i ;$$
$$a\sum x_i + bn = \sum y_i .$$

3. Решить СЛАУ любым известным методом.

4. Построить в координатных осях данные точки (x_i, y_i) и полученную прямую и убедится в адекватности модели объекту.

Аналогичным образом находят оценки неизвестных параметров **a**, **b** и **c** при параболической зависимости между X и Y, т. е. $y = ax^2 + bx + c$, решив СЛАУ:

$$a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 = \sum x_i^2 y_i,$$
$$a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i y_i,$$
$$a\sum x_i^2 + b\sum x_i + cn = \sum y_i .$$

Аналогичным образом находят оценки неизвестных параметров **a** и **b** при гиперболической зависимости между X и Y, т. е. $y = \frac{a}{x} + b$, решив СЛАУ:

$$a\sum \frac{1}{x_i^2} + b\sum \frac{1}{x_i} = \sum \frac{y_i}{x_i},$$
$$a\sum \frac{1}{x_i} + bn = \sum y_i .$$

Статистической называется зависимость Y от X, при которой изменение X влечет за собой изменение распределения Y.

Корреляционной зависимость Y от X называется функциональная зависимость условной средней $M(Y/X)$ от X, то есть зависимость вида:

$$M(Y/X) = f(x), \qquad (12.1)$$

где условное среднее $M(Y/X)$ – это математическое Y, соответствующее значению X = x. Уравнение (12.1) называется **уравнением регрессии**, а функция f(x), называется регрессией Y на X, а ее график – линией регрессии Y на X.

Теория корреляции решает две задачи:

– установление формы корреляционной связи, т. е. вида функции регрессии;

– оценка тесноты корреляционной связи.

Пусть даны результаты измерений $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Уравнение линейной регрессии Y по X имеет вид:

$$\bar{y}_x - \bar{y} = \rho_{YX}(x - \bar{x}).$$

Здесь $\bar{y}_x$ – зависимая переменная (условное среднее значений величины Y, при условии, что X = x);

x – независимая переменная;

$\rho_{YX} = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$ – коэффициент регрессии Y по X;

$\bar{x} = \dfrac{1}{n}\sum x_i$ – среднее по x; $\bar{y} = \dfrac{1}{n}\sum y_i$ – среднее по y;

$\overline{x^2} = \dfrac{1}{n}\sum x_i^2$ – среднее квадратов; $\overline{xy} = \dfrac{1}{n}\sum x_i y_i$ – среднее произведений.

Уравнение линейной регрессии можно записать и через выборочный коэффициент корреляции:

$$\bar{y}_x - \bar{y} = r_{YX}\,\frac{\sigma_Y}{\sigma_X}(x - \bar{x}).$$

Здесь σ_X и σ_Y – выборочные среднеквадратические отклонения величин X и Y.

$r_{YX} = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\sigma_X \sigma_Y}$ – выборочный коэффициент корреляции.

Свойства r_{YX}:

1. $|r_{YX}| \le 1$.

2. Если $|r_{YX}| = 1$, то связь между X и Y наиболее тесная – линейная.

3. Если $r_{YX} = 1$, то связь прямая, если $r_{YX} = -1$, то связь обратная.

4. Если X и Y независимы, то $r_{YX} = 0$.

5. Если $r_{YX} = 0$, то X и Y являются некоррелированными, т.е. между ними нет корреляционной связи.

Пусть эмпирические данные представлены в виде корреляционной таблицы 12.2.

Таблица 12.2

Y \ X	y_1	y_2	…	y_j	…	y_m
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}
…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}
…	…	…	…	…	…	…

x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}

В таблице $x_1, x_2, \ldots, x_k$ – значения случайной величины X; $y_1, y_2, \ldots, y_m$ – значения случайной величины Y; n_{ij} – частота появления в выборке пары (x_i, y_j).

Тогда выборочный коэффициент корреляции может быть найден по формуле:

$$r_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij} - n\bar{x}\bar{y}}{n\sigma_X\sigma_Y},$$

где n – сумма всех частот корреляционной таблицы (объем выборки).