ВАРИАЦИОННЫЕ РЯДЫ И ИХ ХАРАКТЕРИСТИКИ

ГЛАВА 1

Содержание

§ 1. Первичная обработка результатов наблюдений

Информация о работе любой отрасли производственной сферы ставит перед ее руководством и наукой задачу: как, сведя к минимуму расходы по использованию природных, материальных и людских ресурсов, эффективно анализировать работу отрасли, управлять ею, прогнозировать развитие возможных сценариев поведения отрасли как сложной системы.

Это означает, что математическому моделированию подлежит (дискретный, непрерывный, фрактальный) информационный поток — *статистическая совокупность* — в виде случайных событий и случайных величин.

§ 1. Первичная обработка результатов наблюдений

Изучение подобного рода *массовых явлений*, выявление их статических и динамических закономерностей становится *обработки данных*.

Среди полезной информации о статистической совокупности особый интерес представляют *статистические данные*, которые можно записать в виде ряда {*x*(1), *x*(2), ..., *x*(*n*)} числовых значений интересующего нас *признака* (случайной величины *X*).

Обработку этого ряда производят посредством *методов* математической статистики; при этом *точность* статистических методов повышается с ростом n.

§ 1. Первичная обработка результатов наблюдений

Пусть A — некоторое множество (например, множество всех жителей данного города), B Ì A — случайно выбранное подмножество (например, множество случайно выбранных жителей, при этом некий наблюдатель измерил у них рост, скажем, в сантиметрах).

Выборочным методом называется метод исследования общих свойств множества A на основе изучения так называемых статистических свойств лишь множества B.

Множество A называется генеральной совокупностью, а множество B — выборочной совокупностью или выборкой.

Число N = | A | элементов множества A называется объемом генеральной совокупности, а число n = | B | — объемом выборки. При изучении некоторого признака X (в нашем примере — роста) выборки производят испытания или наблюдения (измерение роста).

§ 1. Первичная обработка результатов наблюдений

Пусть в результате независимых испытаний, проведенных в одинаковых условиях, получены числовые значения {x(1), x(2), ..., x(*n*)}, где *n* — объем выборки.

При обработке статистических данных строятся статистики. Статистикой называется функция

$$\begin{array}{ccc} \mathbf{R}^n & \xrightarrow{\ f\ } & \mathbf{R} \\ \cup & & \cup \\ (x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x & \mapsto & f(x), \end{array}$$

которая набору значений (x(1), x(2), ..., x(n)), случайной величины ставит в соответствие по некоторому правилу f действительное число.

§ 1. Первичная обработка результатов наблюдений

Статистика является числовой функцией на множестве реализаций случайной величины. Значения x (*i*) располагают в порядке возрастания:

$$x_1, x_2, \ldots, x_n \quad (x_1 < x_2 < \ldots < x_n).$$

Может оказаться, что некоторые варианты xi в выборке встречаются несколько раз.

Число *ni*, показывающее, сколько раз встречается варианта *xi* в выборочной совокупности, называется ее частотой (эмпирической частотой).

§ 1. Первичная обработка результатов наблюдений

Частоты вариант называются их весами. Отношение $wi = ni / n$ частоты ni к объему n выборки называют относительной частотой (частостью) варианты xi.

Вариационным рядом (или статистическим распределением) называется ранжированный в порядке возрастания или убывания ряд вариант с соответствующими им весами.

§ 1. Первичная обработка результатов наблюдений

Различают дискретные и непрерывные вариационные ряды. Дискретный вариационный ряд записывают в виде табл. 1.

Таблица 1

Варианты, x_i	x_1	x_2	. . .	x_k
Частоты, n_i	n_1	n_2	. . .	n_k

Здесь *ni* — частота появления значения *xi*, причем

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

§ 1. Первичная обработка результатов наблюдений

Если объем *n* выборки большой (n > 30), то результаты наблюдений сводят в интервальный вариационный ряд, который формируется следующим образом. Вычисляют размах *R* варьирования признака *X*, как разность между наибольшим *X*max и наименьшим *X*min значениями признака:

$$R = x_{\max} - x_{\min} .$$

§ 1. Первичная обработка результатов наблюдений

Размах *R* варьирования признака X делится на k равных частей и таким образом определяется число столбцов (интервалов) в таблице. Число k частичных интервалов выбирают, пользуясь одним из следующих правил:

$$1)\ 6 \le k \le 20, \qquad 2)\ k \approx \sqrt{n}, \qquad 3)\ k \approx 1 + \log_2 n \approx 1 + 3{,}221 \cdot \lg n.$$

При небольшом объеме *n* выборки число *k* интервалов принимают равным от 6 до 10.

Длина *h* каждого частичного интервала определяется по формуле: $h = R / k$

§ 1. Первичная обработка результатов наблюдений

Величину h обычно округляют до некоторого значения d. Например, если результаты *xi* признака X — целые числа, то h округляют до целого значения, если *xi* содержат десятичные знаки, то h округляют до значения d, содержащего такое же число десятичных знаков.

Затем подсчитывается частота *ni*, с которой попадают значения xi признака X в *i*-й интервал. Значение *xi*, которое попадает на границу интервала, относят к какому либо определенному концу, например, к левому.

За начало первого интервала рекомендуют выбрать величину

$$x_0 = X\min - 0{,}5h.$$

Конец последнего интервала находят по формуле

$$x_k = X\max + 0{,}5h.$$

§ 1. Первичная обработка результатов наблюдений

Сформированный интервальный вариационный ряд записывают в виде табл. 2.

Таблица 2

Варианты-интервалы, $(x_{i-1}; x_i)$	$(x_0; x_1)$	$(x_1; x_2)$	. . .	$(x_{k-1}; x_k)$
частоты, n_i	n_1	n_2	. . .	n_k

Интервальный вариационный ряд изображают в виде гистограммы частот *ni* или гистограммы относительных частот *wi = ni / n.*

§ 1. Первичная обработка результатов наблюдений

Гистограммой называется ступенчатая фигура, для построения которой по оси абсцисс откладывают отрезки, изображающие частичные интервалы (x_{i-1}; x_i) варьирования признака X, и на этих отрезках, как на основаниях, строят прямоугольники с высотами, равными частотам или *частостям* соответствующих интервалов.

Для расчета статистик (выборочной средней, выборочной дисперсии, асимметрии и эксцесса) переходят от интервального вариационного ряда к дискретному. В качестве вариант x_i этого ряда берут середины интервалов (x_i ; x_{i+1}).

§ 1. Первичная обработка результатов наблюдений

Дискретный вариационный ряд записывается в виде табл. 3 или табл. 4.

Таблица 3

Варианты, x_i	x_1	x_2	. . .	x_k
частоты, n_i	n_1	n_2	. . .	n_k

Здесь $\sum n_i = n$, где n — объем выборки.

Таблица 4

Варианты, x_i	x_1	x_2	. . .	x_k
относительные частоты, $w_i = n_i / n$	w_1	w_2	. . .	w_k

Здесь $\sum_{i=1}^{k} w_i = 1$.

§ 1. Первичная обработка результатов наблюдений

Графически дискретный вариационный ряд изображают в виде *полигона* частот (соответственно в виде полигона относительных частот) следующим образом.

Сначала на числовой плоскости строят точки (*xi* ; *ni*) (точки (*xi* ; *wi*)), где *xi* — *i*-я варианта, число *ni* (число *wi*) называют *частотой* (*частостью*).

Затем строят ломаную, соединяющую построенные точки, которую и называют *полигоном*.

§ 1. Первичная обработка результатов наблюдений

Вариационные ряды графически можно изобразить в виде *кумулятивной кривой* (*кривой сумм* — *кумуляты*).

При построении кумуляты дискретного вариационного ряда на оси абсцисс откладывают варианты *xi* , а по оси ординат соответствующие им *накопленные частоты Wi* .

Соединяя точки (*xi* ; *Wi*) отрезками, получаем ломаную, которую называют *кумулятой*. Для получения накопленных частот и дальнейшего построения точек (*xi* ; *Wi*) составляется расчетная табл. 5.

Таблица 5

Варианты, x_i	x_1	x_2	. . .	x_k
Относительные частоты, $w_i = n_i / n$	$w_1 = n_1 / n$	$w_2 = n_2 / n$	. . .	$w_k = n_k / n$
Накопленные относительные частоты, $W_i = W_{i-1} + w_i$	$W_1 = w_1$	$W_2 = W_1 + w_2$	. . .	$W_k = W_{k-1} + w_k$

§ 1. Первичная обработка результатов наблюдений

Для характеристики свойств статистического распределения вводится понятие эмпирической функции распределения.

Эмпирической функцией распределения или *функцией распределения* называется функция *F*в (*x*), определяемая равенством:

$$F_{\text{в}}(x) = \frac{n_x}{n},$$

где *n* — объем выборки, *nx* — число вариант *xi* , меньших *x*.

§ 1. Первичная обработка результатов наблюдений

Эмпирическая функция *F*в(*x*) служит для оценки теоретической функции распределения генеральной совокупности.

Значения эмпирической функции *F*в(*x*) принадлежат промежутку [0; 1]; ее графиком служит кусочно-постоянная кривая (рис. 1).

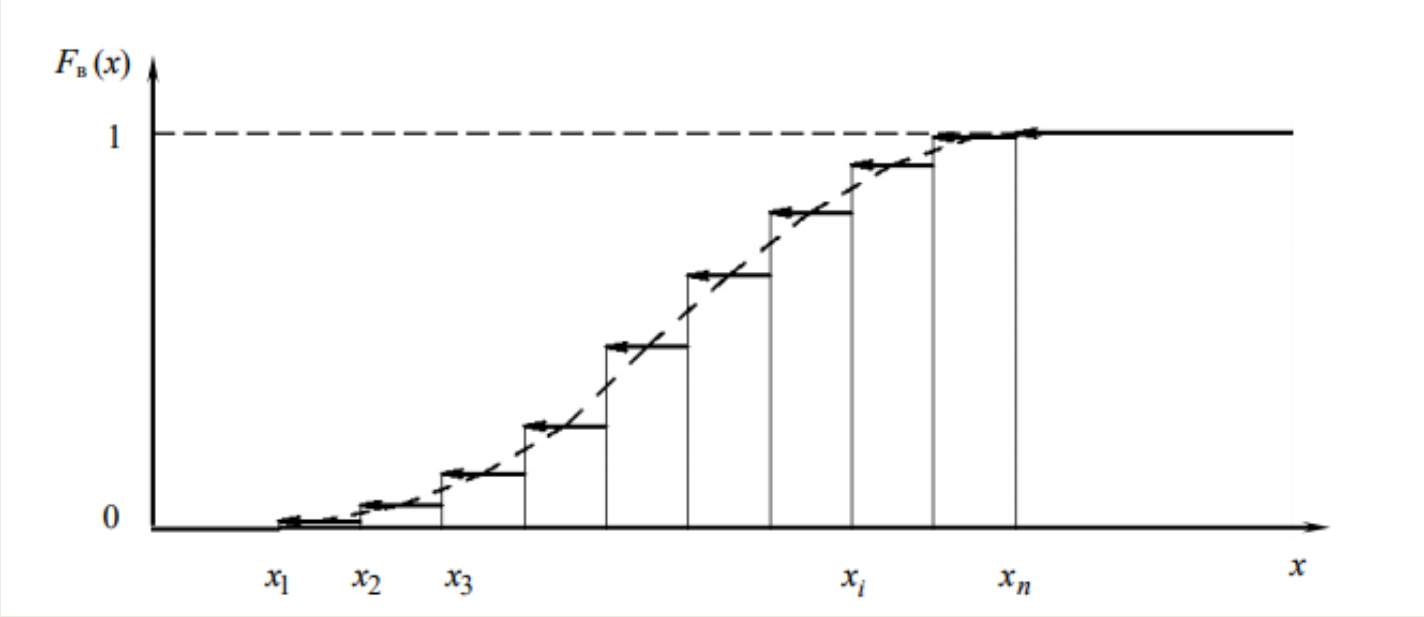

Рис. 1. Кумулята и эмпирическая функция распределения.

§ 2. Расчет выборочных характеристик статистического распределения

Рассмотрим выборку объема n со значениями x1, x2 , . . ., xn признака X. Для характеристики важнейших свойств статистического распределения используют средние показатели, называемые выборочными числовыми характеристиками. Если значения xi признака X не сгруппированы в вариационные ряды (табл. 2, 3, 4) и объем выборки n небольшой, то оценки для неизвестных математического ожидания a и дисперсии s^2 находят по формулам:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

§ 2. Расчет выборочных характеристик статистического распределения

Если результаты наблюдений сгруппированы в дискретный вариационный ряд (табл. 3), то те же оценки находят по формулам:

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

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2 .$$

По последней формуле вычисляют дисперсию в случае, если объем выборки $n \geq 50$.

§ 2. Расчет выборочных характеристик статистического распределения

Если же $n < 50$, то вычисляют исправленную дисперсию по формуле:

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

для простой выборки

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$$

для взвешенной выборки

§ 2. Расчет выборочных характеристик статистического распределения

Выборочное среднее квадратическое отклонение находят по формулам

$$S = \sqrt{S^2} \text{ или } \hat{S} = \sqrt{\hat{S}^2}$$

при различных объемах выборки.

§ 2. Расчет выборочных характеристик статистического распределения

Для анализа вариационных рядов вычисляют такие статистики, как моду и медиану. *Модой* *MoX* называют варианту, которая имеет наибольшую частоту. Например, для вариационного ряда

x_i	4	9	14	19
n_i	3	7	2	5

мода равна *MoX* = 9.

§ 2. Расчет выборочных характеристик статистического распределения

Медианой MeX называют варианту, которая делит вариационный ряд на равные по числу вариант части.

При нечетном объеме выборки $n = 2k+1$ медиана равна $MeX=x_{k+1}$. Например, для вариационного ряда

x_i	3	5	8	12	15
n_i	6	2	4	5	8

медиана равна $MeX = x_{13} = 12$.

§ 2. Расчет выборочных характеристик статистического распределения

При четном объеме выборки $n = 2k$ медиана находится по формуле:

$$M_e X = \frac{x_k + x_{k+1}}{2}.$$

Здесь x_k — варианта, которая находится слева от середины вариационного ряда, а x_{k+1} — справа от нее. Например, для следующего вариационного ряда:

x_i	2	5	7	10	12	14
n_i	3	4	8	2	3	6

медиана равна $MeX = 7$.

§ 2. Расчет выборочных характеристик статистического распределения

Для вычисления выборочной средней $\bar{x}$, выборочной дисперсии S^2, асимметрии *As* и эксцесса *Ex* при достаточно большом объеме выборки ($n > 30$) применяют метод произведений.

При этом вводят условные варианты *ui*, которые вычисляют по формуле:

$$u_i = \frac{x_i - C}{h},$$

где C = *MoX*, h — шаг (длина интервала).

§ 2. Расчет выборочных характеристик статистического распределения

Составляется расчетная табл. 6.

Таблица 6

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i u_i^3$	$n_i u_i^4$	контрольный столбец $n_i(u_i+1)^2$
строка сумм:	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$

Контроль вычислений ведут по формуле:

$$\sum n_i + 2\sum n_i u_i + \sum n_i u_i^2 = \sum n_i (u_i + 1)^2 .$$

§ 2. Расчет выборочных характеристик статистического распределения

Пользуясь табл. 6, вычисляют условные начальные моменты по формулам:

$$M_1^* = \frac{1}{n}\sum n_i u_i \,,$$

$$M_2^* = \frac{1}{n}\sum n_i u_i^2 \,,$$

$$M_3^* = \frac{1}{n}\sum n_i u_i^3 \,,$$

$$M_4^* = \frac{1}{n}\sum n_i u_i^4 \,.$$

§ 2. Расчет выборочных характеристик статистического распределения

Тогда выборочную среднюю находят по формуле:

$$\bar{x} = M_1^* h + C\,.$$

Выборочную дисперсию находят по формуле:

$$S^2 = (M_2^* - M_1^{*2})h^2\,.$$

Выборочное среднее квадратическое отклонение находят по формуле:

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

§ 2. Расчет выборочных характеристик статистического распределения

Асимметрию и эксцесс находят по формулам:

$$A_s = \frac{m_3}{S^3}, \qquad E_x = \frac{m_4}{S^4} - 3,$$

где

$$m_3 = (M_3^* - 3M_2^* M_1^* + 2M_1^{*3})h^3$$

— условный центральный момент третьего порядка

$$m_4 = (M_4^* - 4M_3^* M_1^* + 6M_2^* M_1^{*2} - 3M_1^{*4})h^4$$

— условный центральный момент четвертого порядка

§ 2. Расчет выборочных характеристик статистического распределения

Для характеристики колеблемости признака *X* используют относительный показатель — *коэффициент вариации V*, который для положительной случайной величины *X* вычисляют по формуле:

$$V = S / \bar{x}.$$

Коэффициент вариации подобного вида был предложен Пирсоном (1895) в несколько иной форме:

$$V' = 100S / \bar{x}.$$

§ 3. Интервальные (доверительные) оценки параметров распределения

Выборочные характеристики $\bar{x}$ и S^2 ,являются надежными количественными оценками генеральных характеристик *a* и σ^2 только при большом объеме выборки. При ограниченных объемах выборки возникает необходимость указать степень точности и надежности оценок генеральных характеристик.

При решении практических задач, связанных со статистическим анализом характеристик изучаемого признака *X* значения генеральной дисперсии и математического ожидания неизвестны.

§ 3. Интервальные (доверительные) оценки параметров распределения

Для оценки генеральной средней $M(X)=a$ и генерального среднеквадратического отклонения σ по выборочной средней x и выборочному средне квадратическому отклонению S находят *доверительные интервалы* по формулам:

$$\bar{x} - \frac{S}{\sqrt{n}} \cdot t_\gamma < a < \bar{x} + \frac{S}{\sqrt{n}} \cdot t_\gamma ,$$

Где t_γ находят из таблицы (см. приложение) по заданным n и γ (γ — уровень доверия или надежность, которая задается заранее).

§ 3. Интервальные (доверительные) оценки параметров распределения

Для генерального среднего квадратического отклонения доверительные интервалы находят по формулам:

$$S(1-q) < \sigma < S(1+q) \ (\text{при } q < 1),$$

или

$$0 < \sigma < S(1+q) \ (\text{при } (q > 1).$$

Величину q находят по таблице значений $q = (\gamma , n)$ (см. приложение) по заданным n и γ.

Спасибо за внимание!