

Лекция 2. Статистические методы обработки информации в нефтегазовом деле.

Составитель асс. каф. БНГС СамГТУ, магистр Никитин В.И.

1.5. Первичное представление эмпирических данных

Информация о работе любой отрасли производственной сферы (добыча нефти и газа, ремонт скважин, нефтехимическое производство, машиностроение и т.д.) 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. Изучение подобного рода массовых явлений выявление их статических и динамических закономерностей становится предметом математической статистики. Среди полезной информации о статистической совокупности особый интерес представляют статистические данные, которые можно записать в виде ряда $X_1, X_2, X_3, \dots, X_n$ 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 .

Пусть A — некоторое множество (например, множество всех жителей данного города). $B \subset A$ — 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 элементов множества A называется *объемом генеральной совокупности*, а число n — *объемом выборки*. При изучении

некоторого признака X (в нашем примере — роста) выборки производят испытания или наблюдения (измерение роста).

Пусть в результате независимых испытаний, проведенных в одинаковых условиях, получены числовые значения $x_1, x_2, x_3, \dots, x_n$, где n — объем выборки.

При обработке статистических данных строятся статистики. *Статистикой* называется функция

$$R^n \xrightarrow{f} R$$
$$x_1, x_2, x_3, \dots, x_n = x \in R^n, f(x) \in R$$

Таким образом, перед тем как подводить статистику, необходимо определиться о виде представления исходных данных. Рассмотрим наиболее удобные и часто встречающиеся варианты — это *выборка* (или) *массив данных*, представляющий собой неупорядоченные или упорядоченные в произвольном порядке величины. Далее рассмотрим методы построения и структуру вариационных рядов, которые можно построить на основании исходного массива данных.

1.6. Вариационные ряды

Последовательность n значений $x_1, x_2, \dots, x_n$ полученных в результате наблюдения (эксперимента) некоторого процесса, мы будем рассматривать как совокупность значений распределенных независимых случайных величин представляющих собой n экземпляров одного и того же признака X .

По этой выборке можно оценить основные числовые характеристики генеральной совокупности. Различные элементы выборки x_i , называются *вариантами*.

Число n_i , показывающее, сколько раз встречается варианта x_i в выборочной совокупности, называется ее частотой (эмпирической частотой). Частоты вариантов называются их *весами*. Отношение

$$h_i = \frac{n_i}{n} \quad (1.1)$$

относительной частотой (частостью) варианты x_i

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

Различают дискретные и непрерывные вариационные ряды. Дискретный вариационный ряд записывают в виде табл. 1.1. Отметим, что $\sum_{i=1}^k n_i = n$.

Таблица 1.1.

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

Общий вид дискретного вариационного ряда

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

$$R = x_{\max} - x_{\min} \quad (1.2)$$

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) $k \approx \sqrt{n}$;
- 2) $k \approx 1 + 3.221 \cdot \ln(n)$ - формула Стерджеса.

Данные формулы дают приблизительно одинаковый результат, знак « $\approx$ » используется, т.к. следует число интервалов k округлить до ближайшего целого значения. Длина Δ каждого частичного интервала определяется по формуле:

$$\Delta = \frac{R}{k}. \quad (1.3)$$

Величину Δ обычно округляют до некоторого значения d . Например, если результаты x_i признака X — целые числа, то Δ округляют до целого значения, если x_i содержат десятичные знаки, то Δ округляют до значения d , содержащего такое же число десятичных знаков. Затем подсчитывается частота

n_i , с которой попадают значения x_i признака X в i -й интервал. Значение x_i , которое попадает на границу интервала, относят к какому-либо определенному концу, например, к левому. За начало x_0 первого интервала рекомендуется брать величину $x_0 = x_{\min} - 0.5\Delta$. Конец $x_k = x_{\max} + 0.5\Delta$ последнего интервала находят по формуле. Также, для работы с интервальным вариационным рядом, в более простом виде, можно осуществить переход к дискретной записи, используя вместо интервалов $(x_{i-1}; x_i)$, их середины.

Сформированный интервальный вариационный ряд записывают в виде табл. 1.2.

Таблица 1.2

Интервалы- варианты $(x_{i-1}; x_i)$	$(x_0; x_1)$	$(x_1; x_2)$	...	$(x_{k-1}; x_k)$
Частоты n_i	n_1	n_2	...	n_k
Середины интервалов $(x_{i-1}; x_i)$	x_1^*	x_2^*	...	x_n^*
Относительные частоты $h_i = n_i/n$	h_1	h_2	...	h_n

Общий вид интервального вариационного ряда, с возможностью перехода к дискретному виду

Иногда данные для обработки поступают уже в интервальной группировке или представляется невозможным использовать одинаковые интервалы (например, в экономике).

Интервальный вариационный ряд изображают в виде гистограммы частот n_i или гистограммы относительных частот $h_i = n_i/n$.

Примечание. Если за границы интервального ряда принять $x_{\max}$ и $x_{\min}$, то значительной ошибки вычислений не последует. Поэтому в зависимости от задачи следует самостоятельно выбирать границы интервалов. Более того, используя формулы указанные выше, в ряде задач может возникнуть проблема

выхода за физические значения исследуемых параметров, что будет противоречить здравому смыслу задачи. К примеру, скорость проходки не может быть отрицательной.

ПРИМЕР 1.

Имеется выборка значений механической скорости бурения v_m коронкой И4ДП-59 в трещиноватых и абразивных породах X — XI категории по буримости:

0,67 0,78 0,74 0,69 0,7 0,82 0,76 0,69 0,7 0,68 0,76 0,76 0,82 0,71 0,74 0,81 0,75
0,85 0,8 0,79 0,85 0,85 0,76 0,81 0,72 0,74 0,75 0,73 0,8 0,78 0,76 0,76 0,71 0,77
0,74 0,78 0,72 0,75 0,78 0,78 0,8 0,71 0,74 0,73 0,77 0,75 0,84 0,72 0,78 0,77 0,81
0,75 0,79 0,73 0,75 0,71 0,77 0,72 0,79 0,76 0,83 0,72 0,74 0,83 0,71 0,84 0,75 0,77
0,77 0,73 0,73 0,73 0,74 0,76 0,71 0,75 0,75 0,75 0,82 0,77

Представить выборку в виде статистического дискретного вариационного ряда и в виде интервального вариационного ряда.

Решение

Для удобства следует отсортировать выборку по неубыванию и подсчитать её объём:

0,67 0,68 0,69 0,69 0,7 0,7 0,71 0,71 0,71 0,71 0,71 0,71 0,72 0,72 0,72 0,72 0,72
0,73 0,73 0,73 0,73 0,73 0,73 0,74 0,74 0,74 0,74 0,74 0,74 0,74 0,75 0,75 0,75 0,75
0,75 0,75 0,75 0,75 0,75 0,75 0,76 0,76 0,76 0,76 0,76 0,76 0,76 0,76 0,77 0,77 0,77
0,77 0,77 0,77 0,77 0,78 0,78 0,78 0,78 0,78 0,78 0,79 0,79 0,79 0,8 0,8 0,8 0,81 0,81
0,81 0,82 0,82 0,82 0,83 0,83 0,84 0,84 0,85 0,85 0,85

$$n = 80$$

Составим дискретный вариационный ряд, согласно табл.1.1

Варианты, x_i , м/ч,	0,67	0,68	0,69	0,7	0,71	0,72	0,73	0,74	0,75	0,76	0,77
Частоты, n_i	1	1	2	2	6	5	6	7	10	8	7
Варианты,	0,78	0,79	0,8	0,81	0,82	0,83	0,84	0,85			

$x_i, \text{м/ч},$								
Частоты, n_i	6	3	3	3	3	2	2	3

Всего $k=19$ различных вариантов. $\sum_{i=1}^k n_i = n = 80$

Составим интервальный статистический ряд механической скорости бурения, для этого определим размах варьирования по формуле (1.2) и длину интервала по формуле (1.3).

$$R = x_{\max} - x_{\min} = 0.85 - 0.67 = 0.18$$

$$\Delta = \frac{R}{K} = \frac{0.18}{5} = 0.036 \approx 0.04$$

Δ округлили до сотых, так как все элементы выборки представлены в этой же разрядности.

Запишем интервальный вариационный ряд согласно таблице 1.2. За начало первого интервала возьмем $x_0 = x_{\min} - 0.5\Delta = 0.67 - 0.02 = 0.65$, за конец последнего возьмем $x_{\max}$ так как если воспользуемся формулой $x_k = x_{\max} + 0.5\Delta = 0.87$, то последний интервал будет самым широким. Это не приведет к вычислительной ошибке, но может быть не совсем достоверно с физической точки зрения.

(Интервалы- варианты) Скорость проходки, м/ч	0.65 – 0.69	0.69 – 0.73	0.73 – 0.77	0.77 – 0.81	0.81 – 0.85
Частоты n_i	4	19	32	15	10
Середины интервалов x_i^*	0.67	0.71	0.75	0.79	0.83
Относительные частоты $h_i = n_i/n$	0.05	0.2375	0.4	0.1875	0.125

