ПЕРВОЕ ВЫСШЕЕ ТЕХНИЧЕСКОЕ УЧЕБНОЕ ЗАВЕДЕНИЕ РОССИИ
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования

САНКТ-ПЕТЕРБУРГСКИЙ ГОРНЫЙ УНИВЕРСИТЕТ

Кафедра маркшейдерского дела

МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В МАРКШЕЙДЕРСКОМ ОБЕСПЕЧЕНИИ

Методические указания к практическим занятиям
для студентов специальности 21.05.04

САНКТ-ПЕТЕРБУРГ
2018

МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В МАРКШЕЙДЕРСКОМ ОБЕСПЕЧЕНИИ

Методические указания к практическим занятиям
для студентов специальности 21.05.04

САНКТ-ПЕТЕРБУРГ
2018

УДК 519.2 (076.5)

МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В МАРКШЕЙДЕРСКОМ ОБЕСПЕЧЕНИИ. Статистическая оценка распределения показателя и его параметров: Методические указания к лабораторным работам / Национальный минерально-сырьевой университет «Горный» Сост**.** Е.А Нестеренко СПб, 2018.18 с.

Методические указания включают теоретические основы, практические рекомендации, задания и примеры обработки статистических данных, определение числовых характеристик статистических значений показателя. Рассмотрена возможности обработки данных в программе Microsoft Excel, приведены примеры расчетов.

Предназначены для студентов специальности 130402 «Маркшейдерское дело».

Табл. 3. Библиогр.: 3 назв.

Научный редактор проф. В Н Гусев,

Лабораторная работа №1.
ХАРАКТЕРИСТИКИ СЛУЧАЙНОГО РАСПРЕДЕЛЕНИЯ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙИ ГОРНО-ГЕОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Случайные величины и выборки

Случайной величиной обычно называют такую, которая в результате опыта может принимать то или иное значение, не известное нам заранее в пределах требуемой точности ее определения. Случайные величины бывают дискретными и непрерывными.

Численные значения случайных величин получают в процессе *опыта*, в качестве которого могут выступать маркшейдерские съемки, разведка и опробование залежи, опробование полезного ископаемого, определение различного рода показателей производственной деятельности горнодобывающего предприятия, изучение распределения горно-геометрических и квалиметрических показателей недр, геомеханические показатели процесса сдвижения и деформаций горных пород и др.

Определенное количество значений случайной величины, характеризующее изучаемый процесс или объект, называют *совокупностью*, а число этих значений в совокупности – ее *объемом*. Весь набор множества однородных показателей, характеризующих исследуемый объект или процесс, называют *генеральной совокупностью*.

При больших объемах генеральной совокупности объекты или явления изучают путем извлечения из нее части показателей, которую называют *выборочной совокупностью* или *выборкой.*

Пусть выборка представлена объемом, исчисляемым тысячами значений определяемого показателя. В этом случае для удобства их использования, анализа и получения практических выводов пользуются *вариационным* (*статистическим, сгруппированным*) *рядом* распределения значений показателя.

В качестве примера рассмотрим выборку результатов определения содержания свинца в полиметаллических рудах (в %), объёмом 100 (таблица 1):

Таблица 1

0,12	1,27	0,17	1,68	0,28	3,36	0,76	0,81	0,94
0,21	1,12	0,30	1,78	0,38	4,76	0,17	0,89	0,44
0,41	1,82	0,55	2,25	0,60	4,20	0,64	0,67	0,81
0,76	1,75	0,24	2,10	0,30	5,08	0,35	0,40	0,30
0,79	1,66	0,94	2,71	0,20	4,20	0,31	0,47	0,55
0,52	1,69	0,58	2,60	0,61	0,22	0,35	0,47	0,26
1,44	0,18	2,41	0,12	0,84	0,70	0,51	0,37	1,52
0,31	2,45	0,21	0,74	0,67	0,80	0,73	1,39	0,79
2,74	0,41	0,77	0,88	0,14	0,26	1,44	0,30	2,65
0,37	0,35	0,31	0,26	0,51	1,30	0,44	3,54	0,54
1,99	0,58	0,67	0,73	0,77	1,36	0,83	3,47	0,86
0,88								

Построение вариационного ряда

Содержание полезного компонента в рудном теле, так же как, например, плотность угля, может принимать любые значения от минимального до максимального для изучаемой залежи, поэтому данные случайные величины являются непрерывными [1].

Для *непрерывной случайной* величины составляют *интервальные вариационные ряды*. Для этого весь диапазон значений случайной величины разбивают на ряд интервалов, равных по размеру. Частотой каждого интервала будет являться число значений показателя, попадающих в пределы данного интервала. Среднее значение каждого интервала (середина интервала) является как бы дискретным значением показателя. Вариационный ряд в этом случае записывают в следующем виде:

Интервал	$X_{\min}$-X_1	X_1-X_2	X_{k-1}-X_k
Середина интер-вала	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_k$
Частота	n_1	n_2	n_k
Частость	p_1	p_2	p_k

Здесь $(X_{\min} \div X_1) = (X_1 \div X_2) = \cdots = (X_{k-1}\text{-}X_k) = h$, где h – величина интервала; $\bar{x}_1 = (X_1 + X_{\min})/2$, $\bar{x}_2 = (X_2 + X_1)/2$, $\bar{x}_k = (X_k + X_{k-1})/2$

Величина интервала подсчитывается по формуле Стерджеса

$$h = (x_{\max} - x_{\min})/(1 + 3{,}2 \lg N). \quad (1)$$

Расчёты для примера:

Соответственно, если в выборке, взятой для примера, $x_{\max}$=5,08 и $x_{\min}$=0,12, то по формуле 1:

$h = (5{,}08 - 0{,}12)/(1 + 3{,}2 \lg 100) = 0{,}67 \approx 0{,}7$.

На основании величины h составляются интервалы:

$x_{\min} + h = X_1$

$X_1 + h = X_2$

……

$X_n + h = x_{max}$, где $x_{\min}$ и X_1 – границы первого интервала, X_1 и X_2 – границы второго интервала и т.д.

В выборке значения показателя x_1, x_2, …, x_k называются *вариантами*, а объем каждого варианта n_1, n_2, …, n_k – его *частотой*. В случае с непрерывными случайными величинами для определения частоты необходимо посчитать, сколько вариантов попадает в каждый интервал. Их число и будет частотой интервала.

Для сопоставимости вариационных рядов с различными объемами выборок используют значения *относительных частот*, или *частостей*, вычисляемых по формуле

$p_i = n_i/N$

В рассматриваемом примере вариационный ряд будет выглядеть следующим образом (табл.2):

Таблица 2

Интервал	0,12-0,82	0,82-1,52	1,52-2,22	2,22-2,92	2,92-3,62	3,62-4,32	4,32-5,02	5,02-5,72
Середина интервала	0,47	1,17	1,87	2,57	3,27	3,97	4,67	5,37
Частота	62	15	9	7	3	2	1	1
Частость	0,62	0,15	0,09	0,07	0,03	0,02	0,01	0,01

Для упрощения процедуры расчётов использовалась программа Microsoft Excel (таблица 3).

Для упрощения вычислений при обработке вариационных рядов удобно пользоваться условными значениями вариантов:

$$a_i = (\bar{x}_i - \bar{x}_0)/h,$$

где $\bar{x}_i$ – среднее значение показателя в i-м интервале (значение середины интервала); $\bar{x}_0$ – среднее значение интервала с наибольшей частотой (нулевого интервала).

При резко асимметричном распределении нулевой интервал выбирается независимо от частоты ближе к середине вариационного ряда. Нетрудно убедиться, что при постоянном значении интервала вариационного ряда условные значения вариантов есть целые числа, возрастающие от нуля (интервал со средним значением $\bar{x}_0$) на единицу (в сторону уменьшения показателя – отрицательные, в сторону увеличения –

положительные). Абсолютная величина $(\bar{x}_i - \bar{x}_0)$ всегда получается кратной h.

В примере условные значения вариантов были рассчитаны в столбце 5 таблицы 3:

$$a_1 = (\bar{x}_1 - \bar{x}_0)/h = (0{,}47 - 3{,}27)/0{,}7 = -4,$$

$$a_2 = (\bar{x}_2 - \bar{x}_0)/h = (1{,}17 - 3{,}27)/0{,}7 = -3 \text{ и т.д.,}$$

при этом за величину $\bar{x}_0$ была взято не среднее значение интервала с наибольшей частотой, а среднее значение интервала, взятого примерно посередине вариационного ряда.

Определение числовых характеристик статистических значений показателя

Для обобщенной количественной оценки распределения значений изучаемого показателя вариационного ряда используют характеристики, которые определяют положение центра распределения и размах рассеивания отдельных значений относительно центра.

Положение центра распределения определяют математическим ожиданием (средним), модой и медианой.

Для *оценки рассеивания* (вариации, изменчивости) отдельных значений относительно центра пользуются дисперсией, средним квадратическим отклонением (стандартом), коэффициентом вариации, асимметрией, эксцессом.

Положение центра распределения

Среднее выборочное значение показателя $\bar{x}$ рассчитывается по формулам

$$\bar{x} = \frac{\sum n_i \bar{x}_i}{N} = \sum p_i \bar{x}_i$$

Для расчета величин $\sum n_i \overline{x_i}$ и $\sum p_i \overline{x_i}$ в представленном примере в таблице 3 были добавлены соответствующие столбцы 9 и 10. Отсюда

$$\bar{x} = \frac{\sum n_i \overline{x_i}}{N} = \frac{109{,}3}{100} = 1{,}093$$

Модой Мо называют значение показателя, которое имеет наибольшую вероятность (частоту или частость). Интервал вариационного ряда, имеющий наибольшую частоту (частость), называют модальным. Рассчитывается мода по формуле

$$Mo = x_{нMo} + h\frac{p_2 - p_1}{2p_2 - p_1 - p_3} = x_{нMo} + h\frac{n_2 - n_1}{2n_2 - n_1 - n_3}, \qquad (2)$$

где $p_1(n_1)$, $p_2(n_2)$ и $p_3(n_3)$ – частости (частоты) предыдущего, модального и последующего интервалов вариационного ряда; $x_{нМо}$ – нижняя граница модального интервала.

В примере модальным является первый интервал. Пример расчёта моды:

$$Mo = x_{нMo} + h\frac{p_2 - p_1}{2p_2 - p_1 - p_3} = 0{,}12 + 0{,}7\frac{0{,}62 - 0}{2 \cdot 0{,}62 - 0 - 0{,}15} = 0{,}518$$

Медианой Ме называют значение показателя в выборке, которое делит ряд распределения по частоте или частости на две равные части. Медианным интервалом является интервал, в котором находится половина накопленных частот или частостей. В интервальном вариационном ряду медиану определяют, исходя из пропорциональности значений показателя частоты или частости (изменение по линейному закону). На этом основании получена формула

$$Me = x_{нMe} + h\frac{0{,}5 - \sum p_{нMe}}{p_{Me}} = x_{нMe} + h\frac{0{,}5N - \sum n_{нMe}}{n_{Me}}, \qquad (3)$$

где $x_{\text{нMe}}$ – нижняя граница медианного интервала; $\sum n_{\text{нMe}}, \sum p_{\text{нMe}}$ – накопленная частота или частость до начала медианного интервала; $n_{\text{Me}}, p_{\text{Me}}$ – частота или частость медианного интервала.

Для упрощения вычислений в таблице 3 были созданы столбцы 7 и 8. Пример определения накопленной частоты (столбец 7 табл. 3):

$$\sum n_5 = n_1 + n_2 + n_3 + n_4 + n_5 = 62 + 15 + 9 + 7 + 3 = 96$$

Накопленная частость была рассчитана в восьмом столбце таблицы 3 аналогичным образом.

В примере медианным является первый интервал. Отсюда медиана по формуле 3:

$$Me = x_{нMe} + h\frac{0{,}5N - \sum n_{нMe}}{n_{Me}} = 0{,}12 + 0{,}7\frac{0{,}5 \cdot 100 - 0}{62} = 0{,}684$$

Моду и медиану используют для приближенной оценки симметричности распределения. При симметричном распределении $\bar{x} \approx \text{Mo} \approx \text{Me}$. При левосторонней асимметрии $\bar{x} > \text{Mo}$, $\bar{x} > \text{Me}$, а при правосторонней $\bar{x} < \text{Mo}$, $\bar{x} < \text{Me}$. Проанализировав полученные величины, делаем вывод о левосторонней асимметрии распределения.

Характеристика оценки рассеивания

Выборочная дисперсия вычисляется по формулам

$$\sigma^2 = \overline{x}^2 - (\overline{x})^2;$$

$$\sigma^2 = h^2\left[\alpha_2' - (\alpha_1')^2\right], \qquad (4)$$

где $\overline{x}^2$ – средний квадрат показателя в выборке (начальный момент второго порядка); $\overline{x}$ – среднее выборочное значение показателя (начальный момент первого порядка), вычисляемый по тем же формулам; h^2 – интервал вариационного ряда; α_2', α_1' – условные моменты соответственно второго и первого порядка, определяемые по формуле:

$$\alpha_k' = \frac{\sum_1^N n_i a_i^k}{N} = \sum_1^N p_i a_i^k \qquad (5)$$

Отсюда $$\alpha_1' = \frac{\sum_1^N n_i a_i}{N} = \sum_1^N p_i a_i ,$$

$$\alpha_2' = \frac{\sum_1^N n_i a_i^2}{N} = \sum_1^N p_i a_i^2 .$$

Расчёты для примера:

величины $\sum_1^N n_i a_i$ и $\sum_1^N n_i a_i^2$ рассчитаны в столбцах 11 и 12 таблицы 3: $\sum_1^N n_i a_i = -311, \sum_1^N n_i a_i^2 = 1185$.

Отсюда $\alpha'_1 = \frac{\sum_1^N n_i a_i}{N} = -3{,}11$; $\alpha'_2 = \frac{\sum_1^N n_i a_i^2}{N} = 11{,}85.$

По формуле 4 рассчитываем дисперсию:

$\sigma^2 = h^2\left[\alpha'_2 - (\alpha'_1)^2\right] = 0{,}7^2 \cdot (11{,}85 - (-3{,}11)^2) = 1{,}07$

Среднее квадратическое отклонение (стандарт) равно корню квадратному из дисперсии $\sigma = \sqrt{\sigma^2}$ и выражается в тех же единицах, что и изучаемый показатель.

Для предложенного примера:

$\sigma = \sqrt{\sigma^2} = 1{,}033$

Коэффициент вариации ν характеризует относительную величину рассеивания значений показателя. Этот коэффициент позволяет сопоставлять размах рассеивания показателей, разнородных как по абсолютной величине, так и по единицам. Коэффициент вариации обычно вычисляют в процентах по формуле

$$\nu = \frac{\sigma}{\bar{x}} 100. \qquad (6)$$

Для предложенного примера:

$\nu = \frac{\sigma}{\bar{x}} 100 = \frac{1{,}033}{1{,}093} \cdot 100 = 94{,}5\%$

Асимметрия. Рассеивание отдельных значений показателя относительно среднего может быть симметричным и асимметричным.

Для симметричных распределений центральные моменты нечетных порядков равны нулю. Несоблюдение этого условия свидетельствует об асимметричности (скошенности) распределения. В этом случае возрастает влияние крайних значений и центральные моменты нечетной степени (при $k > 1$)

не будут равны нулю. Для суждения об этом достаточно получить центральный момент третьего порядка.

Для удобства сопоставления разнородных показателей обычно используют безразмерный коэффициент асимметрии, вычисляемый по формуле

$$A = \alpha_3/\sigma^3, \qquad (7)$$

где α_3 – центральный момент третьего порядка, определяемый по формуле 8

$$\alpha_3 = h^3(\alpha'_3 - 3\alpha'_2\,\alpha'_1 + 2(\alpha'_1)^3), \qquad (8)$$

где условные эмпирические моменты первого, второго и третьего порядка определяются по формуле 5.

Положительный знак величины A ($A > 0$) указывает на левостороннюю асимметрию, отрицательный ($A < 0$) – на правостороннюю, равенство нулю ($A = 0$) – на отсутствие асимметрии (распределение случайных величин симметричное).

При обработке даже симметричных эмпирических распределений в силу влияния случайных факторов почти всегда $A \neq 0$, поэтому асимметрию считают *существенной*, если по абсолютному значению она превышает значение $3\sigma_A$, где σ_A – среднее квадратическое отклонение величины A, определяемое по формуле

$$\sigma_A = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}} \approx \sqrt{\frac{6}{N}}\,. \qquad (9)$$

Асимметрию считают *несущественной*, если $A < 3\sigma_A$.

Решение для примера (см.табл 3, столбцы 11-13):

$$\alpha'_1 = \frac{\sum_1^N n_i a_i}{N} = -3{,}11;$$

$$\alpha'_2 = \frac{\sum_1^N n_i a_i^2}{N} = 11{,}85\,;$$

$$\alpha'_3 = \frac{\sum_1^N n_i a_i^3}{N} = -44{,}15\,.$$

Отсюда получаем

$$\alpha_3 = h^3(\alpha'_3 - 3\alpha'_2\,\alpha'_1 + 2(\alpha'_1)^3 =$$

$$= 0{,}7^3(-44{,}15 + 3 \cdot 11{,}85 \cdot 3{,}11 + 2 \cdot 3{,}11^2) = 1{,}94$$

$$A = \frac{\alpha_3}{\sigma^3} = \frac{1{,}94}{1{,}03} = 1{,}88$$

$A>0$ – асимметрия левосторонняя.

$\sigma_A = \sqrt{\dfrac{6}{N}} = \sqrt{\dfrac{6}{100}} = 0{,}245$, $A > 3\sigma_A$ - асимметрия существенна.

Эксцесс. Для оценки «крутизны» («вершинности») эмпирической кривой распределения по сравнению с кривой нормального распределения используют показатель эксцесса.

Коэффициент эксцесса вычисляют по формуле 10

$$Э = \frac{\alpha_4}{\sigma^4} - 3, \tag{10}$$

где α_4 – центральный момент четвертого порядка, определяемый по формуле 11

$$\alpha_4 = h^4(\alpha'_4 - 4\alpha'_3\,\alpha'_1 + 6\alpha'_2(\alpha'_1)^2 - 3(\alpha'_1)^4), \tag{11}$$

где условные эмпирические моменты первого, второго, третьего и четвёртого порядка определяются по формуле 5.

Для нормального распределения $Э = 0$. Невыполнение равенства (10) свидетельствует об отклонении вершины фактического распределения от кривой нормального распределения. При $Э > 0$ вершина эмпирической кривой расположена выше вершины нормальной кривой, а при $Э < 0$ – ниже.

Эксцесс считается существенным, если он по абсолютному значению превышает $3\sigma_э$, где $\sigma_э$ – среднее квадратическое отклонение эксцесса, вычисляемое по формуле

$$\sigma_э = \sqrt{\frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}} \approx \sqrt{\frac{24}{N}} \approx 2\sigma_A. \qquad (12)$$

Решение для примера (см.табл. 3, столбцы 11-14):

$$\alpha'_1 = \frac{\sum_1^N n_i a_i}{N} = -3{,}11;$$

$$\alpha'_2 = \frac{\sum_1^N n_i a_i^2}{N} = 11{,}85;$$

$$\alpha'_3 = \frac{\sum_1^N n_i a_i^3}{N} = -44{,}15.$$

$$\alpha'_4 = \frac{\sum_1^N n_i a_i^4}{N} = 173{,}37.$$

Отсюда получаем

$$\alpha_4 = h^4(\alpha'_4 - 4\alpha'_3\alpha'_1 + 6\alpha'_2(\alpha'_1)^2 - 3(\alpha'_1)^4) = 7{,}49$$

$Э = \dfrac{\alpha_4}{\sigma^4} - 3 = 3{,}57$, $Э>0$ – вершина эмпирической кривой лежит выше вершины нормальной кривой.

$$\sigma_э = \sqrt{\frac{24}{N}} = 0{,}49$$

$Э > 3\sigma_э$ – эксцесс существенный.

<u>Задание:</u>

По результатам лабораторного определения объёмного веса известняков месторождения республики Якутия (Саха) (т/м^3) выполнить следующее:

1) Построить вариационный ряд;
2) Определить среднее, среднеквадратичнеское значение показателя, коэффициент вариации, моду, медиану, дисперсию, эксцесс, асимметрию;
3) Построить эмпирическую и теоретическую кривые распределения показателя.

Таблица 3

Характеристики эмпирического распределения

1	2	3	4	5	6	7	8	9	10	11	12	13	14
интервалы		$\bar{x}_i$	n_i	p_i	a_i	$\sum n_i$	$\sum p_i$	$n_i \bar{x}_i$	$p_i \bar{x}_i$	n_i	$n_i a_i^2$	$n_i a_i^3$	$n_i a_i^4$
0,12	0,82	0,47	62	0,62	-4	62	0,62	29,14	0,2914	-248	992	-3968	15872
0,82	1,52	1,17	15	0,15	-3	77	0,77	17,55	0,1755	-45	135	-405	1215
1,52	2,22	1,87	9	0,09	-2	86	0,86	16,83	0,1683	-18	36	-72	144
2,22	2,92	2,57	7	0,07	-1	93	0,93	17,99	0,1799	-7	7	-7	7
2,92	3,62	3,27	3	0,03	0	96	0,96	9,81	0,0981	0	0	0	0
3,62	4,32	3,97	2	0,02	1	98	0,98	7,94	0,0794	2	2	2	2
4,32	5,02	4,67	1	0,01	2	99	0,99	4,67	0,0467	2	4	8	16
5,02	5,72	5,37	1	0,01	3	100	1	5,37	0,0537	3	9	27	81
Сумма		23,36	100		28,7			109,3	1,093	-311	1185	-4415	17337

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Гусев В.Н.* Математическая обработка маркшейдерской информации статистическими методами: Учебное пособие/ В.Н.Гусев,А.Н.Шеремет; СПб.гос.горн.ин-т. СПб.,2004;

2. *Гудков В.М.* Математическая обработка маркшейдерско-геодезических измерений: Учебник для вузов / Гудков В.М., Хлебников А.В.; М., Недра, 1990;

3. *Павловский З.* Введение в математическую статистику. М., Статистика, 1966.

СОДЕРЖАНИЕ

МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В МАРКШЕЙДЕРСКОМ ОБЕСПЕЧЕНИИ. Статистическая оценка распределения показателя и его параметров

Лабораторные работы для студентов очного обучения специальности 060800

Составители: *В.Н.Гусев, Е.А.Нестеренко, Р.А.Такранов*

Редактор и корректор И.В.Неверова

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 16.04.04. Формат 60 х 84/16.
Бум. для копировальной техники. Отпечатано на ризографе.
Усл.печ.л. 1,51. Усл.кр.-отт. 1,51. Уч.-изд.л. 1,47.
Тираж 100 экз. Заказ 176. С 58.

Санкт-Петербургский государственный горный институт им.Г.В.Плеханова
РИЦ Санкт-Петербургского государственного горного института
Адрес института и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2