МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

КАФЕДРА «МАРКШЕЙДЕРСКОЕ ДЕЛО ИМ. Д. Н. ОГЛОБЛИНА»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по выполнению лабораторных и самостоятельных работ по дисциплинам

«Математическая статистика в горном деле»

уровень профессионального высшего образования «специалист»
специальность 21.05.04 «Горное дело»
специализация «Маркшейдерское дело»

РАССМОТРЕНО
на заседании кафедры
маркшейдерского дела им. Д. Н. Оглоблина
Протокол № 7 от 13 января 2020 г.

УТВЕРЖДЕНО
на заседании Учебно-издательского
совета ДОННТУ
Протокол № от

Донецк
2020

УДК 528.3:622.1(076)
ББК 26.12:33.12я73
М54

Рецензент:
Хохлов Борис Валентинович - кандидат технических наук, старший научный сотрудник Республиканского академического научно-исследовательского и проектно-конструкторского института геомеханники и маркшейдерского дела.

Составители:
Филатова Ирина Викторовна - кандидат технических наук, доцент кафедры маркшейдерского дела им. Д. Н. Оглоблина ГОУВПО «ДОННТУ»;
Канавец Александра Андреевна - ассистент кафедры маркшейдерского дела им. Д. Н. Оглоблина ГОУВПО «ДОННТУ».

Методические указания по выполнению лабораторных и самостоятельных работ по дисциплинам «Математическая статистика в горном деле» [Электронный ресурс]: уровень проф. высш. образования «специалист» специальность 21.05.04 «Горное дело» специализация «Маркшейдерское дело» / ГОУВПО «ДОННТУ», Каф. маркшейдерского дела им. Д. Н. Оглоблина; сост.: И. В. Филатова, А. А. Канавец – Электрон. дан. (1 файл). - Донецк: ДОННТУ, 2020. – Систем. требования: Acrobat Reader.

Методические указания разработаны по дисциплине «Математическая статистика в горном деле», составлены на основе рабочей программы, охватывают основные темы учебной дисциплины.

Методические указания рекомендованы к изданию методической комиссией специальности 21.05.04 "Горное дело" специализации «Маркшейдерское дело» (протокол №. 7) и предназначены для подготовки специалистов специальности 21.05.04 «Горное дело» специализации "Маркшейдерское дело".

УДК 528.3:622.1(076)
ББК 26.12:33.12я73
М54

ВВЕДЕНИЕ

Математическая статистика разрабатывает методы сбора, систематизации и обработки статистических данных. Применение этих методов при научных исследованиях или решении производственных задач позволяет выявить закономерности, присущие статистическим данным, и на этой основе получить научные или практические выводы и рекомендации.

Знание методов математической статистики имеет важное значение для успешной деятельности горного инженера-маркшейдера, которому часто приходится иметь дело с величинами, подчиненными вероятностно-статистическим закономерностям. К таким величинам относятся ошибки измерений, различные структурные и качественные показатели месторождения, проявления сдвижения горных пород и их влияния на состояние горных выработок и объектов поверхности. Обоснованное использование этих величин в практике горного дела возможно только после изучения присущих им закономерностей, что требует применения методов математической статистики.

В настоящих методических указаниях рассмотрена методика решения трех статистических задач, которые наиболее часто встречаются в практической деятельности маркшейдера:

1) исследование распределения признака путем нахождения его числовых характеристик или установления закона распределения;

2) нахождение и исследование корреляционной зависимости между двумя признаками;

3) проверка статистических гипотез о математических ожиданиях, дисперсиях, законе распределения.

В приведенных трех практических работах, соответствующих перечисленным задачам, сначала дается детальное решение реального примера с указанием всех необходимых формул и подробными вычислениями, а затем приводятся варианты задач для самостоятельного решения. Объем

практических работ рассчитан на 14 часов аудиторных занятий и 7 часов самостоятельной работы.

ЛАБОРАТОРНАЯ РАБОТА №1
«НАХОЖДЕНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ПРИЗНАКА»

1. Основные положения.

Значения некоторого количественного признака объекта, полученные в результате наблюдений или экспериментов, называются статистическими данными. Они образуют выборочную совокупность (выборку). Нд основе выборки проводится статистическое исследование признака, целью которого является выявление присущих данному признаку закономерностей.

С математической точки зрения понятия признака и случайной величины тождественны. Поэтому проведение статистического исследования признака в сущности означает исследование по данным выборки его закона распределения.

В зависимости от поставленной задачи и объема выборки исследование распределения признака может быть проведено с разной степенью детальности:

1) приближенное исследование, при котором устанавливаются лишь существенные особенности распределения признака; для этого строится интервальный вариационный ряд (ИВР) и находятся числовые характеристики (такому исследованию посвящена настоящая практическая работа);

2) полное исследование, в результате которого устанавливается закон распределения признака (будет рассмотрено в практической работе 3).

Числовыми характеристиками называются числовые значения, характеризующие наиболее существенные особенности распределения признака.

К ним относятся математическое ожидание, мода, медиана, дисперсия, среднее квадратическое отклонение, коэффициент вариации, асимметрия и эксцесс.

Приближенные значения числовых характеристик, найденные по данным выборки, принято называть статистическими характеристиками, или статистическими аналогиями числовых характеристик. В данной работе рассматривается методика их нахождения.

2. Цель работы.

Целью работы является:

1) Освоение методики составления ИВР по данным выборки.

2) Освоение методики вычисления статистических характеристик признака по данным ИВР.

3) Уяснение реального смысла ИВР и найденных характеристик и их значения для решения практических задач.

3. Методика выполнения работы.

Для выполнения практической работы каждому студенту дается индивидуальный вариант с выборочными данными некоторого признака. По этим данным необходимо провести статистическое исследование распределения признака, для чего выполнить следующее:

1) составить ИВР признака;

2) построить гистограмму и статистическую функцию распределения признака (кумулятивную кривую);

3) вычислить статистические характеристики признака: среднее (математическое ожидание), моду, медиану, дисперсию, среднее квадратическое отклонение, коэффициент вариации, ассиметрию, эксцесс;

4) сравнить полученные результаты с результатами смежного варианта и проанализировать, как влияют расхолухения между ними на решение реальных производственных задач.

Методика выполнения работы подробно рассмотрена на примере, который приводится ниже.

Пример:

На полиметаллическом руднике в забоях горных выработок было взято 87 проб, равномерно расположенных на площади залежи. В результате химических анализов этих проб на цинк были получены данные, приведенные в таблице 1.

Таблица 1.

Номер пробы	Содержание цинка, %	Номер пробы	Содержание цинка, %	Номер пробы	Содержание цинка, %
1	2	3	4	5	6
1	2,62	30	1,79	59	1,28
2	1,98	31	0,42	60	5,18
3	3,15	32	1,91	61	0,95
4	0,88	33	2,33	62	1,72
5	1,17	34	1,83	63	2,81
6	0,11	35	3,44	64	1,84
7	2,13	36	0,25	65	3,98
8	1,35	37	1,64	66	1,15
9	4,19	38	2,51	67	0,55
10	3,27	39	1,09	68	1,32
11	1,87	40	4,59	69	2,34
12	5,31	41	0,18	70	1,56
13	0,93	42	1,27	71	6,97
14	1,62	43	2,45	72	0,61
15	1,02	44	1,18	73	1,44
16	2,29	45	1,35	74	2,23
17	1,43	46	0,46	75	1,27
18	4,46	47	1,48	76	0,86
19	0,74	48	3,61	77	1,66
20	1,69	49	1,31	78	4,87
21	1,24	50	2,74	79	1,93
22	2,06	51	0,31	80	2,11
23	1,32	52	1,39	81	0,92
24	0,61	53	6,43	82	1,78
25	1,06	54	1,57	83	3,49
26	2,17	55	2,96	84	1,81

Продолжение таблицы 1.

1	2	3	4	5	6
27	0,56	56	0,24	85	5,68
28	3,08	57	1,86	86	1,62
29	1,72	58	4,71	87	3,73

По этим данным необходимо выполнить статистическое исследование распределения цинка в соответствии с указанными выше пп.1-4.

Решение

1. Строим ИВР признака. Для этого прежде всего рассчитываем ширину интервала по формуле:

$$h = \frac{X_{\max} - X_{\min}}{1 + 3{,}2 \lg n}, \qquad (1)$$

где X_{max} и X_{min} – максимальное и минимальное значения признака (содержания цинка) в выборке;

n – объем выборки.

Подставляя фактические значения в формулу (1), получим:

$$h = \frac{6{,}97 - 0{,}11}{1 + 3{,}2 \lg 87} = 0{,}95\%$$

Тогда количество интервалов составит:

$$l = \frac{X_{\max} - X_{\min}}{h} = \frac{6{,}97 - 0{,}11}{0{,}95} = 7{,}2$$

Принимаем количество интервалов $l = 7$ и уточняем ширину интервала:

$$h = \frac{X_{\max} - X_{\min}}{l} = \frac{6{,}97 - 0{,}11}{7} = 0{,}98\%$$

Теперь можно принять X_{min} в качестве левой границы первого интервала, и последовательно добавляя h=0,98, рассчитать границы интервалов. При этом правой границей последнего интервала должно получиться значение X_{max}.

Однако для большей наглядности ИВР и удобства последующих расчетов целесообразно сделать некоторые округления: ширину интервала принять равной 1%, а в качестве левой границы первого интервала принять 0 вместо X_{min}=0,11%. Тогда правая граница последнего интервала окажется равной 7% вместо 6,97%. Полученные при этом границы всех интервалов приведены в графе 1 таблицы 2.

Таблица 2.

Границы интервалов	Середина интервала	Подсчет частоты m_i	Частота m_i	Частость	Накопленная частость
1	2	3	4	5	6
0-1	0,5		17	0,20	0,20
1-2	1,5		38	0,44	0,64
2-3	2,5		14	0,16	0,80
3-4	3,5		8	0,09	0,89
4-5	4,5		5	0,06	0,95
5-6	5,5		3	0,03	0,98
6-7	6,5		2	0,02	1,00
			$\sum m_i = 87$	$\sum W_i = 1$	

В графу 2 записываем средние значения каждого интервала. В графе 3 производим подсчет частоты m_i, т,е. количества проб, попавших в данный интервалу, и полученные значения заносим в графу 4, Контролем подсчета служит соотношение $\sum_{1}^{l} m_i = n$.

В графу 5 записываются частоты (относительные частоты или статистические вероятности), вычисляемые по формуле:

$$W_i = \frac{m_i}{n}, \qquad (2)$$

Частость показывает, какую долю среди всех значений признака составляют значения, попавшие в данный интервал. Сумма частостей должна равняться единице.

Наконец, в графу 6 заносятся накопленные частости, вычисляемые по формуле:

$$W_i' = W_{i-1}' + W_i \qquad (3)$$

По сравнению с исходной таблицей 1 полученный ИВР гораздо нагляднее характеризует закономерности распределения цинка в месторождении. В частности, он показывает, с какой частотой встречается в залежи то или иное содержание цинка, что дает возможность более рационально проектировать добычу и обогащение руды.

При решении любых задач, касающихся исследуемого признака, ИВР полностью заменяет исходную выборку. Другими словами, в последующих расчетах m_i значений признака, попавших в i-й интервал, заменяются m_i значениями, равными середине интервала.

2. Строим гистограмму и кумулятивную кривую для более наглядного представления ИВР.

Для построения гистограммы (рисунок 1) по оси абсцисс откладываем в масштабе длины интервалов и на каждом из них, как на основании, строим прямоугольник, площадь которого равна частости соответствующего интервала. Практически для этого необходимо взять высоту прямоугольника, равной W_i/h. В рассматриваемом примере h=1, поэтому высоты прямоугольников равны соответствующим частостям.

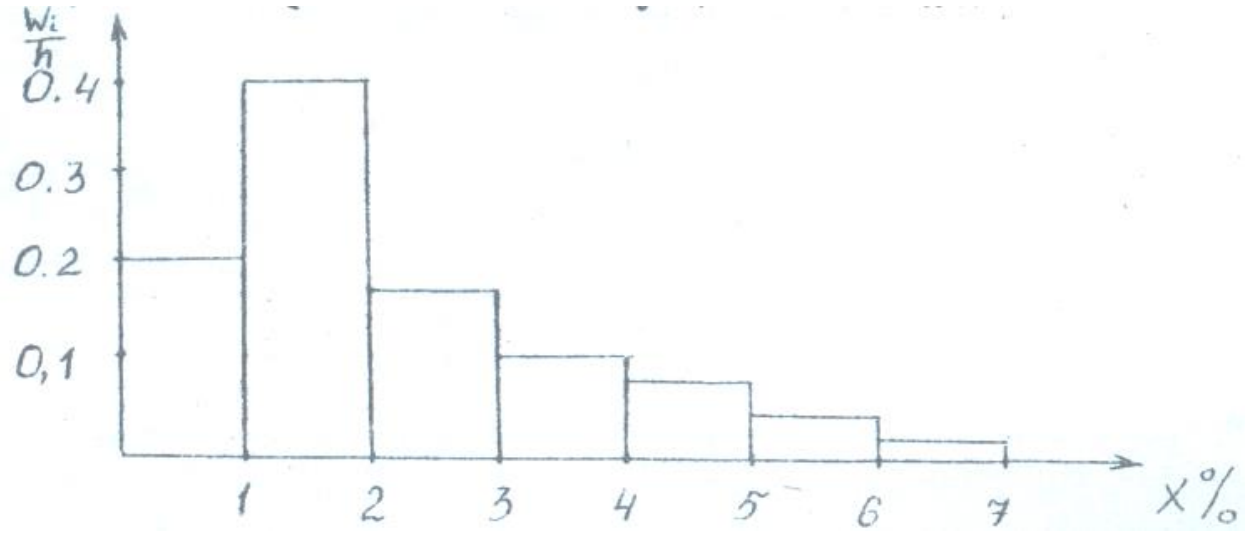

Рисунок 1. – Построение гистограммы.

При достаточно большом объеме выборки ступенчатая линия гистограммы наглядно показывает форму кривой функции плотности вероятности признака. Если статистическое исследование проводится для установления закона распределения признака, то выбор аналитического выражения для функции плотности вероятности осуществляется на основе построенной гистограммы.

Для построения кумулятивной кривой (рисунок 2) по оси абсцисс откладываем в масштабе границы всех интервалов, а из них по оси ординат - соответствующие накопленные частости. Соединяя полученные точки плавной линией, получаем кумулятивную кривую.

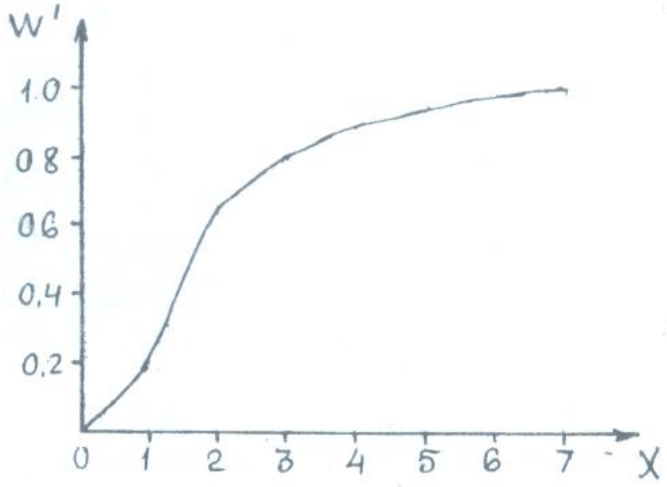

Рисунок 2. – Кумулятивная кривая.

3. Вычисляем статистические характеристики признака, используя для этого данные таблицы 2.

Статистической аналогией математического ожидания является среднее, вычисляемое по формуле:

$$X = \frac{\sum_{1}^{l} X_i \cdot m_i}{n}, \qquad (4)$$

где X_i – значение середины i-го интервала.

Величины X_i могут представлять собой многозначные числа. В этом случае определение среднего по формуле (4), а особенно вычисление последующих числовых характеристик становится весьма громоздким. Для упрощения вычислений переходят к условным единицам. Значения середин интервалов в условных единицах вычисляют по формуле:

$$X_i' = \frac{X_i - X_0}{h}, \qquad (5)$$

где X_0 – значение середины одного из интервалов, принятое за условный нуль.

Полученные значения середин интервалов в условных единицах при X_0=2,5 приведены в таблице 3.

Таблица 3.

Границы интервалов	m_i	X_i	X_i'	$X_i' \cdot m_i$	$X_i'^2 \cdot m_i$	$X_i'^3 \cdot m_i$	$X_i'^4 \cdot m_i$
1	2	3	4	5	6	7	8
0-1	17	0,5	-2	-34	68	-136	272
1-2	38	1,5	-1	-38	38	-38	38
2-3	14	2,5	0	0	0	0	0
3-4	8	3,5	1	8	8	8	8

Продолжение таблицы 3.

1	2	3	4	5	6	7	8
4-5	5	4,5	2	10	20	40	80
5-6	3	5,5	3	9	27	81	243
6-7	2	6,5	4	8	32	128	512
Итого				-37	193	83	1153

Вычисленную по данным таблицы 3 сумму $\sum_{1}^{l} X_i^{'} \cdot m_i$ подставляем в формулу (4) и получаем значение среднего в условных единицах:

$$X^{'} = \frac{-37}{87} = -0{,}425$$

Значение среднего в натуральных единицах определяется по формуле:

$$X = X_0 + h \cdot X^{'}, \qquad (6)$$

После подстановки в формулу (6) численных значений получим:

$$X = 2{,}5 - 1 \cdot 0{,}425 = 2{,}075\%$$

Модой признака называется такое его значение, которому соответствует наибольшее значение функции плотности вероятности.

Для нахождения моды по данным ИВР сначала определяют модальный интервал. Им является интервал, имеющий наибольшую частость. В рассматриваемом примере это интервал (1,2).

Если распределение симметрично и частости двух интервалов, прилегающих к модальному, одинаковы, то мода равна середине модального интервала.

Если распределение несимметрично, как это имеет место в примере, то моду вычисляют по формуле:

$$Mod = X_{Mod(\min)} + \frac{m_{Mod} - m_{Mod-1}}{m_{Mod} - m_{Mod-1} + m_{Mod} - m_{Mod+1}}, \quad (7)$$

где $X_{Mod(min)}$ – левая граница модального интервала;

m_{Mod}, m_{Mod-1}, m_{Mod+1} – частоты соответственно модального, предшествующего и последующего интервалов.

Подставляя численные данные в формулу (7), получим:

$$Mod = 1 + 1 \cdot \frac{38 - 17}{38 - 17 + 38 - 14} = 1{,}47\%$$

Это означает, что наиболее часто встречающимся в залежи содержанием цинка является 1,47%.

Медианой признака называется такое значение, для которого выполняется соотношение:

$$P(X > M_l) = P(X < M_l) = 0{,}5$$

т.е. вероятность для наблюденного значения признака оказаться больше и меньше медианы одинакова.

При наличии ИВР вычисление медианы производят по формуле:

$$M_l = X_{Ml(\min)} + h \cdot \frac{\frac{n}{2} - S}{m_{Ml}}, \quad (8)$$

где $X_{Ml(min)}$ – левая граница интервала, в котором лежит медиана (медианного);

S – сумма частот интервалов, предшествующих медианному;

m_{Ml} – частота медианного интервала.

Медианный интервал ищут подбором, прикидывая, какое количество значений признака окажется больше и меньше предполагаемой медианы, если она будет лежать в выбранном интервале.

Предположим, например, что медианным является интервал (3,2). Тогда, если даже в качестве медианы взять левую границу интервала (M_l =2), то количество значений признака, меньших медианы, составит 17+38=55, а больших 14+8+5+2+3=32. Естественно, для правой границы (M_l =3) это соотношение окажется еще менее благоприятным. Следовательно, интервал (2,3) выбран в качестве медианного неверно.

Если же выбрать в качестве медианного интервал (1,2) и повторить вышеуказанные рассуждения, то можно увидеть, что при переходе от левой границы к правой большее количество значений признака перемещается с правой стороны медианы в левую. Это означает, что медиана лежит в интервале (1,2).

Подставляя данные в формулу (8), вычисляем значения медианы:

$$M_l = 1 + 1 \cdot \frac{\frac{87}{2} - 17}{38} = 1{,}70\%$$

Остальные вычисляемые характеристики относятся к характеристикам рассеивания. Они характеризуют с разных сторон рассеивание признака относительно математического ожидания.

Для их нахождения необходимо предварительно вычислить начальные моменты признака. Основная часть из них не имеет самостоятельного значения и играет в вычислениях промежуточную роль.

Начальный момент K-го порядка вычисляется по формуле:

$$\alpha_K = \frac{\sum_1^l X_i^k \cdot m_i}{n}, \qquad (9)$$

Нетрудно видеть, что начальный момент 1-го порядка - это среднее значение признака *X*, которое вычислено ранее.

Вычисление моментов 2, 3 и 4-го порядков производим по формуле (9) в условных единицах. Для этого предварительно в таблице 3 подсчитываем соответствующие суммы $\sum X_i^{'k} m_i$:

$$\alpha_2^{'} = \frac{193}{87} = 2{,}218;$$
$$\alpha_3^{'} = \frac{83}{87} = 0{,}954;$$
$$\alpha_4^{'} = \frac{1153}{87} = 13{,}25.$$

Вычисление центральных моментов K-го порядка можно произвести по общей формуле:

$$\alpha_k^0 = \frac{\sum_1^l (X_i - X)^k \cdot m_i}{n}, \qquad (10)$$

Однако в случае, когда начальные моменты уже найдены, удобней вычислить центральные моменты по формулам:

$$\begin{aligned} \alpha_2^{0'} &= \alpha_2^{'} - \alpha_1^{'2}; \\ \alpha_3^{0'} &= \alpha_3^{'} - 3\alpha_1^{'}\alpha_2^{'} + 2\alpha_1^{'3}; \\ \alpha_4^{0'} &= \alpha_4^{'} - 4\alpha_1^{'}\alpha_3^{'} + 6\alpha_2^{'}\alpha_1^{'2} - 3\alpha_1^{'4}. \end{aligned} \qquad (11)$$

Поскольку начальные моменты выражены в условных единицах, центральные моменты также получаются в условных единицах:

$$\alpha_2^{0'} = 2{,}218 - (-0{,}425)^2 = 2{,}037;$$
$$\alpha_3^{0'} = 0{,}954 \cdot 3 \cdot (-0{,}425) \cdot 2{,}218 + 2 \cdot (-0{,}425)^3 = 3{,}6284;$$
$$\alpha_4^{0'} = 13{,}253 - 4 \cdot (-0{,}425) \cdot 0{,}954 + 6 \cdot 2{,}218 \cdot (-0{,}425)^2 - 3 \cdot (-0{,}425)^4 = 17{,}181.$$

Из центральных моментов важное самостоятельное значение имеет момент 2-го порядка, называемый дисперсией. Статистическая аналогия дисперсии (эмпирическая дисперсия) обозначается S^2 и может быть вычислена по общей формуле (10). Однако, если центральный момент 2-го порядка уже найден в условных единицах из выражения (11), то вычисление эмпирической дисперсии производят по формуле:

$$S^2 = h^2 \cdot \alpha_2^{0'}, \tag{12}$$

В нашем случае это составит:

$$S^2 = 1^2 \cdot 2{,}037 = 2{,}037$$

Поскольку дисперсия имеет размерность квадрата признака, она недостаточно наглядно характеризует его рассеивание. Поэтому на практике для характеристики рассеивания признака относительно среднего обычно пользуются средним квадратическим отклонением. В условных единицах оно находится из выражения:

$$S' = \sqrt{\alpha_2^{0'}} = \sqrt{2{,}037} = 1{,}427 \tag{13}$$

а в натуральных –

$$S = S' \cdot h = 1{,}427 \cdot 1 = 1{,}427\% \tag{14}$$

Среднее квадратическое отклонение выражается в тех же единицах, что и признак, и поэтому является абсолютной характеристикой рассеивания.

В качестве относительной характеристики рассеивания на практике используют коэффициент вариации. Его статистическая аналогия вычисляется по формуле:

$$\vartheta = \frac{S}{X} \cdot 100\%, \qquad (15)$$

из которой видно, что коэффициент вариации показывает в процентах соотношение между средним квадратическим отклонением и средним признака.

В рассматриваемом примере коэффициент вариации составляет:

$$\vartheta = \frac{1{,}427}{2{,}075} \cdot 100 = 69\%$$

Важно отметить, что при его вычислении значения S и X должны быть выражены в натуральных единицах.

Для оценки симметричности кривой распределения служит асимметрия, вычисляемая по формуле:

$$A = \frac{\alpha^{0'}}{S^{'3}}, \qquad (16)$$

В нашем примере она составит:

$$A = \frac{3{,}628}{1{,}427^3} = 1{,}25$$

Как видим, распределение имеет большую положительную асимметрию. Это говорит о том, что расположенная справа от среднего часть кривой функции плотности вероятности значительно длиннее, чем расположенная слева.

Для оценки плосковершинности или островершинности кривой распределения по отношению к кривой нормального закона служит эксцесс, который вычисляется по формуле:

$$E = \frac{\alpha_4^{0'}}{S'^4} - 3, \qquad (17)$$

Подставив численные значения, получим:

$$E = \frac{17{,}181}{1{,}427^4} - 3 = 1{,}14$$

Положительный эксцесс свидетельствует об островершинности распределения по сравнению с нормальным.

4. Рассмотрим реальное значение найденных характеристик, т.е. как они используются для решения практических задач. При этом будем учитывать следующие два обстоятельства.

Во-первых, при проведении статистического исследования вовсе не обязательно вычисляются все рассмотренные характеристики. Как правило, находятся только некоторые из них, которые нужны для решаемой практической задачи. Почти всегда вычисляются главные характеристики: среднее и среднее квадратическое отклонение. Другие характеристики вычисляются реже для решения некоторых частных вопросов. Часть характеристик может для исследуемого признака вообще не иметь реального значения. (В рассмотренном примере вычисление всех характеристик проведено для иллюстрации методики их нахождения).

Во-вторых, практическое значение той или иной характеристики наглядно проявляется при сопоставлении ее значений для двух различных генеральных совокупностей.

Если, например, для двух залежей вычислены средние содержания металла *X* и *У,* то даже без знания других характеристик можно сопоставлением средних оценить, какая из залежей богаче и, следовательно, какую при прочих равных условиях предпочтительней разрабатывать. Среднее имеет и самостоятельное значение. В частности, зная среднее содержание металла и запасы руды, можно очень просто определить количество металла в месторождении.

Если средние содержания металла двух залежей примерно одинаковы ($X \approx У$), то при прочих равных условиях ценности залежей можно сравнить, сопоставив средние квадратические отклонения S_X и $S_У$.

Если, например, $S_У > S_X$, то это говорит о том, что содержание металла во второй залежи значительно колеблется на разных участках. Чтобы обеспечить при отработке постоянный выход чистого металла, потребуется соответственно комбинировать объемы выемки с участков с разным содержанием металла, что, естественно, осложнит горные работы. Поэтому отработка первой залежи, где колебания содержания металла меньше, является более предпочтительной.

Подобным образом, сравнивая и другие характеристики, можно проанализировать, как влияет разница между ними на решение практических вопросов.

При выполнении практической работы необходимо сравнить найденные статистические характеристики с характеристиками другого варианта, указанного преподавателем, и сделать соответствующие выводы.

В отчете по практической работе необходимо привести условие и цели работы, подробный расчет ИВР и всех статистических характеристик, графики гистограммы и кумуляты, а также анализ сравнения результатов расчетов с данными другого варианта.

ЛАБОРАТОРНАЯ РАБОТА №2
«НАХОЖДЕНИЕ УРАВНЕНИЯ ПРЯМОЙ ЛИНИИ РЕГРЕССИИ И ОЦЕНКА ЕГО ТОЧНОСТИ»

I. Основные положения

Зависимость между случайными величинами X и $У$ называется корреляционной, если при изменении одной величины меняется

математическое ожидание другой. Его называют условным математическим ожиданием и обозначают Y_X.

Аналитическое выражение корреляционной зависимости называется, корреляционным уравнением, или уравнением регрессии. Если графически уравнение регрессии представляет собой прямую линию, то корреляционную зависимость называют линейной, в противном случае - нелинейной.

Нахождение и исследование корреляционного уравнения называется корреляционным анализом. При корреляционном анализе решаются следующие вопросы:

1) оценка тесноты связи, т.е. определение, какую долю в общем изменении величины *У* составляет влияние изменения величины *Х* и какую - остальных факторов;

2) нахождение параметров уравнения регрессии и оценка их точности;

3) оценка возможности практического использования найденного уравнения регрессии (обычно для целей прогноза).

Решение последнего вопроса зависит от тесноты связи и точности нахождения параметров уравнения, а также от характера той реальной задачи, для которой будет использоваться найденная зависимость.

2. Цель работы

Целью работы является:

1) овладение методикой нахолщения уравнения регрессии и оценки его точности;

2) овладение приемами практического использования корреляционных уравнений для решения производственных задач.

3. Методика выполнения работы

В практической работе необходимо исследовать зависимость между

двумя признаками объекта в соответствии с индивидуальным вариантом, выданным преподавателем.

Порядок выполнения работы подробно рассматривается на примере, который приводится ниже.

Пример:

При разведке и отработке поля шахты № 2 "Ганзовка" отобрано 124 пробы угля марки *К*. Для калщой пробы были определены содержание внутренней золы *А* и объемный вес γ. Полученные результаты приведены в таблицу 4. По этим данным необходимо выполнить следующее:

1) оценить наличие прямолинейной корреляционной зависимости между содержанием золы и объемным весом угля;

2) найти уравнение прямой линии регрессии;

3) оценить точность найденного уравнения регрессии и возможность его практического использования.

Таблица 4.

№ п/п	А, %	γ, т/м^3	№ п/п	А, %	γ, т/м^3	№ п/п	А, %	γ, т/м^3	№ п/п	А, %	γ, т/м^3
1	2	3	4	5	6	7	8	9	10	11	12
1	8,3	1,30	32	11,3	1,38	63			94	12,9	1,36
2	4,5	1,26	33	9,6	1,36	64	4,3	1,24	95	10,1	1,32
3	13,8	1,38	34	1,0	1,21	65	1,0	1,24	96	3,6	1,27
4	6,0	1,26	35	11,5	1,38	66	6,5	1,30	97	1,6	1,26
5	10,4	1,36	36	4,9	1,27	67	12,1	1,32	98	7,0	1,32
6	2,1	1,22	37	13,8	1,40	68	14,1	1,38	99	15,0	1,40
7	0,1	1,18	38	6,2	1,30	69	2,6	1,26	100	5,2	1,30
8	6,1	1,28	39	4,9	1,28	70	9,9	1,35	101	10,2	1,34
9	4,6	1,26	40	8,6	1,33	71	5,6	1,30	102	1,7	1,26
10	7,9	1,31	41	9,7	1,37	72	7,7	1,33	103	9,0	1,34

Продолжение таблицы 4.

1	2	3	4	5	6	7	8	9	10	11	12
11	10,6	1,36	42	0,9	1,22	73	8,8	1,35	104	7,1	1,32
12	8,4	1,32	43	2,4	1,24	74	1,3	1,24	105	3,5	1,28
13	9,1	1,28	44	11,8	1,37	75	6,6	1,30	106	5,3	1,30
14	4,2	1,24	45	8,0	1,28	76	12,2	1,32	107	7,2	1,32
15	0,5	1,20	46	4,3	1,24	77	7,8	1,36	108	1,9	1,27
16	2,1	1,22	47	11,9	1,37	78	2,8	1,26	109	7,6	1,35
17	4,7	1,26	48	8,1	1,28	79	6,7	1,30	110	3,8	1,30
18	10,9	1,36	49	1,1	1,23	80	1,4	1,24	111	7,3	1,34
19	9,3	1,36	50	13,9	1,40	81	9,8	1,36	112	13,1	1,38
20	13,9	1,38	51	6,3	1,30	82	14,0	1,36	113	9,1	1,35
21	6,2	1,28	52	5,0	1,28	83	5,8	1,30	114	2,0	1,20
22	0,7	1,20	53	2,4	1,24	84	6,8	1,32	115	3,9	1,32
23	11,0	1,36	54	8,2	1,30	85	1,4	1,25	116	15,3	1,42
24	2,2	1,24	55	8,7	1,34	86	3,1	1,26	117	4,0	1,22
25	4,8	1,28	56	1,1	1,23	87	12,6	1,34	118	5,9	1,30
26	1,0	1,20	57	9,5	1,36	88	10,0	1,36	119	7,4	1,32
27	8,5	1,32	58	12,0	1,31	89	8,9	1,33	120	4,0	1,23
28	9,4	1,36	59	1,2	1,24	90	3,4	1,27	121	2,1	1,22
29	11,1	1,38	60	2,5	1,26	91	1,5	1,26	122	9,2	1,35
30	2,2	1,24	61	5,1	1,28	92	14,6	1,40	123	7,5	1,32
31	1,0	1,24	62	6,4	1,34	93	6,9	1,32	124	5,9	1,32

Для общности приводимых в примере формул обозначим содержание золы через *X*, а объемный вес угля через *У*.

Основная часть вычислений выполняется в корреляционной таблице (таблица 5). Для ее составления рассчитываем ширину интервала для *X* и *У*:

$$h_X = \frac{X_{\max} - X_{\min}}{1 + 3{,}2 \lg n} = \frac{15{,}3 - 0{,}1}{1 + 3{,}2 \lg 124} = 1{,}98\%;$$
$$h_y = \frac{У_{\max} - У_{\min}}{1 + 3{,}2 \lg n} = \frac{1{,}42 - 1{,}18}{1 + 3{,}2 \lg 124} = 0{,}031 т / м^3. \quad (18)$$

Определяем количество интервалов для *X* и *У*:

$$K = \frac{X_{\max} - X_{\min}}{h_X} = \frac{15{,}3 - 0{,}1}{1{,}98} = 7{,}68 \approx 8;$$

$$l = \frac{Y_{\max} - Y_{\min}}{h_y} = \frac{1{,}42 - 1{,}18}{0{,}031} = 7{,}74 \approx 8.$$

Тогда окончательная ширина интервала составит:

для X

$$h_X = \frac{15{,}3 - 0{,}1}{8} = 1{,}9\%;$$

для Y

$$h_y = \frac{1{,}42 - 1{,}18}{8} = 0{,}03 m / м^3.$$

Используя полученные значения h_x и h_y, составляем корреляционную таблицу. За левую границу первого интервала принимаем минимальное значение признака. Прибавляя к ней многократно ширину интервала, получаем границы остальных интервалов и заносим их в корреляционную таблицу. Затем вычисляем средние значения всех интервалов X_i и Y_j и также записываем их в таблицу. Подсчитываем, сколько пар значений X и Y попало в каждую клетку таблицы и полученную величину m_{xy} записываем в левом верхнем углу соответствующей клетки.

Суммируем значения m_{xy} по каждому столбцу и определяем частоты m_x для каждого интервала по X. Суммируя те же величины m_{xy} по строчкам, определяем частоты m_y. Контроль вычислений осуществляется с помощью соотношения:

$$\sum_1^K m_X = \sum_1^l m_y = n.$$

Для упрощения последующих вычислений осуществляем переход к условным единицам. Значение середины каждого интервала в условных единицах вычисляется по формулам:

$$X_i' = \frac{X_i - X_0}{h_X}; \qquad Y_j' = \frac{Y_j - Y_0}{h_y}.$$

где X_0 и Y_0 – значения середин интервалов, выбранные за условный ноль.

Таблица 5.

Границы интервалов			0,1--2,0		2,0--3,9		3,9--5,8		5,8--7,7		7,7--9,6		9,6--11,5		11,5--13,4		13,4--15,3		m_y	$m_y Y_j$	$m_y Y_j^2$	$\sum m_{xy} X_i Y_j$
Средние X_i			1,05		2,95		4,85		6,75		8,65		10,55		12,45		14,35					
Средние Y_j			-3		-2		-1		0		+1		+2		+3		+4					
1,18--1,21	1,195	-3	4		1														5	-15	45	42
				36		6																
1,21--1,24	1,225	-2	5		3		2												10	-20	40	46
				30		12		4														
1,24--1,27	1,255	-1	8		8		6		1										23	-23	23	46
				24		16		6		0												
1,27--1,30	1,285	0	1		3		5		3		2								14	0	0	0
				0		0		0		0		0										
1,30--1,33	1,315	+1			2		3		16		5		1		3				30	30	30	9
						-4		-3		0		5		2		9						
1,33--1,36	1,345	+2							3		8		2		1				14	28	56	30
										0		16		8		6						
1,36--1,39	1,375	+3									4		10		5		4		23	69	207	165
												12		60		45		48				
1,39--1,42	1,405	+4															5		5	20	80	80
																		80				
m_X			18		17		16		23		19		13		9		9		124	89	481	
$m_X X_i'$			-54		-34		-16		0		19		26		27		36		$\sum m_X X_i' = 4$			
$m_X X_i'^2$			162		68		16		0		19		52		81		144		$\sum m_X X_i'^2 = 542$			
$\sum_{j=1}^{j=l} m_{XY} X_i Y_j$			90		30		7		0		33		70		60		128		$\sum_{1}^{Kl} m_{XY} X_i Y_j = 418$			
Y_X			1,235		1,259		1,272		1,312		1,337		1,366		1,352		1,392					

Полученные значения середин интервалов в условных единицах заносим в таблицу и все дальнейшие вычисления ведем в этих единицах.

Для каждого интервала по *X* находим произведения $X_i' m_X$ и $X_i'^2 m_X$, а затем их суммы:

$$\sum_1^K X_i' m_X = 4;\ \sum_1^K X_i'^2 m_X = 542.$$

Для *У* находим аналогично:

$$\sum_1^l Y_j' m_y = 89;\ \sum_1^l Y_j'^2 m_y = 481.$$

Для каждой клетки таблицы вычисляем произведение $m_{XY} X_i' Y_j'$ и записываем его в правом нижнем углу клетки. Суммируя произведения всех клеток таблицы, получаем:

$$\sum_1^{Kl} m_{XY} X_i Y_j = 418$$

Для контроля суммирование выполняется независимо по строчкам и столбцам. Полученные два значения должны быть одинаковыми.

На этом вычисления в таблице заканчиваются. Все остальные необходимые величины находим, пользуясь подсчитанными в таблице суммами.

Вычисляем средние значения зольности и объемного веса сначала в условных единицах:

$$X' = \frac{\sum_1^K X_i' m_X}{n} = \frac{4}{124} = 0{,}032;\ Y' = \frac{\sum_1^l Y_j' m_y}{n} = \frac{89}{124} = 0{,}718.$$

а затем в натуральных единицах:

$$X = X_0 + X' h_X = 6{,}75 + 0{,}032 \cdot 1{,}9 = 6{,}80\%;$$
$$Y = Y_0 + Y' h_Y = 1{,}285 + 0{,}718 \cdot 0{,}03 = 1{,}307 m/м^3.$$

Вычисляем средние квадратические отклонения в условных единицах:

$$X'^2 = \frac{\sum_1^K X_i'^2 m_X}{n} = \frac{542}{124} = 4{,}371;$$
$$Y'^2 = \frac{\sum_1^l Y_j'^2 m_Y}{n} = \frac{481}{124} = 3{,}879;$$
$$S_X'^2 = X'^2 - X'^2 = 4{,}371 - 0{,}032^2 = 4{,}370;$$
$$S_Y'^2 = Y'^2 - Y'^2 = 3{,}879 - 0{,}718^2 = 3{,}363;$$
$$S_X' = 2{,}09;$$
$$S_Y' = 1{,}83.$$

Переходим к натуральным единицам:

$$S_X = S_X' \cdot h_X = 2{,}09 \cdot 1{,}9 = 3{,}97\%;$$
$$S_Y = S_Y' \cdot h_Y = 1{,}83 \cdot 0{,}03 = 0{,}055 m/м^3.$$

Вычисляем коэффициент корреляции:

$$r = \frac{\sum_1^{Kl} m_{XY} X_i' Y_j' - nX'Y'}{nS_X' S_Y'} = \frac{418 - 124 \cdot 0{,}032 \cdot 0{,}718}{124 - 2 \cdot 0{,}9 \cdot 1{,}83} = 0{,}875 \qquad (19)$$

Находим среднее квадратическое отклонение коэффициента корреляции:

$$S_r = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - 0{,}86^2}{\sqrt{124}} = 0{,}023 \qquad (20)$$

Строим доверительный интервал для истинного значения коэффициента корреляции r_0 при доверительной вероятности $\beta = 0{,}997$:

$$r - 3S_r \le r_0 \le r + 3S_r;$$
$$0{,}79 \le r_0 \le 0{,}93.$$

Построенный доверительный интервал показывает, что истинное значение коэффициента корреляции не может быть равно 0. Это говорит о том, что корреляционная связь между признаками действительно существует, а величина коэффициента корреляции показывает, что связь тесная. Следовательно, нахождение уравнения регрессии имеет практический смысл.

Уравнение прямой регрессии имеет вид:

$$Y - \overline{Y} = r\frac{S_Y}{S_X} \cdot (X - \overline{X}) \qquad (21)$$

где переменными являются X и Y, а остальные величины определены из предыдущих вычислений. Подставляя их значения в (21), получим:

$$Y - 1{,}307 = 0{,}875 \cdot \frac{0{,}055}{3{,}97} \cdot (X - 6{,}8);$$
$$Y = 0{,}0121X + 1{,}224.$$

Пользуясь найденным уравнением, построим прямую регрессии (рис. 3) Для среднего значения каждого интервала вычислим условное среднее $\overline{Y}_{Xi}$ по формуле:

$$\overline{Y}_{Xi} = \frac{\sum_{1}^{l} m_{XY} Y_j}{m_X}. \qquad (22)$$

Полученные значения приведены в последней строке таблицы 5. Нанесем их на рисунок 3. В результате получится ломаная линия. Взаимное расположение ломаной и найденной прямой регрессии в определенной степени отражает правильность проведенных вычислений. Как известно, уравнение прямой находится исходя из принципа наименьших квадратов, т.е. таким образом, чтобы сумма квадратов уклонений вершин ломаной от прямой регрессии была минимальной. Положение прямой относительно ломаной на рисунке 3 соответствует этому принципу.

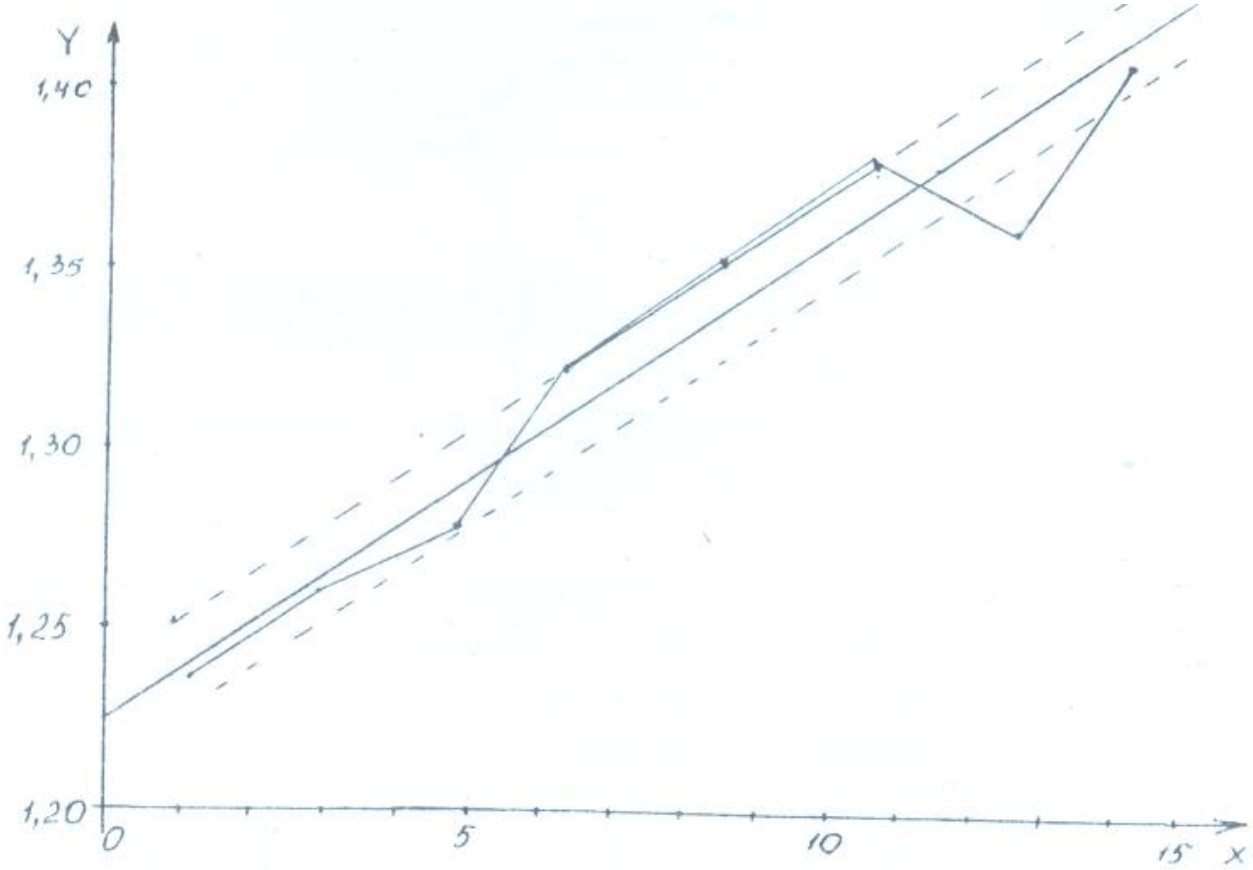

Рисунок 3 – Прямая регрессии.

Найденное корреляционное уравнение используется маркшейдерской службой шахты для определения объемного веса угля при подсчете добычи. Вместо трудоемкого определения объемного веса пробной вырубкой или лабораторным методом его находят с помощью корреляционного уравнения

по известному значению зольности, которая периодически определяется отделом технического контроля во всех очистных забоях.

При этом, естественно, возникает вопрос о точности определения объемного веса из уравнения регрессии.

Этот способ исходит из допущения, что зольность и объемный вес угля в пределах ограниченного участка (площади выемки лавы за 2-3 месяца) можно считать распределенными по нормальному закону с примерно постоянным математическим ожиданием. Однако на разных участках шахтного поля математические ожидания этих признаков меняются. Для определения месячной добычи маркшейдеру необходимо знать значение математического ожидания объемного веса на площади, вынутой за отчетный месяц. Поскольку прямая регрессии является изображением условных математических ожиданий, ее использование является подходящим способом нахождения объемного веса. Однако определение должно производиться не по случайному значению зольности для данного участка, а по ее математическому ожиданию. Поэтому рассматриваемый способ имеет смысл лишь в том случае, если для оцениваемого участка надежно определено математическое ожидание зольности путем взятия совокупности проб и вычисления ее среднего значения. Тогда точность определения математического ожидания объемного веса с помощью уравнения регрессии будет зависеть только от точности нахождения его параметров.

Средняя квадратическая ошибка вычисления условного математического ожидания Y_X с помощью корреляционного уравнения выражается формулой:

$$S_{YX} = S_{\Delta}\sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{S_X^2 n}}, \qquad (23)$$

где X – значение, для которого из корреляционного уравнения определяется соответствующее значение Y;

S_Δ – среднее квадратическое отклонение фактических значений величины $У$ от прямой регрессии, обусловленное действием побочных случайных факторов.

Значение S_Δ вычисляется по формуле:

$$S_\Delta = S_У \sqrt{1-r^2} = 0{,}055\sqrt{1-0{,}875^2} = 0{,}0266 \qquad (24)$$

Пользуясь формулой (23), определим величину $S_{УX}$ для средних значений интервалов X_i. Полученные значения приведены в таблице 6.

Таблица 6.

X_i	1,05	2,95	4,85	6,75	8,65	10,55	12,45	14,35
$S_{УX}$	0,0042	0,0033	0,0027	0,0024	0,0027	0,0033	0,0042	0,0051

Найденные величины $S_{УX}$ показывают, с какой средней квадратической ошибкой определяется величина $У_X$ из корреляционного уравнения для разных значений X_i. Как видим, точность вычисления $У_X$ различна: она выше при $X_i \approx \overline{X}$ и понижается по мере удаления от $\overline{X}$.

Задавшись уровнем значимости $q = 0{,}997$, построим доверительный интервал для $У_X$ шириной $\pm 3S_{УX}$. Графическое изображение интервала представлено на рисунке 3.

Ширина доверительного интервала для разных X_i показывает, что предельная ошибка определения объемного веса составляет 0,007- 0,015 т/м^3 (в относительной мере 0,6-1,1%). Это вполне приемлемо, поскольку добычу допускается определять с ошибкой до 3%.

Отчет по практической работе необходимо оформить аналогично рассмотренному примеру с детальным приведением всех вычислений.

ЛАБОРАТОРНАЯ РАБОТА №3
«ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ»

1. Основные положения

Статистическими гипотезами называются различного рода предположения о законах распределения случайных величин. Исходным материалом для проверки статистической гипотезы являются статистические данные наблюденные значения случайной величины (или нескольких случайных величин). Сущность проверки заключается в том, что путем соответствующего анализа статистических данных проверяется предположение (гипотеза) о виде закона распределения исследуемой случайной величины или о значениях параметров этого закона.

Выдвинутую гипотезу называют нулевой и обозначают H_0. Противоположную гипотезу называют конкурирующей и обозначают H_1.

Дня проверки выдвинутой гипотезы H_0 используют специально подобранную случайную величину, называемую критерием (для каждого вида статистических гипотез подбирается свой критерий). Выбор критерия осуществляется таким образом, чтобы:

1) при справедливости гипотезы H_0 он подчинялся известному закону распределения, а при справедливости гипотезы H_1 вид или параметры закона изменялись;

2) отдельное фактическое значение критерия можно было вычислить по имеющимся статистическим данным.

Общая схема проверки статистической гипотезы состоит в следующем. Предполагая гипотезу H_0 справедливой, по таблицам распределения критерия определяют, какое значение он может принять (допустимое значение). Затем по статистическим данным вычисляют фактическое значение критерия. Сравнивая фактическое и допустимое значения критерия, принимают или отклоняют нулевую гипотезу.

2. Цель работы

Целью работы является:

1) освоение методики проверки статистических гипотез о равенстве математических ожиданий, о равенстве дисперсий, о законе распределения;

2) уяснение практического значения освоенной методики и возможностей ее использования для решения производственных маркшейдерских задач.

3. Методика выполнения работы

Практическая работа состоит в решении шести задач с использованием методов проверки указанных статистических гипотез.

Порядок выполнения работы подробно рассматривается ниже на примерах. После каждого примера приводятся задачи для самостоятельного решения.

3.1. Проверка статистической гипотезы о равенстве математических ожиданий

Пример:

Для исследования гирокомпасов двух разных типов каждым из них было выполнено многократное ориентирование одной и той же стороны. Полученные результаты приведены в таблице 7.

Проверить, имеется ли между результатами ориентирования 1-м и 2-м гирокомпасами систематическая разность или расхождения между ними носят случайный характер.

Таблица 7.

1-й гирокомпас				2-й гирокомпас			
№ п/п	Значения дирекцион-ного угла X_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	№ п/п	Значения дирекцион-ного угла Y_i	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})^2$
1	164°25’35”	1	1	1	164°26’03”	12	144
2	42”	8	64	2	25’35”	-16	256
3	30”	-4	16	3	26’18”	27	729
4	14”	-20	100	4	25’43”	-8	64
5	58”	24	576	5	25’29”	-22	484
6	47”	13	169	6	26’10”	19	361
7	25”	-9	81	7	26’05”	14	196
8	18”	-16	256	8	25’16”	-35	1225
$\overline{X} = 164°25'34''$ $\sum (X_i - \overline{X})^2 = 1563$				9	25’40”	-11	121
				10	25’56”	5	25
				11	26’08”	17	289
				$\overline{Y} = 164°25'51''$ $\sum (Y_i - \overline{Y})^2 = 3894$			

Задачу решить для двух вариантов исходных данных:

а) средние квадратические ошибки ориентирования каждым гирокомпасом известны с высокой точностью и равны соответственно $\sigma_x = 15''$ и $\sigma_y = 20''$;

б) средние квадратические ошибки ориентирования неизвестны и определяются по данным, приведенным в таблице 7.

Уровень значимости *q* для обоих случаев принять равным 0,05.

Решение

Результаты измерений 1-м и 2-м гирокомпасами будем рассматривать как случайные величины *X* и *Y*, каждая из которых распределена по нормальному закону со средними квадратическими отклонениями соответственно σ_x и σ_y. Отклонения каждой случайной величины от своего математического ожидания носят случайный характер. Очевидно, разности между значениями *X* и *Y* также будут иметь случайный характер (будет отсутствовать систематическая составляющая), если колебания обеих величин

будут происходить вокруг одного и того же центра рассеивания, т.е. если будут равны их математические ожидания. Неравенство математических ожиданий будет означать наличие в разностях измерений 1-м и 2-м гирокомпасом систематической составляющей.

Таким образом, решение поставленной задачи сводится к проверке статистической гипотезы о равенстве математических ожиданий двух случайных величин X и $У$, распределенных по нормальному закону. Рассмотрим оба варианта решения, придерживаясь общего порядка проверки статистических гипотез.

Первый вариант решения

1. Выдвигаем нулевую гипотезу. В качестве гипотезы H_0 предположим, что $MX = MУ$. Тогда конкурирующая гипотеза H_1 : $MX \neq MУ$.

2. Выбираем критерий. В случае, если средние квадратические отклонения обеих нормальных совокупностей известны, для проверки гипотезы о равенстве их математических ожиданий используют критерий:

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma_d}, \tag{25}$$

где $\overline{X}$ и $\overline{Y}$ – средние, найденные соответственно из n_1 и n_2 наблюденных значений случайных величин X и $У$;

σ_d – средняя квадратическая ошибка разности средних, определяемая по формуле:

$$\sigma_d = \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}, \tag{26}$$

При справедливости гипотезы H_0 критерий Z имеет нормальное распределение с $MZ=0$ и $\sigma_z=1$. В случае справедливости гипотезы H_1 распределение остается нормальным, но $MZ\neq 0$.

3. Вычисляем фактическое значение критерия. Для этого сначала находим средние значения случайных величин:

$$\overline{X} = \frac{[X]}{n_1} = 164°25'34'';$$
$$\overline{Y} = \frac{[Y]}{n_2} = 164°25'51''.$$

Далее вычисляем по формуле (26) величину σ_d:

$$\sigma_d = \sqrt{\frac{15^2}{8} + \frac{20^2}{11}} = 8''$$

и находим фактическое значение критерия:

$$Z_\phi = \frac{164°25'34''-164°25'51''}{8} = \frac{-17}{8} = -2{,}12$$

4. Находим допустимое значение критерия Z_q, используя выражение:

$$\Phi(Z_q) = \frac{1-q}{2}, \quad (27)$$

где q – уровень значимости, равный согласно условию 0,05;

Ф – функция Лапласа, значения которой задаются таблично в зависимости от Z [I].

Для нахождения Z_q сначала из выражения (27) вычисляем значение функции:

$$\Phi(Z_q) = \frac{1-0{,}05}{2} = 0{,}475$$

а затем из таблиц по $\Phi(Z_q)$ находим значение аргумента Z_q= 1,96.

5. Сравниваем фактическое и допустимое значения критерия. При этом оказывается, что $|Z_\Phi| > Z_q$, т.е. фактическое значение критерия попадает в критическую область. На основании этого отклоняем нулевую гипотезу и принимаем гипотезу H_1, т.е. считаем, что $MX \neq MY$. Следовательно, разность между средними значениями дирекционного угла, полученными 1-м и 2-м гирокомпасами, объясняется не только случайными погрешностями их определения из-за ограниченного количества данных, а и наличием систематической разности между результатами измерений двумя гирокомпасами.

Второй вариант решения

1. Для второго варианта задачи, нулевая и конкурирующие гипотезы остаются прежними, поэтому решение начинаем со второго шага - выбора критерия.

2. Во втором варианте точные значения σ_x и σ_y отсутствуют и вместо них предполагается использовать эмпирические средние квадратические отклонения S_X и S_Y, найденные соответственно из n_1=8 и n_2=11 измерений. Вследствие ограниченности количества измерений значения S_X и S_Y получаются с погрешностями и соответственно неточной окажется и средняя квадратическая разность S_d, найденная по формуле (26). В результате критерий Z, определяемый выражением (25), уже не будет распределен по нормальному закону и поэтому не может использоваться для проверки статистической гипотезы в описанном выше порядке.

При использовании вместо σ_x и σ_y эмпирических средних квадратических отклонений S_X и S_Y проверка статистической гипотезы о равенстве математических ожиданий производится с помощью критерия:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1) \cdot S_X^2 + (n_2 - 1) \cdot S_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \qquad (28)$$

При справедливости гипотезы H_0, т.е. при M_X=M_Y критерий t имеет распределение Стьюдента с числом степеней свободы $K = n_1 + n_2 - 2.$

3. Вычисляем фактическое значение критерия $t_Ф$.

Для этого сначала находим эмпирические средние квадратические отклонения S_X и S_Y, используя данные таблица 7:

$$S_X = \sqrt{\frac{\left[\left(X_i - \overline{X}\right)^2\right]}{n_1 - 1}} = \sqrt{\frac{1563}{8-1}} = 14'',9;$$

$$S_Y = \sqrt{\frac{\left[\left(Y_i - \overline{Y}\right)^2\right]}{n_2 - 1}} = \sqrt{\frac{3894}{11-1}} = 19'',7.$$

Вычисляем фактическое значение критерия:

$$t_Ф = \frac{164°25'34'' - 164°25'51''}{\sqrt{(8-1) \cdot 14,9^2 + (11-1) \cdot 19,7^2}} \sqrt{\frac{8 \cdot 11 \cdot (8+11-2)}{8+11}} = -2,04.$$

4. По таблицам распределения Стьюдента находим допустимую величину критерия t_q для уровня значимости q=0,05 и числа степеней свободы $K = 8+11-2 = 17$. Получаем t_q=2,11.

5. Сравнивая $t_Ф$ и t_q, видим, что фактическое значение критерия попадает в область допустимых значений (– t_q, + t_q). На основании этого принимаем нулевую гипотезу, т.е. считаем, что *MX*=*MУ* и расхождения между результатами ориентирования могут быть объяснены случайными факторами.

Хотя в обоих вариантах задачи разность средних значений дирекционного угла осталась одной и той же $(\overline{X} - \overline{Y} = -17'')$, в первом случае на основании ее делается вывод о наличии систематической составляющей в разности, а во втором случае такой вывод сделать нельзя. Это объясняется тем, что для второго варианта допустимое значение t_q при том же q=0,05 оказалось несколько большим, так оно учитывает не только случайные ошибки измерений обоими гирокомпасами, но и возможную неточность определения S_X и S_Y по ограниченному количеству измерений.

Задачи для самостоятельного выполнения

1. Перед началом работ по ориентированию сторон в шахте поправка гирокомпаса была определена как среднее из n_1, пусков и оказалась равной δ_1. По окончании работ в шахте поправка была определена как среднее из n_2 пусков и оказалась равной δ_2. Средняя квадратическая ошибка определения поправки одним пуском равна m. Приняв уровень значимости равным q, проверить, изменилась ли поправка гирокомпаса за истекший период. Численные значения исходных данных приведены в таблице 8.

Таблица 8.

Вариант	n_1	δ_1	n_2	δ_2	m	q%
1	2	3	4	5	6	7
1	9	1°05’16”	12	1°05’27”	20	5
2	10	02”	16	25”	15	2
3	12	24”	9	48”	20	1
4	15	51”	7	32”	15	10
5	16	17”	8	59”	20	5
6	8	36”	15	12”	15	2
7	10	41”	13	03”	20	1
8	12	19”	11	40”	15	10
9	14	29”	9	45”	20	1

Продолжение таблицы 8.

1	2	3	4	5	6	7
10	16	38”	8	15”	15	2
11	18	15”	7	29”	20	5
12	9	54”	15	37”	15	10
13	12	43”	13	28”	20	1
14	14	20”	11	41”	15	2
15	16	08”	9	15”	20	5
16	8	12”	7	24”	15	10
17	10	26”	16	41”	20	1
18	12	34”	14	52”	15	2
19	14	53”	12	38”	20	5
20	16	47”	10	28”	15	10
21	9	1°05’17”	8	1°05’31”	20	1
22	11	39”	7	25”	15	2
23	13	22”	9	34”	20	5
24	15	46”	11	32”	15	10
25	16	53”	13	40”	20	5

2. Технический теодолит исследовался путем многократного измерения эталонного угла, истинное значение которого равно β_0. Всего было выполнено n измерений, среднее из которых составило $\overline{\beta}$. Средняя квадратическая ошибка отдельного измерения, вычисленная по уклонениям от среднего, составила m. Приняв уровень значимости равным q, проверить наличие систематической ошибки в результатах измерений исследуемым теодолитом. Проверку произвести с использованием нормального распределения и распределения Стьюдента.

Численные значения исходных данных приведены в таблице 9.

Таблица 9.

Вариант	β_0	$\overline{\beta}$	n	m	q, %
1	2	3	4	5	6
1	84°17’38”,1	84°17’29”,8	22	18”,7	2
2	91°06’32”,2	91°06’43”,2	25	20”,1	5
3	112°54’08”,4	112°54’15”,6	30	19”,4	10

Продолжение таблицы 9.

1	2	3	4	5	6
4	167°41’37”,3	167°41’25”,1	34	22”,3	6
5	123°19’08”,5	123°19’20”,3	29	18”,8	1
6	79°25’14”,9	79°25’09”,1	26	17”,6	2
7	85°39’21”,6	85°39’32”,8	24	19”,1	3
8	129°49’57”,1	129°49’50”,4	28	21”,9	7
9	93°35’20”,8	93°35’10”,7	32	16”,8	10
10	151°08’27”,7	151°08’20”,3	35	17”,1	6
11	93°11’28”,6	93°11’20”,6	40	17”,4	5
12	98°18’06”,4	98°18’20”,8	20	21”,7	3
13	148°53’29”,3	148°53’36”,4	38	20”,9	2
14	108°16’32”,9	108°16’27”,6	36	21”,5	3
15	99°21’39”,7	99°21’51”,9	21	17”,6	2
16	87°15’09”,8	87°15’20”,8	23	17”,9	1
17	83°09’31”,9	83°09’39”,6	27	21”,6	5
18	86°51’39”,4	86°51’28”,1	29	18”,1	6
19	83°17’08”,5	83°17’19”,6	31	18”,9	4
20	79°14’32”,1	79°14’39”,7	33	18”,4	3
21	102°06’52”,3	102°06’41”,4	32	19”,6	2
22	117°23’14”,2	117°23’05”,1	36	19”,9	6
23	108°27’15”,7	108°27’29”,3	35	20”,2	8
24	96°33’49”,6	96°33’35”,7	32	20”,5	10
25	95°01’17”,4	95°01’06”,3	34	21”,6	6

3. У некоторых гирокомпасов МВТ2 в процессе пуска имеет место температурная разбалансировка, т.е. смещение центра тяжести чувствительного элемента из-за повышения температуры, вызванного работой гиромотора. Разбалансировка вызывает одностороннее отклонение положения равновесия колебаний и приводит к систематической ошибке в определении гироскопического азимута.

При выполнении на ориентируемой стороне двух пусков подряд действие разбалансировки наиболее сильно проявляется в первом пуске, когда интенсивно идет процесс нагревания. Во втором пуске ее действие значительно ослабевает вследствие стабилизации температуры.

Для проверки наличия разбалансировки у гирокомпаса МВТ2 № 07 было проанализировано n разностей двойных определений гироскопического

азимута ориентируемых сторон. При этом среднее арифметическое значение разности составило $\overline{d}$, а ее среднее квадратическое значение m_d (по уклонениям от среднего арифметического).

Приняв уровень значимости равным q, проверить наличие разбалансировки у указанного гирокомпаса. Численные значения исходных данных приведены в таблице 10.

Таблица 10.

Вариант	n	$\overline{d}$	m_d	q, %
1	30	20”,1	45”,0	4
2	55	-9”,8	44”,1	5
3	42	14”,2	39”,8	2
4	32	13”,8	30”,6	1
5	53	-6”,2	27”,3	4
6	41	14”,1	32”,9	1
7	37	-8”,3	38”,6	5
8	51	-9”,4	37”,5	2
9	48	13”,2	34”,1	3
10	39	8”,9	35”,3	4
11	50	-15”,4	38”,6	3
12	40	6”,3	39”,4	6
13	36	-12”,7	41”,6	1
14	35	10”,0	25”,3	2
15	48	-8”,1	26”,9	5
16	47	7”,9	27”,3	4
17	53	-9”,3	29”,1	6
18	41	12”,4	33”,8	1
19	43	-10”,2	31”,7	2
20	35	-12”,3	35”,6	3
21	46	8”,3	32”,7	5
22	42	-10”,9	31”,8	6
23	38	9”,7	30”,2	4
24	33	-13”,4	38”,7	2
25	44	12”,1	37”,4	1

3.2. Проверка статистической гипотезы о равенстве дисперсий

Пример

Для сравнения точности двух теодолитов каждым из них проводились многократные измерения одного угла. Первым теодолитом было выполнено n_1=50 измерения, вторым – n_1=80 измерений. При этом средние квадратические отклонения, вычисленные по отклонениям от среднего, составили соответственно S_1 =14",8 и S_2= 12",6. Приняв уровень значимости q=10%, проверить, действительно ли теодолиты имеют разную точность.

Решение

1. В качестве нулевой гипотезы выберем предположение, что точность теодолитов одинакова, т.е. равны их дисперсии:

$$H_0 : \sigma_1^2 = \sigma_2^2,$$

Тогда

$$H_1 : \sigma_1^2 \neq \sigma_2^2,$$

2. Для проверки гипотезы о равенстве дисперсий в качестве критерия используется величина:

$$F = \frac{S_1^2}{S_2^2}, \qquad (29)$$

причем в числителе ставят большую из эмпирических дисперсий.

При справедливости гипотезы H_0 величина F подчинена F – распределению (распределению Фишера) с числом степеней свободы $K_1 = n_1 - 1$ и $K_2 = n_2 - 1$, где n_1 – количество данных для большей дисперсии.

3. Вычисляем фактическое значение критерия.

Для этого подставляем в формулу (29) фактические значения средних квадратических отклонений:

$$F_\Phi = \frac{14",8^2}{12",0^2} = 1{,}52.$$

4. По числу степеней свободы $K_1 = 50 - 1 = 49$ и $K_2 = 80 - 1 = 79$ находим в таблицах F – распределения [3] допустимое значение критерия F_q=1,51.

5. Сравниваем фактическое и допустимое значения критерия. Поскольку $F_Ф > F_q$, отклоняем нулевую гипотезу и принимаем конкурирующую, т.е. $\sigma_1^2 \neq \sigma_2^2$.

Другими словами, разность между эмпирическими дисперсиями S_1^2 и S_2^2 настолько существенна, что не может быть объяснена неточностью их определения по ограниченному числу данных, а свидетельствует о неодинаковой точности теодолитов.

Задачи для самостоятельного решения

1. На руднике проводилось сравнение точности бороздового и шпурового методов опробования. Для этого при проходке выработки отбирались пробы на содержание металла бороздовым и шпуровым методами (одна рядом с другой). Всего было отобрано n пар проб. По полученным данным для каждого метода были вычислены эмпирические средние квадратические отклонения, которые составили для бороздового метода $S_б$ и для шпурового $S_ш$.

Приняв уровень значимости равным q, проверить предположение, что точность обоих способов одинакова. Числовые значения исходных данных приведены в таблице 11.

Таблица 11.

Вариант	n	$S_б$	$S_ш$	q, %
1	2	3	4	5
1	21	0,225	0,285	10
2	17	0,184	0,275	2
3	15	0,249	0,369	10
4	21	0,245	0,342	10

Продолжение таблицы 11.

1	2	3	4	5
5	31	0,189	0,261	2
6	26	0,234	0,302	10
7	21	0,238	0,304	2
8	15	0,191	0,254	10
9	13	0,192	0,173	2
10	41	0,232	0,201	2
11	26	0,227	0,189	2
12	31	0,199	0,241	10
13	21	0,202	0,253	10
14	17	0,223	0,314	2
15	15	0,215	0,308	2
16	13	0,219	0,296	10
17	26	0,193	0,231	2
18	36	0,242	0,314	10
19	46	0,251	0,308	10
20	17	0,188	0,207	2
21	31	0,199	0,231	2
22	21	0,214	0,188	10
23	46	0,218	0,314	10
24	36	0,229	0,317	2
25	21	0,234	0,175	10

2. Применительно к гирокомпасу МВТ2 производилось сравнение точности двух методов определения гироскопического азимута: метопа реверсии и метода прохождений. Для этого каждым из способов было произведено соответственно n_p и n_n определений гироскопического азимута одной и той же стороны и по отклонениям от средних вычислены эмпирические средние квадратические отклонения. Они оказались равными для метода реверсий S_p, для метода прохождений S_n.

Приняв уровень значимости равным q, выяснить, являются ли способы равноценными по точности или один из них точнее. Числовые значения исходных данных приведены в таблице 12.

Таблица 12.

Вариант	n_p	n_n	S_p	S_n	q, %
1	30	15	15”,4	17”,2	10
2	29	17	14”,8	18”,6	2
3	21	20	15”,9	14”,3	10
4	31	25	27”,1	19”,3	2
5	27	41	22”,4	33”,6	10
6	26	21	18”,3	16”,3	2
7	16	31	24”,2	35”,6	10
8	28	41	17”,3	28”,6	2
9	19	17	23”,5	36”,1	10
10	23	15	18”,3	25”,4	2
11	25	21	15”,2	17”,3	10
12	24	15	24”,8	32”,7	10
13	28	13	13”,1	21”,5	2
14	18	17	15”,8	22”,8	10
15	19	21	17”,6	20”,3	10
16	36	26	15”,3	22”,8	2
17	17	31	20”,3	32”,5	10
18	41	17	22”,1	17”,8	10
19	32	26	20”,7	29”,5	10
20	34	17	21”,5	23”,8	2
21	29	21	19”,8	27”,8	10
22	38	31	18”,3	30”,8	2
23	27	41	17”,6	23”,8	10
24	40	26	14”,3	15”,2	10
25	15	42	15”,8	14”,3	10

3.3. Проверка статистической гипотезы о законе распределения

При проверке двух ранее рассмотренных статистических гипотез закон распределения исследуемых статистических данных предполагался известным. Однако в ряде случаев он оказывается неизвестным. Тогда задача, как правило, формулируется следующим образом: имеется совокупность статистических данных (значения случайной величины), необходимо определить, какому закону распределения подчинена эта случайная величина.

Такая задача решается путем проверки статистической гипотезы о законе распределения. Проверка осуществляется в обычном порядке, т.е.

прежде всего выдвигается нулевая гипотеза - делается предположение о законе распределения случайной величины. Такое предположение можно сделать, построив ИВР признака и вычислив числовые характеристики.

Проверка гипотезы о законе распределения может производиться с помощью различных критериев. Наиболее употребительным из них является критерий Пирсона:

$$X^2 = \sum_{1}^{l} \frac{(m_i - nP_i)^2}{n \cdot P_i}, \qquad (30)$$

где n – количество статистических данных;

l – количество интервалов при построении ИВР;

m_i – частота попадания в интервал;

p_i – теоретическая вероятность попадания в интервал, определяемая с помощью функции плотности вероятности предполагаемого закона распределения.

Пирсон показал, что при совпадении предполагаемого и фактического законов распределения (при справедливости гипотезы H_0) критерий (30) имеет распределение X^2 с числом степеней свободы:

$$K = l - r - 1 \qquad (31)$$

где r – число параметров предполагаемого закона распределения, вычисляемых по статистическим данным. Если параметры находятся без использования исследуемой совокупности данных, то r =0.

Идея использования критерия Пирсона состоит в следующем. Если гипотеза H_0 верна, и следовательно, критерий (30) имеет распределение X^2, то фактическое значение критерия X_{Φ}^2 не должно превысить допустимого значения X_q^2, найденного по таблицам для уровня значимости q и числа степеней свободы K . Если же предполагаемый закон распределения выбран

неверно, то фактическая частота попадания в каждый интервал m_i и теоретическая частота $n \cdot P_i$ будут существенно различаться между собой. В результате значение критерия (30) получится увеличенным и превзойдет допустимое значение X_q^2. При этом нулевая гипотеза будет отвергнута.

Пример

В процессе эксплуатации гирокомпаса МВТ2 было выполнено 289 двойных определений гироскопических азимутов сторон и для каждой пары определений вычислена разность d_i. Используя полученные разности, необходимо выяснить, подчиняются ли ошибки измерений нормальному закону с математическим ожиданием, равным 0. Уровень значимости принять q = 5%.

Решение

Если ошибки измерений гироскопических азимутов подчинены нормальному закону с *MO*=0, то разности каждой пары измерений d_i также распределены нормально с *M(d)*=0. Следовательно, решение поставленной задачи сводится к проверке статистической гипотезы, что разности d_i распределены по нормальному закону с *M(d)*=0.

Для проверки статистической гипотезы о законе распределения необходимо для имеющейся совокупности разностей предварительно построить ИВР. Построение производится по известной методике, которая детально рассмотрена в практической работе I. Поэтому ниже не приводятся промежуточные вычисления, а даны лишь окончательные результаты (таблица 13).

Таблица 13.

Границы интервалов		Частоты m_i
1		2
-90	-70	4
-70	-50	13
-50	-30	29

Продолжение таблицы 13.

1		2
-30	-10	60
-10	+10	79
+10	+30	53
+30	+50	31
+50	+70	17
+70	+90	3

Кроме того, по данным ИВР было вычислено среднее квадратическое отклонение разности $S_d = 28''$.

Имея перечисленные исходные данные, можем приступить собственно к проверке гипотезы о законе распределения.

1. В качестве нулевой выдвигаем гипотезу, что разности подчинены нормальному закону распределения с *M(d)*=0 и средним квадратическим отклонением $S_d = 28''$. Другими словами, предполагаем, что функция плотности вероятности разности имеет вид:

$$f(d) = \frac{1}{28\sqrt{2\pi}} lxp\, \frac{d^2}{2 \cdot 28^2}, \qquad (32)$$

2. В качестве критерия для проверки выдвинутой гипотезы используем критерий Пирсона (30).

3. Вычислим фактическое значение критерия. Для этого предварительно найдем входящие в выражение (30) теоретические вероятности P_i. Это вероятности попадания фактической разности в каждый интервал при условии, что разность подчинена нормальному закону, определяемому выражением (32).

Вычисление вероятностей производим по формуле:

$$P(d_{i\pi} \le d \le d_{in}) = \Phi\left(\frac{d_{i\pi} - M(d)}{S_d}\right) - \Phi\left(\frac{d_{in} - M(d)}{S_d}\right) \qquad (33)$$

где $d_{iл}$ и d_{in} – соответственно левая и правая границы *i*-го интервала;

Ф – функция Лапласа.

Вычисление значений P_i производим в таблице 14.

Таблица 14.

Границы интервала		$\frac{d_{in}-M(d)}{S_d}$	$\frac{d_{iл}-M(d)}{S_d}$	$\Phi\left(\frac{d_{in}-M(d)}{S_d}\right)$	$\Phi\left(\frac{d_{iл}-M(d)}{S_d}\right)$	P_i
$-\infty$	-70	-2,50	-	-0,494	-0,500	0,006
-70	-50	-1,79	-2,50	-0,463	-0,494	0,031
-50	-30	-1,07	-1,79	-0,358	-0,463	0,105
-30	-10	-0,36	-1,07	-0,141	-0,358	0,217
-10	10	0,36	-0,36	0,141	-0,141	0,282
10	30	1,07	0,36	0,358	0,141	0,217
30	50	1,79	1,07	0,463	0,358	0,105
50	70	2,50	1,79	0,494	0,463	0,031
70			2,50	0,500	0,494	0,006
						$\sum P_i = 1{,}000$

В рассматриваемом примере для предполагаемого закона распределения параметр *M(d)*=0. В общем случае математическое ожидание отлично от 0. Тогда в качестве его предполагаемого значения принимается среднее.

Вычисление фактического значения критерия производим в таблице 15.

Таблица 15.

Интервал		m_i	P_i	nP_i	$m_i - nP_i$	$\frac{(m_i - nP_i)^2}{n \cdot P_i}$
$-\infty$	-70	4	0,006	1,73	2,27	2,98
-70	-50	13	0,031	8,96	4,04	1,82
-50	-30	29	0,105	30,34	-1,34	0,06
-30	-10	60	0,217	62,71	-2,71	0,12
-10	+10	79	0,282	81,50	-2,50	0,08
+10	+30	53	0,217	62,71	-9,71	1,50
+30	+50	32	0,105	30,34	1,66	0,09
+50	+70	16	0,031	8,96	7,04	5,53
+70		3	0,006	1,73	1,27	0,93
						$X^2_{\Phi} = 13{,}11$

4. Для нахождения допустимого значения критерия предварительно рассчитываем число степеней свободы по формуле (31). При этом значение r принимаем равным 1, так, как только один из параметров предполагаемого закона распределения (среднее квадратическое отклонение S_d) вычислен по данным выборки. В результате получаем:

$$K = 9 - 1 - 1 = 7$$

Используя найденное число степеней свободы и уровень значимости q=0,05, по таблицам распределения X^2[3] находим допустимое значение критерия:

$$X_q^2 = 14{,}1$$

5. Сравнивая фактическое и допустимое значения критерия, получаем $X_Ф^2 < X_q^2$. Следовательно, гипотеза H_0 не противоречит опытным данным и поэтому принимается. Другими словами, можно считать, что исследуемые разности подчинены нормальному закону.

Задание для самостоятельного выполнения

Используя результаты практической работы 1 (ИВР, среднее и среднее квадратическое отклонение), необходимо проверить гипотезу, что исследованный в работе признак распределен по нормальному закону. Проверку произвести для уровней значимости 2% и 10%.

Список рекомендуемой литературы

1. Рыжов П.А. Математическая статистика в горном деле. - М.: Высш. шк., 1973. - 287 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высш. шк., 1972. - 368 с.

3. Смирнов Н.В., Белугин Д.А. Теория вероятностей и математическая статистика в приложении к геодезии. - М.: Недра, 1969. - 379 с.