МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение
высшего образования
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ»

Н.А. Богородская Е.М. Лукина

Кафедра №85

СТАТИСТИКА

Методические указания к практическим занятиям

Санкт-Петербург

1. МЕТОД ГРУППИРОВОК

1.1. Методические указания к решению задач по теме «Метод группировок»

Для изучения структуры статистической совокупности, наличия взаимосвязей между явлениями, их признаками используется статистическая группировка.

Группировка заключается в образовании групп единиц совокупности, однородных в каком-либо существенном отношении, а также имеющих одинаковые или близкие значения группировочного признака. В зависимости от задач исследования используются различные виды группировки статистической информации (табл. 1.1).

Таблица 1.1

Виды группировок

Признак	Вид группировки
Содержание статистической информации	Первичная Вторичная
Число группировочных признаков	Простая Комбинационная Многомерная
Задачи, решаемые с помощью группировки	Типологическая Структурная Аналитическая Территориальная

Первичная группировка производится непосредственно по первичным данным статистического наблюдения.

Вторичная группировка используется для образования новых групп на основе ранее произведенной первичной группировки. Необходимость в перегруппировке данных возникает в тех случаях, когда первичная группировка содержит больше (или меньше) групп, чем это необходимо для характеристики типичных отношений и связей, и когда необходимо получить сопоставимые данные по нескольким группировкам.

При простой группировке объединение единиц совокупности в группы производится по одному какому-либо признаку.

При комбинационной (комбинированной) группировке производится разбиение статистической совокупности на группы по двум и более признакам, взятым в сочетании (комбинации). Сначала образуются группы по одному признаку, затем выделенные группы подразделяются на подгруппы по другому признаку, в свою очередь выделенные подгруппы разделяются на подгруппы по следующему признаку и т.д.

Многомерная группировка производится по величине средней многомерной.

Аналитическая группировка служит для выявления наличия взаимосвязей между изучаемыми явлениями и их признаками. Взаимосвязанные признаки делятся на факторные и результативные признаки. При этом группы образуются по факторному признаку, а для каждой выделенной группы рассчитывается либо среднее значение результативного признака, либо относительные величины.

С использованием типологической группировки в изучаемой совокупности явлений выделяются однокачественные в существенном отношении группы, прежде всего классы и социально-экономические типы.

Структурная группировка выявляет состав (строение) однородной в качественном отношении статистической совокупности.

При территориальной группировке осуществляется распределение сводных статистических данных по экономико-географическому и административно-территориальному признакам.

В процессе группировки выполняются две таблицы: разработочная (рабочая) и результативная (итоговая). В разработочной таблице должны быть представлены упорядоченные (в соответствии с установленными интервалами) данные отдельно по каждой единице совокупности и итоговые данные по группам. В результативной таблице приводятся суммарные и средние показатели по группам и по всей совокупности в целом.

Интервалы группировки могут быть равными и неравными, закрытыми и открытыми. Закрытыми называются интервалы группировки, у которых обозначены обе границы интервалов, открытыми – интервалы, у которых указана только одна граница: нижняя – у первого, верхняя – у последнего интервала группировки. При обработке статистических данных открытые интервалы необходимо закрывать. Величина первого интервала принимается равной величине второго, а последнего – величине предыдущего.

Величина равных интервалов определяется по формуле

$$\Delta = \frac{x_{\max} - x_{\min}}{n},$$

где $x_{\max}, x_{\min}$ – максимальное и минимальное значение признака;

n – заданное количество интервалов группировки.

Для определения величины интервала группировок может быть также использована формула Стерджесса (обычно при незначительной вариации признаков)

$$\Delta = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg N},$$

где N – число единиц совокупности.

1.2. Примеры решения задач по теме «Метод группировок»

Задача 1.1

Имеются данные о стаже и выработке рабочих (табл. 1.2).

Таблица 1.2

Данные о стаже и выработке рабочих

Номер рабочего	Стаж работы, лет	Дневная выработка рабочего, тыс. р.	Номер рабочего	Стаж работы, лет	Дневная выработка рабочего, тыс. р.
1	1,0	2,00	16	10,5	2,76
2	1,0	2,02	17	1,0	2,34
3	3,0	2,05	18	9,0	2,70
4	6,5	2,90	19	9,0	2,64
5	9,2	2,93	20	6,5	2,52
6	4,4	2,50	21	5,0	2,41
7	6,9	2,80	22	6,0	2,56
8	2,5	2,30	23	10,1	2,62
9	2,7	2,23	24	5,5	2,45
10	16,0	3,10	25	2,5	2,40
11	13,2	2,84	26	5,0	2,44
12	14,0	3,20	27	5,3	2,52
13	11,0	2,95	28	7,5	2,53
14	12,0	2,79	29	7,0	2,52
15	4,5	2,22	30	8,0	2,62

По данным табл. 1.2 выполнить следующие группировки:

– **структурную группировку** рабочих по стажу, образовав пять групп с равными интервалами. Рассчитать удельный вес (относительную величину структуры) рабочих в каждой группе;

– **простую аналитическую группировку** рабочих по стажу для изучения наличия взаимосвязи между стажем работы и месячной выработкой рабочих, образовав пять групп с равными интервалами. Каждую группу охарактеризовать: числом рабочих; средним стажем работы; месячной выработкой продукции – всего и в среднем на одного рабочего;

– **комбинационную группировку** по двум признакам: стажу работы и месячной выработке на одного рабочего.

Решение

Структурная группировка

Величина интервала группировочного признака (стажа работы) при изменении стажа работы от одного года до 16 лет и заданном числе групп – 5 определяется следующим образом

$$\Delta = \frac{x_{\max} - x_{\min}}{n} = \frac{16-1}{5} = 3 \text{ года},$$

где $x_{\max}, x_{\min}$ – максимальное и минимальное значение признака ряда распределения;

n – число групп.

Следовательно, первая группа рабочих имеет стаж 1–4 года; вторая 4–7 и т.д. По каждой группе определяется численность рабочих.

Ряд распределения характеризуется численностью единиц изучаемого признака, для наглядности численность рабочих, может быть представлена в процентах (табл. 1.3).

Таблица 1.3

Распределение рабочих по стажу работы

Номер группы рабочих	Группы рабочих по стажу, лет	Число рабочих, чел.	Удельный вес рабочих в группе, %
1	1–4	7	23,33
2	4–7	10	33,34
3	7–10	6	20,00
4	10–13	4	13,33
5	13–16	3	10,00
Итого:		30	100,00

Результаты группировки показывают, что более половины рабочих, т.е. 53,34% имеют стаж работы от 4 до 10 лет. Равное число рабочих имеет стаж работы до 4 лет и свыше 10 лет, удельный вес которых равен 23,3%.

Простая аналитическая группировка

Применяя метод группировок для анализа наличия взаимосвязей признаков, необходимо прежде всего определить факторный признак,

оказывающий влияние на взаимосвязанные с ним признаки. В данной задаче таким признаком является стаж работы, который должен быть положен в основание группировки. По условию задачи требуется выделить пять групп рабочих по стажу с равными интервалами. В основание аналитической группировки используются такие же группы, как и в табл. 1.3.

Для оформления результатов группировки предварительно составляется макет итоговой таблицы (табл. 1.4), который после расчета заполняется сводными групповыми показателями.

Таблица 1.4

Макет итоговой таблицы группировки рабочих по стажу работы

Номер группы рабочих	Группы рабочих по стажу, лет	Число рабочих, чел.	Средний стаж работы, лет	Дневная выработка, тыс. р.	
				всего	на одного рабочего
1	1–4				
2	4–7				
3	7–10				
4	10–13				
5	13–16				
Итого:					

Для заполнения табл. 1.4 составляется рабочая (разработочная) табл. 1.5.

Таблица 1.5

Рабочая таблица группировки рабочих по стажу

Номер группы рабочих	Группы рабочих по стажу, лет	Номер рабочего	Число рабочих	Стаж работы, лет	Дневная выработка, тыс. р.
1	1– 4	1	7	1,0	2,00
		2		1,0	2,02
		3		3,0	2,05
		8		2,5	2,30
		9		2,7	2,23
		17		1,0	2,34
		25		2,5	2,40
Итого по группе:			**7**	**13,7**	**15,34**

Окончание табл. 1.5

Номер группы рабочих	Группы рабочих по стажу, лет	Номер рабочего	Число рабочих	Стаж работы, лет	Дневная выработка, тыс. р.
2	4 – 7	4	10	6,5	2,90
		6		4,4	2.50
		7		6,9	2.80
		15		4.5	2,22
		20		6,5	2,52
		21		5,0	2,41
		22		6,0	2,56
		24		5,5	2,45
		26		5,0	2,44
		27		5,3	2,52
Итого по группе:			**10**	**55,6**	**25,32**
3	7 – 10	5	6	9,2	2,93
		18		9,0	2,70
		19		9,0	2,64
		28		7,5	2,53
		29		7,0	2,52
		30		8,0	2,62
Итого по группе:			**6**	**49,7**	**15,94**
4	10 – 13	13	4	11,0	2,95
		14		12,0	2,79
		16		10,5	2,76
		23		10,1	2,62
Итого по группе:			**4**	**43,6**	**11,12**
5	13 – 16	10	3	16,0	3,10
		11		13,2	2,84
		12		14,0	3,20
Итого по группе:			**3**	**43,2**	**9,14**
Всего по совокупности:			**30**	**205,8**	**76,86**

Групповые показатели табл. 1.5 и вычисленные на их основе средние показатели заносятся в соответствующие графы макета (табл. 1.4). Результаты простой аналитической группировки представлены в табл. 1.6.

Таблица 1.6

Группировка рабочих по стажу работы

Номер группы рабочих	Группы рабочих по стажу, лет	Число рабочих, чел.	Средний стаж работы, лет	Дневная выработка, тыс. р.	
				всего	на одного рабочего
1	1–4	7	1,96	15,34	2,191
2	4–7	10	5,56	25,32	2,532
3	7–10	6	8,28	15,94	2,657
4	10–13	4	10,90	11,12	2,780
5	13–16	3	14,40	9,14	3,047
Итого:		**30**	**6,86**	**76,86**	**2,562**

Из сравнения данных о среднем стаже и выработке на одного рабочего табл. 1.6 видно, что с увеличением стажа работы растет дневная выработка продукции одного рабочего. Следовательно, между изучаемыми признаками (показателями) имеется прямая зависимость.

Для полноты анализа результативные показатели по каждой группе сравниваются с показателями первой группы (рассчитываются базисные абсолютные и относительные приросты). Результаты представлены в табл. 1.7.

Таблица 1.7

Зависимость выработки продукции от стажа работы

Номер группы	Группы рабочих по стажу, лет	Число рабочих	Базисный прирост дневной выработки продукции на одного рабочего	
			абсолютный, тыс. р.	относительный, %
1	1 – 4	7	–	–
2	4 – 7	10	0,341	15,6
3	7 –10	6	0,466	21,3
4	10 –13	4	0,589	26,9
5	13 –16	3	0,856	39,1
Итого:		**30**	–	–

Данные табл. 1.7 показывают тенденцию роста выработки продукции в зависимости от стажа работы. С ростом стажа постепенно увеличивается прирост продукции. Рабочие пятой группы наиболее квалифицированные произвели продукции в среднем на 0,856 тыс.р., или на 39,1% больше по сравнению с рабочими первой группы. Следовательно, подтверждается вывод, сделанный ранее, что между изучаемыми признаками имеется прямая связь.

Комбинационная группировка

Чтобы произвести группировку по двум признакам (комбинационную группировку) – по стажу работы и средней дневной выработке продукции, необходимо в каждой группе рабочих по стажу выделить подгруппы по второму признаку – средней дневной выработке продукции на одного рабочего и охарактеризовать подгруппы требуемыми признаками.

Результаты комбинационной группировки при выделении пяти групп по стажу и трех подгрупп по выработке представлены в табл. 1.8.

Таблица 1.8

Комбинационная группировка рабочих по стажу и средней дневной выработке продукции

Номер группы	Группы рабочих		Число рабочих	Средний стаж работы	Средняя дневная выработка продукции, тыс. р.	
	по стажу	по средней дневной выработке, тыс. р.			всего	на одного рабочего
1	1–4	2,0–2,4	7	1,96	15,34	2,191
		2,4–2,8	–	–	–	–
		2,8–3,2	–	–	–	–
Итого:			**7**	**1,96**	**15,34**	**2,191**
2	4–7	2,0–2,4	1	4,50	2,22	2,220
		2,4–2,8	8	5,58	20,20	2,525
		2,8–3,2	1	6,50	2,90	2,900
Итого:			**10**	**5,56**	**25,32**	**2,532**

Окончание табл. 1.8

Номер группы	Группы рабочих		Число рабочих	Средний стаж работы	Средняя дневная выработка продукции, тыс. р.	
	по стажу	по средней дневной выработке, тыс. р.			всего	на одного рабочего
3	7–10	2,0–2,4	–	–	–	–
		2,4–2,8	5	8,10	13,01	2,602
		2,8–3,2	1	9,20	2,93	2,930
Итого:			**6**	**8,28**	**15,94**	**2,657**
4	10–13	2,0–2,4	–	–	–	–
		2,4–2,8	3	10,87	8,17	2,723
		2,8–3,2	1	11,00	2,95	2,950
Итого:			**4**	**10,90**	**11,12**	**2,780**
5	13–16	2,0–2,4	–	–	–	–
		2,4–2,8	–	–	–	–
		2,8–3,2	3	14,40	9,14	3,047
Итого:			**3**	**14,40**	**9,14**	**3,047**
Итого по подгруппам		**2,0–2,4**	**8**	**2,28**	**17,56**	**2,195**
		2,4–2,8	**16**	**7,36**	**41,38**	**2,586**
		2,8–3,2	**6**	**11,65**	**17,92**	**2,987**
Всего по совокупности:			**30**	**6,86**	**76,86**	**2,562**

Анализ данных табл. 1.8 показывает, что выработка продукции рабочих находится в прямой зависимости от стажа работы.

2. ПОКАЗАТЕЛИ ВАРИАЦИИ

2.1 Методические указания к решению задач по теме «Показатели вариации»

Для измерения степени варьирования (колеблемости) признака служит вариация, показателями которой являются: размах вариации, среднее линейное отклонение, среднее квадратическое отклонение, средний квадрат отклонений (дисперсия), коэффициент вариации.

Размах вариации

Размах вариации (R) характеризует пределы вариации (изменения) индивидуальных значений (или вариантов) признака (x) в статистической совокупности

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

где $x_{\max}$, $x_{\min}$ – наибольшее и наименьшее значение признака.

Среднее линейное отклонение

Среднее линейное отклонение вычисляется по формулам средней арифметической:

– простой (невзвешенной)

$$\overline{d} = \frac{\sum |x_i - \overline{x}|}{n},$$

где x_i – i-е значение признака x;

$\overline{x}$ – средняя величина признака x;

f_i – статистический вес i-го значения признака;

n – число членов совокупности;

– взвешенной

$$\overline{d} = \frac{\sum | x_i - \overline{x} | f_i}{\sum f_i}.$$

Среднее квадратическое отклонение

Среднее квадратическое отклонение рассчитывается по формулам:
– невзвешенной

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}};$$

– взвешенной

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}.$$

Дисперсия количественного признака

Дисперсия количественного признака определяется по формулам средней арифметической:
– невзвешенной

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n};$$

– взвешенной

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}.$$

Дисперсия может быть рассчитана следующим образом:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} = \frac{\sum x_i^2 f_i}{\sum f_i} - 2\bar{x}\frac{\sum x_i f_i}{\sum f_i} + (\bar{x})^2 =$$

$$= \overline{x^2} - 2(\bar{x})^2 + (\bar{x})^2 = \overline{x^2} - (\bar{x})^2,$$

где $\overline{x^2}$ – средний квадрат значений признака;

$(\bar{x})^2$ – квадрат средней величины признака.

Дисперсии количественного признака в совокупности, разделенной на группы

Для анализа связей количественных признаков в статистической совокупности, разделенной на группы, рассчитываются следующие дисперсии: групповая, межгрупповая, внутригрупповая и общая.

Групповая дисперсия (частная) характеризует вариацию признака в группе, обусловленную действием на него всех прочих факторов, кроме признака, положенного в основание группировки (группировочного признака):

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 f_{ij}}{\sum_{i=1}^{n_j} f_{ij}},$$

где x_{ij} – i-е значение признака в j-й группе;

$\bar{x}_j$ – частная (групповая) средняя величина признака в j-й группе;

f_{ij} – статистический вес i-го значения признака в j-й группе;

n_j – число различных значений признака в j-й группе.

Межгрупповая дисперсия измеряет степень колеблемости (вариацию) признака во всей статистической совокупности за счет фактора, положенного в основание группировки (группировочного признака):

$$\delta^2 = \frac{\sum_{j=1}^{J} \left(\bar{x}_j - \bar{x}\right)^2 F_j}{\sum_{j=1}^{J} F_j},$$

где $\bar{x}$ – среднее значение признака в совокупности (общая средняя);

F_j – вес j-й группы, представляющий собой численность единиц в j-й группе;

J – количество групп в статистической совокупности.

Внутригрупповая дисперсия (средняя групповых дисперсий) измеряет степень колеблемости признака во всей совокупности в целом за счет действия на него всех прочих факторов (признаков), кроме группировочного признака:

$$\overline{\sigma^2} = \frac{\sum_{j}^{J} \sigma_j^2 F_j}{\sum_{j}^{J} F_j}.$$

Общая дисперсия измеряет степень колеблемости признака, за счет влияния всех действующих на него факторов:

$$\sigma^2 = \overline{\sigma^2} + \delta^2.$$

Общая дисперсия признака в статистической совокупности, разделенной на группы, может быть определена по основной формуле дисперсии

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}.$$

Межгрупповая и общая дисперсии применяются для определения показателей тесноты связи показателей в совокупности, разделенной на группы.

Коэффициент вариации

Коэффициент вариации вычисляется по формуле

$$v = \frac{\sigma}{\bar{x}},$$

где σ – среднее квадратическое отклонение;

$\bar{x}$ – средняя величина признака.

Коэффициент вариации выражается обычно в процентах и дает представление о степени однородности статистической совокупности. Если коэффициент меньше 25–30%, то статистическую совокупность по изучаемому признаку можно считать однородной.

«Показатели вариации»

Задача 2.1

Имеются данные об индивидуальной производительности труда рабочих в двух бригадах (табл. 2.1).

Таблица 2.1

Производительность труда рабочих

Номер рабочего	Производство продукции за смену, шт.	
	в первой бригаде	во второй бригаде
1	20	80
2	30	90
3	120	100
4	150	110
5	180	120
Итого	500	500

Определить среднюю производительность труда в бригадах, размах вариации, среднее линейное отклонение.

Решение

1. Средняя производительность труда в двух бригадах одинаковая

$$\bar{x}_1 = \bar{x}_2 = \frac{500}{5} = 100 \text{ шт.}$$

2. Размах вариации производительности труда

– в первой бригаде

$$R_1 = x_{\max 1} - x_{\min 1} = 180 - 20 = 160 \text{ шт.};$$

– во второй бригаде

$$R_2 = x_{\max 2} - x_{\min 2} = 120 - 80 = 40 \text{ шт},$$

где $x_{\max}$, $x_{\min}$ – наибольшее и наименьшее значение признака.

3. Среднее линейное отклонение вычисляется по формуле средней арифметической простой (невзвешенной)

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n},$$

где x_i – i-е значение признака x;

$\bar{x}$ – средняя величина признака x;

n – число членов совокупности.

Расчет индивидуальных линейных отклонений производительности труда приведен в табл. 2.2.

Таблица 2.2

Линейное отклонение производительности труда

Номер рабочего	Первая бригада			Вторая бригада		
	x_{i1}, шт.	$x_{i1} - \bar{x}_{i1}$, шт.	$\lvert x_{i1} - \bar{x}_{i1} \rvert$, шт.	x_{i2}, шт.	$x_{i2} - \bar{x}_{i2}$, шт.	$\lvert x_{i2} - \bar{x}_{i2} \rvert$, шт.
1	20	–80	80	80	–20	20
2	30	–70	70	90	–10	10
3	120	20	20	100	0	0
4	150	50	50	110	10	10
5	180	80	80	120	20	20
Итого	500	0	300	500	0	60

Среднее линейное отклонение

– в первой бригаде

$$\overline{d}_1 = \frac{\sum |x_{i1} - \bar{x}_1|}{n} = \frac{300}{5} = 60 \text{ шт.};$$

– во второй бригаде

$$\overline{d}_2 = \frac{\sum |x_{i2} - \bar{x}_2|}{n} = \frac{60}{5} = 12 \text{ шт.}$$

Величина размаха вариации и среднего линейного отклонения указывает на то, что первая бригада является более неоднородной по производительности труда по сравнению со второй бригадой.

Задача 2.2

Распределение рабочих по тарифным разрядам приведено в графах 1 и 2 табл. 3.3.

Таблица 2.3

Распределение рабочих по тарифным разрядам

Тарифный разряд, x_i	Число рабочих, f_i	$x_i f_i$	$(x_i - \bar{x})$	$(x_i - \bar{x}) f_i$	$(x_i - \bar{x})^2 f_i$
1	2	3	4	5	6
2	1	2	–2,5	–2,5	6,25
3	2	6	–1,5	–3,0	4,50
4	6	24	–0,5	–3,0	1,50
5	8	40	0,5	4,0	2,00
6	3	18	1,5	4,5	6,75
Итого	20	90	–	0	21,00

Определить степень однородности совокупности рабочих по тарифному разряду.

Решение

1. Степень однородности совокупности по изучаемому показателю можно оценить с помощью коэффициента вариации

$$v = \frac{\sigma}{\bar{x}},$$

где σ – среднее квадратическое отклонение тарифного разряда;

$\bar{x}$ – средний тарифный разряд.

2. Расчет среднего значения, дисперсии и среднего квадратического отклонения тарифного разряда производится по следующим формулам:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}; \quad \sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}; \quad \sigma = \sqrt{\sigma^2},$$

где x_i – i-й тарифный разряд;

f_i – число рабочих, имеющих i-й тарифный разряд.

Расчет числителей средней величины и дисперсии приведены в графах 3–6 табл. 3.3.

3. Средний тарифный разряд рабочих

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{90}{20} = 4{,}5 \text{ разряд};$$

дисперсия тарифного разряда

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} = \frac{21}{20} = 1{,}05;$$

среднее квадратическое отклонение тарифного разряда

$$\sigma = \sqrt{\sigma^2} = \sqrt{1{,}05} = 1{,}025 \text{ разряда};$$

коэффициент вариации тарифного разряда

$$v = \frac{\sigma}{\bar{x}} = \frac{1{,}025}{4{,}5} = 0{,}227\ (22{,}7\%).$$

При коэффициенте вариации 22,7% совокупность рабочих можно считать однородной по тарифному разряду.

Задача 2.3

Распределение рабочих по заработной плате приведено в графах 1 и 2 табл. 2.4.

Таблица 2.4

Распределение рабочих по заработной плате

Заработная плата, тыс. р.	Число рабочих, f_i, чел.	Середина интервала, x_i, тыс. р.	$\frac{x_i - A}{h}$	$\left(\frac{x_i - A}{h}\right) f_i$	$\left(\frac{x_i - A}{h}\right)^2 f_i$
1	2	3	4	5	6
До 3,0	2	2	–3	–6	18
3,0 – 5,0	12	4	–2	–24	48
5,0 – 7,0	15	6	–1	–15	15
7,0 – 9,0	64	8	0	0	0
9,0 – 11,0	55	10	1	55	55
11,0 – 13,0	32	12	2	64	128
Свыше 13,0	20	14	3	60	180
Итого	200	–	–	134	444

Определить дисперсию и среднее квадратическое отклонение заработной платы рабочих способом моментов.

Решение

При использовании способа моментов расчет дисперсии осуществляется по формуле

$$\sigma^2 = h^2(m_2 - m_1^2),$$

где m_1, m_2 – моменты первого и второго порядка

$$m_1 = \frac{\sum\left(\frac{x-A}{h}\right)f}{\sum f}; \quad m_2 = \frac{\sum\left(\frac{x-A}{h}\right)^2 f}{\sum f}.$$

Расчет числителей моментов при A = 8 тыс. р. (значение заработной платы, имеющее наибольшую частоту) и h = 2 тыс. р. (величина интервала группировок) приведен в графах 4–6 табл. 3.4.

Момент первого порядка

$$m_1 = \frac{\sum\left(\frac{x-A}{h}\right)f}{\sum f} = \frac{134}{200} = 0{,}67;$$

момент второго порядка

$$m_2 = \frac{\sum\left(\frac{x-A}{h}\right)^2 f}{\sum f} = \frac{444}{200} = 2{,}22;$$

дисперсия заработной платы

$$\sigma^2 = h^2(m_2 - m_1^2) = 2^2(2{,}22 - 0{,}67^2) = 7{,}0844;$$

среднее квадратическое отклонение заработной платы

$$\sigma = \sqrt{\sigma^2} = \sqrt{7{,}0844} = 2{,}66 \text{ тыс. р.}$$

2.2 Методические указания к решению задач по теме «Дисперсионный анализ связей социально-экономических явлений»

Для анализа связей признаков в статистической совокупности, разбитой на группы, рассчитываются следующие дисперсии: групповая, межгрупповая, внутригрупповая и общая.

Групповая дисперсия (частная) характеризует изменение результативного признака в группе, обусловленную действием на него всех прочих факторов, кроме признака, положенного в основание группировки (группировочного признака)

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 f_{ij}}{\sum_{i=1}^{n_j} f_{ij}},$$

где x_{ij} – i-е значение признака в j-й группе;

$\bar{x}_j$ – частная (групповая) средняя величина признака в j-й группе;

f_{ij} – статистический вес i-го значения признака в j-й группе;

n_j – число различных значений признака в j-й группе.

Межгрупповая дисперсия измеряет степень колеблемости (вариацию) результативного признака во всей статистической совокупности за счет изменения группировочного признака

$$\delta^2 = \frac{\sum_{j=1}^{J} (\bar{x}_j - \bar{x})^2 F_j}{\sum_{j=1}^{J} F_j},$$

где $\bar{x}$ – среднее значение признака в совокупности (общая средняя);

F_j – вес j-й группы, представляющий собой численность единиц в j-й группе;

J – количество групп в статистической совокупности.

Внутригрупповая дисперсия измеряет степень колеблемости результативного признака во всей совокупности в целом за счет действия на него всех прочих факторов (признаков) кроме группировочного признака:

$$\overline{\sigma^2} = \frac{\sum_j^J \sigma_j^2 F_j}{\sum_j^J F_j},$$

где σ_j^2 – дисперсия признака в j-й группе.

Общая дисперсия измеряет степень колеблемости результативного признака за счет воздействия на него всех факторов:

$$\sigma^2 = \overline{\sigma^2} + \delta^2.$$

Коэффициент детерминации показывает, какую часть общей вариации изучаемого признака составляет межгрупповая вариация, т.е. обусловленная группировочным признаком. Определяется следующим образом:

$$\eta^2 = \frac{\delta^2}{\sigma^2}.$$

Коэффициент детерминации лежит в пределах

$$0 \le \eta^2 \le 1.$$

Чем ближе коэффициент детерминации к 1, тем больше степень влияния изучаемого факторного признака на результативный.

Эмпирическое корреляционное отношение характеризует степень тесноты связи между факторным (группировочным) и результативным признаками

$$\eta = \pm\sqrt{\frac{\delta^2}{\sigma^2}},$$

где σ^2 – общая дисперсия результативного признака.

Направление связи (прямая, обратная) устанавливается в зависимости от исходной статистической информации.

Эмпирическое корреляционное отношение лежит в пределах

$$-1 \le \eta \le 1.$$

Чем ближе к 1 абсолютное значение эмпирического коэффициента корреляции, тем теснее связь:

При $|\eta| = 0 - 0{,}4$ – связь слабая;

$|\eta| = 0{,}4 - 0{,}7$ – связь средняя (умеренная);

$|\eta| = 0{,}7 - 1$ – связь сильная.

Примеры решения задач по теме «Дисперсионный анализ связей социально-экономических явлений»

Задача 2.4

Имеются данные о стаже и средней часовой выработке рабочих (табл. 2.5).

Таблица 2.5

Стаж и выработка рабочих

Номер группы рабочих	Стаж работы, лет	Среднечасовая выработка продукции одного рабочего, x_{ij}, шт./чел.	Число рабочих, f_{ij},чел
1	До 3	2 3 4	1 4 3
2	3 – 5	3 4 5 6	1 2 6 3
3	5 – 7	6 7 8	1 2 2

Определить взаимосвязь между стажем работы и производительностью труда (часовой выработкой) методом дисперсионного анализа.

Методика статистического исследования

Для решения задач анализа статистической информации необходимо разработать методику статистического исследования. Построение методики начинается с последнего действия задачи. В дисперсионном анализе связей социально-экономических явлений вычисляются следующие показатели:

1. Эмпирическое корреляционное отношение $\eta = \pm\sqrt{\frac{\delta^2}{\sigma^2}}$.

2. Коэффициент детерминации $\eta^2 = \frac{\delta^2}{\sigma^2}$.

3. Общая дисперсия $\sigma^2 = \overline{\sigma^2} + \delta^2$.

4. Межгрупповая дисперсия $$\delta^2 = \frac{\sum_{j=1}^{J}\left(\bar{x}_j - \bar{x}\right)^2 F_j}{\sum_{j=1}^{J} F_j}.$$

5. Внутригрупповая дисперсия $$\overline{\sigma^2} = \frac{\sum \sigma_j^2 F_j}{\sum F_j}.$$

6. Групповая дисперсия $$\sigma_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 f_{ij}}{\sum_{i=1}^{n_j} f_{ij}}.$$

7. Общая средняя $$\bar{x} = \frac{\sum_{j=1}^{J} \bar{x}_j F_j}{\sum_{j=1}^{J} F_j}.$$

8. Групповая средняя $$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij} f_{ij}}{\sum_{i=1}^{n_j} f_{ij}}.$$

Методика статистического исследования представляет собой алгоритм решения задачи с примерами расчета показателей. Методика дисперсионного анализа приведена в решении данной задачи.

Решение

1. Групповая средняя. Средняя часовая выработка рабочего:

– в первой группе

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{i1} f_{i1}}{\sum_{i=1}^{n_1} f_{i1}} = \frac{2\cdot 1 + 3\cdot 4 + 4\cdot 3}{1+4+3} = \frac{26}{8} = 3{,}25 \text{ шт./чел.};$$

– во второй группе

$$\bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{i2} f_{i2}}{\sum_{i=1}^{n_1} f_{i2}} = \frac{3\cdot 1 + 4\cdot 2 + 5\cdot 6 + 6\cdot 3}{1+2+6+3} = \frac{59}{12} = 4{,}92 \text{ шт./чел.};$$

– в третьей группе

$$\bar{x}_3 = \frac{\sum_{i=1}^{n_3} x_{i3} f_{i3}}{\sum_{i=1}^{n_3} f_{i3}} = \frac{6 \cdot 1 + 7 \cdot 2 + 8 \cdot 2}{1 + 2 + 2} = \frac{36}{5} = 7{,}2 \text{ шт./чел.}$$

2. Общая средняя. Средняя часовая выработка рабочего во всей совокупности может быть рассчитана по формулам:

– средней арифметической взвешенной

$$\bar{x} = \frac{\sum_{j=1}^{J} \bar{x}_j F_j}{\sum_{j=1}^{J} F_j} = \frac{3{,}25 \cdot 8 + 4{,}92 \cdot 12 + 7{,}2 \cdot 5}{8 + 12 + 5} = \frac{121}{25} = 4{,}84 \text{ шт./чел.;}$$

– средней агрегатной

$$\bar{x} = \frac{\sum_{j=1}^{J} Q_j}{\sum_{j=1}^{J} F_j} = \frac{26 + 59 + 36}{8 + 12 + 5} = \frac{121}{25} = 4{,}84 \text{ шт./чел.,}$$

где Q_j – часовой объем выпуска продукции в *j*-й группе.

3. Групповые дисперсии определяются по формуле

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 f_{ij}}{\sum_{i=1}^{n_j} f_{ij}}.$$

Расчет суммы квадрата отклонений индивидуальных выработок от средней групповой выработки (числителя групповой дисперсии) по группам приведен в табл. 2.2-2.4.

Таблица 2.6

Расчет дисперсии выработки в первой группе

Выработка рабочих в первой группе, x_{i1}, шт./чел.	Число рабочих в первой группе, f_{i1},чел	$x_{i1} - \bar{x}_1 = x_{i1} - 3{,}25$	$(x_{i1} - \bar{x}_1)^2 = (x_{i1} - 3{,}25)^2$	$(x_{i1} - \bar{x}_1)^2 f_{i1} = (x_{i1} - 3{,}25)^2 f_{i1}$
2	1	–1,25	1,5625	1,5625
3	4	–0,25	0,0625	0,2500
4	3	0,75	0,5625	1,6875
Итого	8	–	–	3,5000

Дисперсии выработки в первой группе $\sigma_1^2 = \frac{3,5}{8} = 0,4375.$

Таблица 2.7

Расчет дисперсии выработки во второй группе

Выработка рабочих во второй группе, x_{i2}, шт./чел.	Число рабочих во второй группе, f_{i2},чел	$x_{i2} - \bar{x}_2 = = x_{i2} - 4,92$	$(x_{i2} - \bar{x}_2)^2 = = (x_{i2} - 4,92)^2$	$(x_{i2} - \bar{x}_2)^2 f_{i2} = = (x_{i2} - 4,92)^2 f_{i2}$
3	1	–1,92	3,6864	3,6864
4	2	–0,92	0,8464	1,6928
5	6	0,08	0,0064	0,0384
6	3	1,08	1,1664	3,4992
Итого	12	–	–	8,9168

Дисперсии выработки во второй группе $\sigma_2^2 = \frac{8,9168}{12} = 0,74307.$

Таблица 2.8

Расчет дисперсии выработки в третьей группе

Выработка рабочих в третьей группе, x_{i3}, шт./чел.	Число рабочих в третьей группе, f_{i3},чел	$x_{i3} - \bar{x}_3 = = x_{i3} - 7,2$	$(x_{i3} - \bar{x}_3)^2 = = (x_{i3} - 7,2)^2$	$(x_{i3} - \bar{x}_3)^2 f_{i3} = = (x_{i3} - 7,2)^2 f_{i3}$
6	1	–1,2	1,44	1,44
7	2	–0,2	0,04	0,08
8	2	0,8	0,64	1,28
Итого	5	–	–	2,80

Дисперсии выработки в третьей группе $\sigma_3^2 = \frac{2,8}{5} = 0,56.$

4. Внутригрупповая дисперсия выработки (средняя из групповых дисперсий)

$$\overline{\sigma^2} = \frac{\sum \sigma_j^2 F_j}{\sum F_j} = \frac{0,4375 \cdot 8 + 0,74307 \cdot 12 + 0,56 \cdot 5}{8 + 12 + 5} = 0,60867.$$

5. Межгрупповая дисперсия выработки (дисперсия средних групповых)

$$\delta^2 = \frac{\sum_{j=1}^{J}\left(\bar{x}_j - \bar{x}\right)^2 F_j}{\sum_{j=1}^{J} F_j} = \frac{(3{,}25-4{,}84)^2 \cdot 8 + (4{,}92-4{,}84)^2 \cdot 12 + (7{,}2-4{,}84)^2 \cdot 5}{8+12+5} = 1{,}92598.$$

6. Общая дисперсия выработки

$$\sigma^2 = \overline{\sigma^2} + \delta^2 = 0{,}60867 + 1{,}92598 = 2{,}53465.$$

Общая дисперсия выработки может быть рассчитана по основной формуле дисперсии (табл. 2.5)

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}.$$

Таблица 2.9

Расчет общей дисперсии выработки

Выработка рабочих во второй группе, x_i, шт./чел.	Число рабочих во второй группе, f_i,чел	$x_i - \bar{x} = x_i - 4{,}84$	$(x_i - \bar{x})^2 = (x_i - 4{,}84)^2$	$(x_i - \bar{x})^2 f_i = (x_i - 4{,}84)^2 f_i$
2	1	–2,84	8,0656	8,0656
3	5	–1,84	3,3856	16,9280
4	5	–0,84	0,7056	3,5280
5	6	0,16	0,0256	0,3360
6	4	1,16	1,3456	5,3824
7	2	2,16	4,6656	9,3312
8	2	3,16	9,9856	19,9712
Итого	25	–	–	63,5424

Общая дисперсии выработки $\sigma^2 = \frac{63{,}5424}{25} = 2{,}541696.$

Отличие общей дисперсии, рассчитанной разными способами , объясняется неточностью округления значений в табл. 2.5.

7. Коэффициент детерминации

$$\eta^2 = \frac{\delta^2}{\sigma^2} = \frac{1{,}92598}{2{,}53465} = 0{,}75986 \ \ (76\%).$$

8. Эмпирическое корреляционное отношение

$$\eta = \pm\sqrt{\frac{\delta^2}{\sigma^2}} = +\sqrt{0{,}76} = +0{,}87.$$

Изменение стажа влияет на изменение выработки на 76%, между стажем работы и производительностью труда в данном статистическом исследовании существует тесная положительная связь.

Задача 2.2

Имеются данные о капитале коммерческих банков (табл. 2.10).

Таблица 2.10

Капитал коммерческих банков

Номер группы банков	Собственный капитал, млн р.	Удельный вес банков d_j, %	Средняя величина привлеченных средств $\bar{x}_j$, млн р.	Дисперсия привлеченных средств σ_j^2
1	200 – 300	10	500	2500
2	300 – 400	30	600	3600
3	400 – 500	50	800	4900
4	500 – 600	10	1200	8100
Итого	–	**100**	–	–

Определить показатели тесноты связи между величиной собственных средств и привлеченными капиталами.

Решение

1. Средний размер привлеченных средств по всей совокупности банков

$$\bar{x} = \frac{\sum_{j=1}^{J} \bar{x}_j d_j}{\sum_{j=1}^{J} d_j} = \frac{500 \cdot 10 + 600 \cdot 30 + 800 \cdot 50 + 1200 \cdot 10}{100} = 750 \text{ млн р.}$$

2. Межгрупповая дисперсия

$$\delta^2 = \frac{\sum_{j=1}^{J}\left(\bar{x}_j - \bar{x}\right)^2 d_j}{\sum_{j=1}^{J} d_j} =$$

$$= \frac{(500-750)^2 \cdot 10 + (600-750)^2 \cdot 30 + (800-750)^2 \cdot 50 + (1200-750)^2 \cdot 10}{100} = 34500.$$

3. Внутригрупповая дисперсия

$$\overline{\sigma^2} = \frac{\sum_{j=1}^{J}\sigma_j^2 d_j}{\sum_{j=1}^{J} d_j} = \frac{2500 \cdot 10 + 3600 \cdot 30 + 4900 \cdot 50 + 8100 \cdot 10}{100} = 4590.$$

4. Общая дисперсия

$$\sigma^2 = \overline{\sigma^2} + \delta^2 = 4590 + 34500 = 39090.$$

5. Коэффициент детерминации

$$\eta^2 = \frac{\delta^2}{\sigma^2} = \frac{34500}{39090} = 0{,}8826.$$

6. Эмпирическое корреляционное отношение

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{0{,}8826} = 0{,}94.$$

Изменение величины собственного капитала влияет на изменение размера привлеченных средств на 88,26%. Между показателями существует тесная прямая связь.

3. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ СВЯЗЕЙ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ

3.1. Методические указания к решению задач по теме «Корреляционно-регрессионный анализ связей социально-экономических явлений»

Одним из методов изучения корреляционных связей является корреляционно-регрессионный анализ, позволяющий определить степень тесноты и форму связи между признаками.

Анализ формы корреляционной связи количественных показателей

Парная корреляция

Линейное уравнение парной регрессии.

Если с увеличением факторного признака результативный признак равномерно возрастает или убывает, то такая зависимость является линейной и выражается уравнением прямой

$$y_x = a_0 + a_1 x = f(x, a_0, a_1),$$

где y_x – теоретическая зависимость (уравнение регрессии) результативного признака от факторного;

x – факторный признак;

a_0, a_1 – параметры уравнения прямой.

Параметры уравнения прямой a_0 и a_1 определяются путем решения системы уравнений, полученных с использованием метода наименьших квадратов:

$$\begin{cases} \sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} x_i; \\ \sum_{i=1}^{n} x_i y_i = a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2, \end{cases}$$

где x_i, y_i – индивидуальные значения факторного и результативного признаков;

n – число индивидуальных значений признака.

Коэффициент уравнения регрессии a_1 показывает изменение средней величины результативного признака при изменении факторного признака на единицу.

Параболическое уравнение регрессии.

Если связь между признаками нелинейная и с возрастанием факторного признака происходит ускоренное возрастание или убывание результативного признака, то корреляционная зависимость может быть выражена параболой

$$y_x = a_0 + a_1 x + a_2 x^2.$$

Значения параметров параболы a_0, a_1, a_2 определяются из решения системы нормальных уравнений

$$\begin{cases} \sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2; \\ \sum_{i=1}^{n} x_i y_i = a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3; \\ \sum_{i=1}^{n} y_i x_i^2 = a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4. \end{cases}$$

Гиперболическое уравнение регрессии.

Если результативный признак с увеличением факторного признака возрастает (или убывает) не бесконечно, а стремится к конечному пределу, то для анализа такого признака применяется уравнение гиперболы

$$y_x = a_0 + a_1 1/x.$$

Для определения параметров этого уравнения используется система нормальных уравнений

$$\begin{cases} \sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} \frac{1}{x_i}; \\ \sum_{i=1}^{n} \frac{1}{x_i} y_i = a_0 \sum_{i=1}^{n} \frac{1}{x_i} + a_1 \sum_{i=1}^{n} (\frac{1}{x_i})^2. \end{cases}$$

Степенное уравнение регрессии.

Если связь между признаками слабая нелинейная, то для характеристики этой связи в экономических исследованиях применяется степенная функция

$$y_x = a_0 x^{a_1}.$$

Для определения параметров производится логарифмирование степенной функции

$$\lg y = \lg a_0 + a_1 \lg x.$$

и строится система нормальных уравнений с использованием метода наименьших квадратов

$$\begin{cases} \sum_{i=1}^{n} \lg y_i = n \lg a_0 + a_1 \sum_{i=1}^{n} \lg x_i; \\ \sum_{i=1}^{n} \lg y_i \lg x_i = \lg a_0 \sum_{i=1}^{n} \lg x_i + a_1 \sum_{i=1}^{n} (\lg x_i)^2. \end{cases}$$

Показательное уравнение регрессии.

При статистическом анализе нелинейной корреляционной связи возможно применение уравнения регрессии в виде показательной функции

$$y_x = a_0 a_1^x.$$

Для решения уравнения производится его логарифмирование

$$\lg y = \lg a_0 + x \lg a_1.$$

С учетом требований метода наименьших квадратов составляется система нормальных уравнений

$$\begin{cases} \sum_{i=1}^{n} \lg y_i = n \lg a_0 + \lg a_1 \sum_{i=1}^{n} x_i; \\ \sum_{i=1}^{n} x_i \lg y_i = \lg a_0 \sum_{i=1}^{n} x_i + \lg a_1 \sum_{i=1}^{n} x_i^2. \end{cases}$$

Логарифмическое уравнение регрессии.

При статистическом анализе криволинейной связи может применяться логарифмическая функция

$$y_x = a_0 + a_1 \lg x.$$

Параметры логарифмической функции определяются из системы нормальных уравнений, отвечающих требованию метода наименьших квадратов:

$$\begin{cases} \sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} \lg x_i; \\ \sum_{i=1}^{n} y_i \lg x_i = a_0 \sum_{i=1}^{n} \lg x_i + a_1 \sum_{i=1}^{n} (\lg x_i)^2. \end{cases}$$

Аналогичным образом с использованием метода наименьших квадратов определяются параметры любой формы связи между результативным и факторным признаками.

Множественная корреляция

Линейное уравнение множественной регрессии.

Статистическая модель, показывающая связь между результативным и несколькими факторными признаками, представляет собой уравнение множественной регрессии. Уравнения множественной регрессии могут быть линейными и нелинейными.

Наиболее простым видом уравнения множественной регрессии является линейное уравнение с двумя независимыми переменными

$$y_{x_1x_2} = a_0 + a_1x_1 + a_2x_2.$$

Параметры этого уравнения определяются решением системы нормальных уравнений, составленных в результате применения метода наименьших квадратов:

$$\begin{cases} na_0 + a_1\sum\limits_{i=1}^{n} x_{1i} + a_2\sum\limits_{i=1}^{n} x_{2i} = \sum\limits_{i=1}^{n} y_i; \\ a_0\sum\limits_{i=1}^{n} x_{1i} + a_1\sum\limits_{i=1}^{n} x_i^2 + a_2\sum\limits_{i=1}^{n} x_{1i}x_{2i} = \sum\limits_{i=1}^{n} y_i x_{1i}; \\ a_0\sum\limits_{i=1}^{n} x_{2i} + a_1\sum\limits_{i=1}^{n} x_{1i}x_{2i} + a_2\sum\limits_{i=1}^{n} x_{2i}^2 = \sum\limits_{i=1}^{n} y_i x_{2i}. \end{cases}$$

В общем виде линейная регрессия $\hat{y}_i$ с m независимыми переменными имеет вид

$$\hat{y}_i = a_0 + a_1x_1 + a_2x_2 + ... + a_jx_j + ... + a_mx_m.$$

Анализ тесноты связи количественных показателей

Парная корреляция

Парные коэффициенты корреляции и детерминации.

При линейной парной зависимости для определения тесноты связи между результативным и факторным признаками и оценки степени влияния факторного признака на результативный используются коэффициенты корреляции и детерминации.

Коэффициент корреляции

$$r_{xy} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\ \sqrt{n\sum_{i=1}^{n} y^2{}_i - (\sum_{i=1}^{n} y_i)^2}}.$$

Величина коэффициента корреляции находится в пределах от –1 до +1. Чем ближе по абсолютной величине коэффициент корреляции к 1, тем теснее связь.

Коэффициент детерминации

$$D_{xy} = r_{xy}^2$$

характеризует долю влияния факторного признака на вариацию результативного.

Парные индексы корреляции и детерминации.

При парных нелинейных зависимостях для определения тесноты связи между результативным и факторным признаками и оценки степени влияния факторного признака на результативный используются индексы корреляции и детерминации.

Индекс корреляции

$$R_{xy} = \sqrt{\frac{\sigma_{y_x}^2}{\sigma_y^2}},$$

где $\sigma_{y_x}^2$ – факторная дисперсия результативного признака y ;

σ_y^2 – общая дисперсия результативного признака.

Величина индекса корреляции находится в пределах от –1 до +1. Чем ближе по абсолютной величине индекс корреляции к 1, тем теснее связь.

Факторная дисперсия результативного признака

$$\sigma_{y_x}^2 = \frac{\sum_{i=1}^{n} (y_{x_i} - \bar{y})^2}{n},$$

где y_{x_i} – теоретические значения результативного признака при значениях факторного признака x_i ;

$\bar{y}$ – среднее значение результативного признака.

Общая дисперсия результативного признака

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n},$$

где y_i – эмпирические значения результативного признака.

Индекс детерминации

$$B_y = R_{xy}^2 = \frac{\sigma_{y_x}^2}{\sigma_y^2}$$

показывает долю факторной дисперсии в общей дисперсии, т.е. характеризует, какая часть общей вариации результативного признака y объясняется изучаемым фактором x.

Множественная корреляция

Множественные коэффициенты корреляции и детерминации.

Множественный коэффициент корреляции r_y характеризует степень тесноты линейной статистической связи результативного y и линейной комбинацией факторных $x_1, x_2, \ldots, x_m$ признаков

$$r_y = \sqrt{\frac{\sigma_{y12\ldots m}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sigma_{y(12\ldots m)}^2}{\sigma_y^2}};$$

$$\sigma_y^2 = \sigma_{y12\ldots m}^2 + \sigma_{y(12\ldots m)}^2,$$

где $\sigma_{y12\ldots m}^2$ – факторная дисперсия результативного признака, полученная с учетом факторов $x_1, x_2, \ldots, x_m$;

σ_y^2 – общая дисперсия результативного признака, полученная с учетом факторных признаков $x_1, x_2, \ldots, x_m$ и всех прочих признаков;

$\sigma_{y(12\ldots m)}^2$ – остаточная дисперсия результативного признака, полученная при элиминации (исключении) влияния факторных признаков $x_1, x_2, \ldots, x_m$.

Факторная, остаточная и общие дисперсии результативного признака определяются следующим образом :

$$\sigma^2_{y12\ldots m} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n};$$

$$\sigma^2_{y(12\ldots m)} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n};$$

$$\sigma^2_y = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n},$$

где $\hat{y}_i$, $\bar{y}$, y_i – расчетное (теоретическое), среднее и эмпирическое (опытное) значения результативного признака.

Множественный коэффициент корреляции никогда не уменьшается с расширением набора факторных признаков, относительно которых измеряется зависимость результативного признака.

Квадрат величины коэффициента множественной корреляции является коэффициентом множественной детерминации

$$B_y = r_y^2 = \frac{\sigma^2_{y12\ldots m}}{\sigma^2_y}$$

и характеризует долю влияния выбранных факторных признаков на результативный фактор.

При статистической оценке тесноты линейной связи между результативным y и двумя факторными признаками x_1, x_2 может быть использована следующая формула :

$$r_{yx_1x_2} = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1} r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}}},$$

где $r_{yx_1}, r_{yx_2}, r_{x_1x_2}$ – парные коэффициенты корреляции.

$$r_{yx_1} = \frac{n\sum_{i=1}^{n} y_i x_{1i} - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_{1i}}{\sqrt{[n\sum_{i=1}^{n} x_{1i}^2 - (\sum_{i=1}^{n} x_{1i})^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}};$$

$$r_{yx_2} = \frac{n\sum_{i=1}^{n} y_i x_{2i} - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_{2i}}{\sqrt{[n\sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{2i})^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}};$$

$$r_{x_1x_2} = \frac{n\sum_{i=1}^{n} x_{1i} x_{2i} - \sum_{i=1}^{n} x_{1i} \sum_{i=1}^{n} x_{2i}}{\sqrt{[n\sum_{i=1}^{n} x_{1i}^2 - (\sum_{i=1}^{n} x_{1i})^2][n\sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{2i})^2]}}.$$

Методика оценки тесноты связи при нелинейной модели регрессии такая же, как при линейной. Коэффициент связи при этом называется индексом множественной корреляции, а его квадрат – индексом множественной детерминации.

Частный коэффициент корреляции.

Частные коэффициенты корреляции служат для оценки вклада во множественный коэффициент корреляции каждого из факторов.

В общем случае формула для определения частного коэффициента корреляции между факторами y и x_m при исключении влияния факторов $x_1, x_2, \ldots, x_{m-1}$ имеет вид

$$R_{ym(12\ldots m-1)} = \sqrt{\frac{\sigma^2_{y12\ldots m} - \sigma^2_{y12\ldots m-1}}{\sigma^2_{y(12\ldots m-1)}}} = \sqrt{\frac{\sigma^2_{y12\ldots m} - \sigma^2_{y12\ldots m-1}}{\sigma^2_{y} - \sigma^2_{y12\ldots m-1}}},$$

где $\sigma^2_{y12\ldots m}$ – факторная дисперсия результативного признака, полученная с учетом влияния факторов $x_1, x_2, \ldots, x_m$;

$\sigma^2_{y12\ldots m-1}$ – факторная дисперсия результативного признака, полученная с учетом влияния факторов $x_1, x_2, \ldots, x_{m-1}$;

$\sigma^2_{y(12\ldots m-1)}$ – остаточная дисперсия результативного признака, полученная с учетом влияния рассматриваемого фактора x_m и прочих факторов;

σ^2_y – общая дисперсия результативного признака.

Абсолютная величина частного коэффициента корреляции лежит в пределах от 0 до 1, а знак определяется знаком соответствующих параметров регрессии.

При статистической оценке тесноты линейной связи между результативным y и двумя факторными признаками x_1, x_2 частный коэффициент корреляции результативного признака и первого фактора x_1 при элиминации второго фактора x_2 равен

$$R_{y1(2)} = \sqrt{\frac{\sigma^2_{y12} - \sigma^2_{y2}}{\sigma^2_y - \sigma^2_{y2}}}.$$

Частный коэффициент корреляции результативного признака и второго фактора x_2 при элиминации первого фактора x_1 равен

$$R_{y2(1)} = \sqrt{\frac{\sigma^2_{y12} - \sigma^2_{y1}}{\sigma^2_y - \sigma^2_{y1}}}.$$

Частный коэффициент корреляции может быть рассчитан через парные коэффициента корреляции r_{xy}, например:

$$r_{y1(2)} = r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}}.$$

Квадрат частного коэффициента корреляции является частным коэффициентом детерминации

$$B_{ym(12\ldots m-1)} = r^2_{ym(12\ldots m-1)}.$$

Анализ тесноты связи качественных показателей

Оценка степени тесноты связей качественных признаков осуществляется для парных и множественных связей, представленных альтернативными значениями и в порядковой шкале признаков.

Коэффициент корреляции рангов

Коэффициент корреляции рангов характеризует статистическую связь двух признаков, измеряемых в порядковой шкале.

Для признаков, измеренных в порядковых шкалах, наиболее известным является коэффициент ранговой корреляции Спирмена. При отсутствии связанных (одинаковых) рангов коэффициент Спирмена вычисляется по формуле

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_k^2}{n^3 - n} = 1 - \frac{6 \cdot \sum_{i=1}^{n} (i_{k1} - i_{k2})^2}{n(n^2 - 1)},$$

где d_k – разность рангов k–го объекта; n – количество объектов; i_{k1}, i_{k2} – ранги k–го объекта по первому и второму признакам.

Ранговые коэффициенты корреляции являются наиболее простыми показателями степени тесноты корреляционной зависимости. Они могут применяться не только для качественных, но и для количественных признаков. Чем ближе по абсолютной величине коэффициент корреляции рангов к 1, тем теснее связь.

Коэффициент конкордации

Коэффициент конкордации характеризует связь между несколькими признаками, измеряемыми в порядковой шкале.

Имеется выборка объемом n из m –мерной генеральной совокупности, признаки x_j которой можно измерить в порядковой шкале (табл.3.2).

Если при ранжировании имеются одинаковые наблюдения, то вместо обычных рангов, определяемых в вариационном ряду, каждому из этих одинаковых значений приписывается одно и то же число, равное средней арифметической их рангов. Получаемые таким образом ранги называются объединенными, или связанными.

Таблица 3.2

Ранги признаков

Номер наблюдения	Номер признака					
	1	2	. . .	j	. . .	m
1	i_{11}	i_{12}	. . .	i_{1j}	. . .	i_{1m}
2	i_{21}	i_{22}	. . .	i_{2j}	. . .	i_{2m}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
k	i_{k1}	i_{k2}	. . .	i_{kj}	. . .	i_{km}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
n	i_{n1}	i_{n2}	. . .	i_{nj}	. . .	i_{nm}

где i_{kj} – ранг k –го наблюдения j –го признака; $k = 1, \ldots, n$ – номер объекта; $j = 1, \ldots, m$ – номер признака.

Коэффициент конкордации (согласованности) для выборки объемом n при отсутствии связанных рангов вычисляется по формуле

$$K = \frac{12}{m^2(n^3 - n)} \sum_{k=1}^{n} [\sum_{j=1}^{m} i_{kj} - \frac{m(n+1)}{2}]^2.$$

Если расчет ведется с учетом связанных рангов, то формула коэффициента конкордации имеет вид

$$K = \frac{\sum_{k=1}^{n} [\sum_{j=1}^{m} i_{kj} - \frac{m(n+1)}{2}]^2}{\frac{1}{12} m^2(n^3 - n) - m\sum_{j=1}^{m} T_j},$$

где $T_j = \frac{1}{12} \sum_{i=1}^{m_j} (n_i^3 - n_i)$;

n_i – число неразличимых элементов (рангов) в i –й группе признака x_j;

m_j – число групп из неразличимых рангов.

3.2. Примеры решения задач по теме «Корреляционно-регрессионный анализ связей социально-экономических явлений»

Задача 3.1

Имеется информация о среднегодовой стоимости основных производственных фондов и объеме товарной продукции по десяти промышленным предприятиям (табл.3.3.).

Построить уравнение зависимости объема товарной продукции от среднегодовой стоимости основных производственных фондов. Определить тесноту связи между показателями.

Таблица 3.3

Исходная информация

Номер предприятия	Объем товарной продукции, y_i, млн р.	Среднегодовая стоимость основных фондов, x_i, млн р.
1	56	100
2	141	235
3	154	200
4	42	17
5	164	175
6	66	70
7	174	240
8	118	156
9	47	74
10	37	62

Решение

Для определения формы связи между показателями исходную информацию представим в виде эмпирического распределения (рис.3.1) и проверим две гипотезы о линейной и степенной зависимостях стоимости продукции от среднегодовой стоимости основных фондов.

Рассчитаем коэффициенты и индексы корреляции и детерминации для определения тесноты связи между показателями.

Линейное уравнение регрессии.

Линейная зависимость выражается уравнением прямой

$$y_x = a_0 + a_1 x,$$

где y_x – теоретическая зависимость результативного признака;

x – факторный признак;

a_0 и a_1 – параметры уравнения прямой (уравнения регрессии).

Параметры уравнения прямой a_0 и a_1 определяются путем решения системы уравнений

$$\begin{cases} \sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} x_i; \\ \sum_{i=1}^{n} x_i y_i = a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2, \end{cases}$$

где x_i, y_i – индивидуальные значения соответственно факторного и результативного признаков;

n – число индивидуальных значений признака.

Линейный коэффициент корреляции.

При линейной зависимости для определения тесноты связи между результативным и факторным признаками и оценки степени влияния факторного признака на результативный используются коэффициенты корреляции и детерминации.

Коэффициент корреляции, характеризующий степень тесноты связи между результативным и факторным признаками, для линейного уравнения корреляционной зависимости рассчитывается по следующей формуле

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\ \sqrt{n\sum_{i=1}^{n} y^2{}_i - (\sum_{i=1}^{n} y_i)^2}}.$$

Коэффициент детерминации

$$D_{xy} = r_{xy}^2$$

характеризует долю влияния факторного признака на вариацию результативного.

Расчет суммарных значений, используемых для расчета коэффициентов корреляционной зависимости и коэффициента корреляции приведен в графах 3-5 табл.3.4.

Таблица 3.4

Расчетная таблица для определения коэффициентов линейного уравнения регрессии

Номер предприятия	Объем товарной продукции, y_i, млн р.	Среднегодовая стоимость основных фондов, x_i, млн р.	$x_i y_i$	x_i^2	y_i^2	y_x
А	1	2	3	4	5	6
1	56	100	5600	10000	3136	78,19
2	141	235	33135	55225	19881	167,29
3	154	200	30800	40000	23716	144,19

A	1	2	3	4	5	6
4	42	17	714	289	1764	23,41
5	164	175	28700	30625	26896	127,69
6	66	70	4620	4900	4356	58,39
7	174	240	41760	57600	30276	170,59
8	118	156	18408	24336	13924	115,15
9	47	74	3478	5476	2209	61,03
10	37	62	2294	3844	1369	53,11
Итого	999	1329	169509	232295	127527	999,04

Значения, полученные в табл.3.4, подставляются в систему уравнений

$$\begin{cases} 999 = 10 \cdot a_0 + a_1 \cdot 1329; \\ 169509 = a_0 \cdot 1329 + a_1 \cdot 232295, \end{cases}$$

из решения которой определяются коэффициенты a_0 и a_1

$$a_0 = \frac{999 - a_1 \cdot 1329}{10} = 99{,}9 - a_1 \cdot 132{,}9$$

$$169509 = (99{,}9 - a_1 \cdot 132{,}9) \cdot 1329 + a_1 \cdot 232295 = 132767 - a_1 \cdot 176624 + a_1 \cdot 232295;$$

$$55671 \cdot a_1 = 36742; \qquad a_1 = 0{,}65998 \approx 0{,}66;$$

$$a_0 = 99{,}9 - a_1 \cdot 132{,}9 = 99{,}9 - 0{,}66 \cdot 132{,}9 = 12{,}186.$$

Корреляционная зависимость (уравнение регрессии) имеет следующий вид

$$y_x = 12{,}19 + 0{,}66 \cdot x,$$

Результаты расчета значений линейного уравнения регрессии приведены в графе 6 табл.3.4. и на графике (рис.3.1).

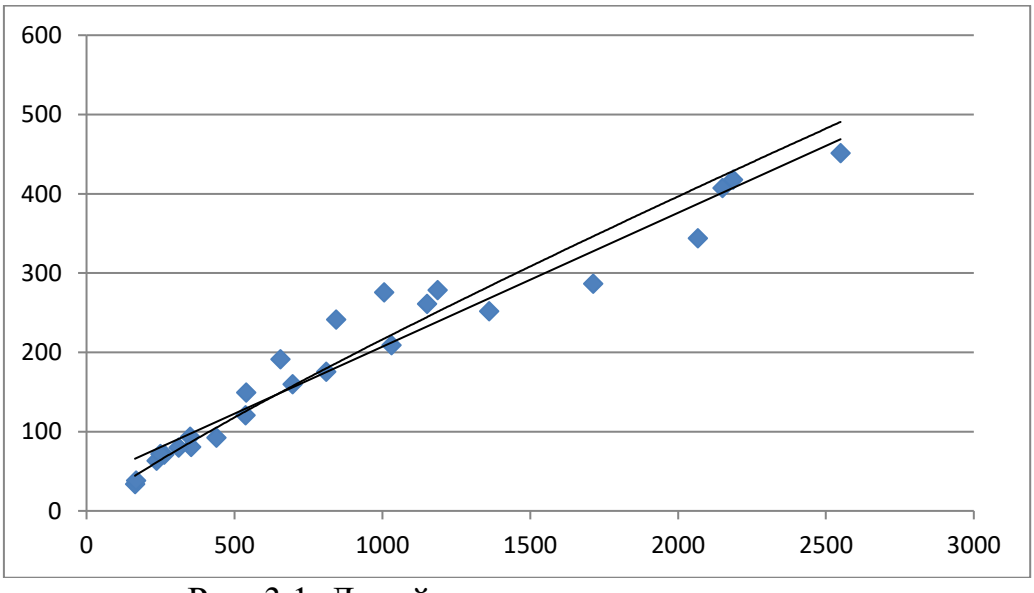

Рис. 3.1. Линейная и степенная регрессия

Линейный коэффициент корреляции равен

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} =$$

$$= \frac{10 \cdot 169509 - 1329 \cdot 999}{\sqrt{10 \cdot 232295 - 1329^2}\sqrt{10 \cdot 127527 - 999^2}} = \frac{367419}{392884} = 0{,}935,$$

что свидетельствует о наличии сильной прямой связи между объемом товарной продукции и среднегодовой стоимостью основных фондов.

Коэффициент детерминации

$$D_{xy} = r_{xy}^2 = 0{,}874$$

показывает, что изменение среднегодовой стоимости основных фондов влияет на изменение объема товарной продукции на 87,4%.

Степенное уравнение регрессии.

Степенная зависимость между показателями имеет следующий вид

$$y_x = a_0 x^{a_1}.$$

Для определения параметров производится логарифмирование степенной функции

$$\lg y = \lg a_0 + a_1 \lg x.$$

и строится система нормальных уравнений с использованием метода наименьших квадратов

$$\begin{cases} \sum_{i=1}^{n} \lg y_i = n \lg a_0 + a_1 \sum_{i=1}^{n} \lg x_i; \\ \sum_{i=1}^{n} \lg y_i \lg x_i = \lg a_0 \sum_{i=1}^{n} \lg x_i + a_1 \sum_{i=1}^{n} (\lg x_i)^2. \end{cases}$$

Расчет суммарных значений, используемых для расчета коэффициентов корреляционной зависимости и коэффициента корреляции приведен в графах 3-5 табл.3.5.

Таблица 3.5

Расчетная таблица для определения коэффициентов степенного уравнения регрессии

Номер предприятия	Объем товарной продукции, y_i, млн р.	Среднегодовая стоимость основных фондов, x_i, млн р.	$\lg y_i$	$\lg x_i$	$\lg y_i \lg x_i$	$(\lg x_i)^2$	y_x
А	1	2	3	4	5	6	7
1	56	100	1,7482	2,0000	3,4964	4,0000	84
2	141	235	2,1492	2,3711	5,0959	5,6220	147
3	154	200	2,1875	2,3010	5,0336	5,2947	132
4	42	17	1,6233	1,2305	1,9973	1,5140	27
5	164	175	2,2148	2,2430	4,9680	5,0312	121
6	66	70	1,8195	1,8451	3,3572	3,4044	67
7	174	240	2,2406	2,3802	5,3330	5,6654	149
8	118	156	2,0719	2,1931	4,5439	4,8098	112
9	47	74	1,6721	1,8692	3,1255	3,4940	69
10	37	62	1,5682	1,7924	2,8108	3,2127	62
Итого	999	1329	19,2953	20,2256	39,7616	42,0482	970

Значения, полученные в табл.3.5, подставляются в систему уравнений

$$\begin{cases} 19{,}2953 = 10 \cdot \lg a_0 + a_1 \cdot 20{,}2256; \\ 39{,}7616 = \lg a_0 \cdot 20{,}2256 + a_1 \cdot 42{,}0482, \end{cases}$$

из решения которой определяются $\lg a_0$ и a_1

$$\lg a_0 = \frac{19{,}2953 - a_1 \cdot 20{,}2256}{10} = 1{,}92953 - a_1 \cdot 2{,}02256$$

$$39{,}7616 = (1{,}92953 - a_1 \cdot 2{,}02256) \cdot 20{,}2256 + a_1 \cdot 42{,}0482 =$$

$$= 39{,}0259 - a_1 \cdot 40{,}9075 + a_1 \cdot 42{,}0482;$$

$$1{,}1407 \cdot a_1 = 0{,}7357; \qquad a_1 = 0{,}64495 \approx 0{,}65;$$

$$\lg a_0 = 1{,}92953 - a_1 \cdot 2{,}02256 = 1{,}92953 - 0{,}64495 \cdot 2{,}02256 = 0{,}62508; \quad a_0 = 4{,}22.$$

Корреляционная зависимость (уравнение регрессии) имеет следующий вид

$$y_x = a_0 x^{a_1} = 4{,}22 \cdot a^{0{,}65}.$$

Результаты расчета значений степенного уравнения регрессии приведены в графе 7 табл.3.5. и на графике (рис.3.1).

Суммарные величины эмпирических и теоретических значений результативного признака (графы 1 и 7 табл.3.5) не совпадают из-за округления рассчитанных коэффициентов уравнения регрессии.

Парные индексы корреляции и детерминации.

При парных нелинейных зависимостях для определения тесноты связи между результативным и факторным признаками и оценки степени влияния факторного признака на результативный используются индексы корреляции и детерминации.

Индекс корреляции

$$R_{xy} = \sqrt{\frac{\sigma^2_{y_x}}{\sigma^2_y}},$$

где $\sigma^2_{y_x}$ – факторная дисперсия результативного признака y ;

σ^2_y – общая дисперсия результативного признака.

Индекс детерминации

$$B_y = R^2_{xy} = \frac{\sigma^2_{y_x}}{\sigma^2_y}.$$

Факторная дисперсия результативного признака

$$\sigma^2_{y_x} = \frac{\sum_{i=1}^{n} (y_{x_i} - \bar{y})^2}{n},$$

где y_{x_i} – теоретические значения результативного признака при значениях факторного признака x_i ;

$\bar{y}$ – среднее значение результативного признака.

Общая дисперсия результативного признака

$$\sigma^2_y = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n},$$

где y_i – эмпирические значения результативного признака.

Среднее значение результативного признака

$$\bar{y} = \frac{\sum_{i=1}^{n} y_{x_i}}{n} = \frac{970}{10} = 97 \text{ млн р.}$$

Расчет числителей факторной и общей дисперсий приведен в табл.3.6.

Таблица 3.6

Расчетная таблица для определения индексов корреляции и детерминации при степенной зависимости между показателями

Номер предприятия	Объем товарной продукции, y_i, млн р.	Среднегодовая стоимость основных фондов, x_i, млн р.	y_x	$(y_{x_i} - \bar{y})^2$	$(y_i - \bar{y})^2$
А	1	2	3	4	5
1	56	100	84	169	1681
2	141	235	147	2500	1936
3	154	200	132	1225	3249
4	42	17	27	4900	3025
5	164	175	121	576	4489
6	66	70	67	900	961
7	174	240	149	2704	5929
8	118	156	112	225	441
9	47	74	69	784	2500
10	37	62	62	1225	3600
Итого	999	1329	970	15208	27811

Факторная дисперсия результативного признака

$$\sigma_{y_x}^2 = \frac{\sum_{i=1}^{n} (y_{x_i} - \bar{y})^2}{n} = \frac{15208}{10} = 1521,$$

Общая дисперсия результативного признака

$$\sigma_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} = \frac{27811}{10} = 2781.$$

Индекс корреляции

$$R_{xy} = \sqrt{\frac{\sigma_{y_x}^2}{\sigma_y^2}} = +\sqrt{\frac{1521}{2781}} = +0,740$$

характеризует степенную взаимосвязь между объемом продукции и среднегодовой стоимостью основных фондов как сильную.

Индекс детерминации

$$B_y = R_{xy}^2 = 0{,}548$$

показывает, что изменение среднегодовой стоимости основных фондов влияет на изменение стоимости товарной продукции на 54,8%.

При сравнении линейного коэффициента корреляции (0,935) и индекса корреляции в случае степенной зависимости (0,74) можно сделать вывод о том, что линейное уравнение регрессии лучше отражает исходную статистическую информацию, чем степенная зависимость.

Задача 3.2

Имеется информация о среднегодовой стоимости основных производственных фондов, объеме товарной продукции и численности персонала по десяти промышленным предприятиям (табл.3.7.).

Построить линейное уравнение зависимости объема товарной продукции от среднегодовой стоимости основных производственных фондов и численности персонала. Определить тесноту связи между показателями.

Таблица 3.7

Исходная информация

Номер предприятия	Объем товарной продукции, y_i, млн р.	Среднегодовая стоимость основных фондов, x_{1i}, млн р.	Численность персонала, , x_{2i}, чел.
1	56	100	357
2	141	235	474
3	154	200	600
4	42	17	27
5	164	175	550
6	66	70	440
7	174	240	734
8	118	156	446
9	47	74	315
10	37	62	339

Решение

Линейное уравнение множественной регрессии

Наиболее простым видом уравнения множественной регрессии является линейное уравнение множественной регрессии с двумя переменными

$$y_{x_1x_2} = a_0 + a_1x_1 + a_2x_2.$$

Параметры этого уравнения определяются решением системы нормальных уравнений, составленных в результате применения метода наименьших квадратов:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_{1i} + a_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i; \\ a_0 \sum_{i=1}^{n} x_{1i} + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} y_i x_{1i}; \\ a_0 \sum_{i=1}^{n} x_{2i} + a_1 \sum_{i=1}^{n} x_{1i} x_{2i} + a_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} y_i x_{2i}. \end{cases}$$

Расчет суммарных значений, используемых для определения коэффициентов корреляционной зависимости и коэффициента корреляции приведен в графах 3-5 табл.3.8.

Таблица 3.8

Расчетная таблица для определения коэффициентов линейного уравнения регрессии

Номер предприятия	y_i, млн р.	x_{1i}, млн р.	x_{2i}, чел.	x_{1i}^2	x_{2i}^2	$x_{1i}x_{2i}$	$x_{1i}y_i$	$x_{2i}y_i$	$y_{x_1x_2}$
А	1	2	3	4	5	6	7	8	9
1	56	100	357	10000	127449	35700	5600	19992	80
2	141	235	474	55225	224676	111390	33135	66834	165
3	154	200	600	40000	360000	120000	30800	92400	149
4	42	17	27	289	729	459	714	1134	18
5	164	175	550	30625	302500	96250	28700	90200	132
6	66	70	440	4900	193600	30800	4620	29040	65
7	174	240	734	57600	538756	176160	41760	127716	178
8	118	156	446	24336	198916	69576	18408	52628	117
9	47	74	315	5476	99225	23310	3478	14805	63
10	37	62	339	3844	114921	21018	2294	12543	56
Итого	999	1329	4285	232295	2160777	684669	169509	507300	1023

Значения, полученные в табл.3.4, подставляются в систему уравнений

$$\begin{cases} 10 \cdot a_0 + a_1 \cdot 1329 + a_2 \cdot 4285 = 999; \\ a_0 \cdot 1329 + a_1 \cdot 232295 + a_2 \cdot 684666 = 169509; \\ a_0 \cdot 4285 + a_1 \cdot 684669 + a_2 \cdot 2160777 = 507300 \end{cases}$$

Разделим каждое уравнение на коэффициенты при a_0

$$\begin{cases} a_0 + a_1 \cdot 132{,}9 + a_2 \cdot 428{,}5 = 99{,}9; \\ a_0 + a_1 \cdot 174{,}79 + a_2 \cdot 515{,}18 = 127{,}55; \\ a_0 + a_1 \cdot 159{,}78 + a_2 \cdot 504{,}27 = 118{,}39. \end{cases}$$

Вычитаем из второго первое уравнение и из второго – третье.

$$\begin{cases} a_1 \cdot 41{,}89 + a_2 \cdot 86{,}68 = 27{,}65; \\ a_1 \cdot 15{,}01 + a_2 \cdot 10{,}91 = 9{,}16. \end{cases}$$

Разделим каждое уравнение на коэффициенты при a_1

$$\begin{cases} a_1 + a_2 \cdot 2{,}069 = 0{,}66; \\ a_1 + a_2 \cdot 0{,}727 = 0{,}61. \end{cases}$$

Вычитаем из первого второе уравнение и определяет коэффициент a_2

$$a_2 \cdot 1{,}342 = 0{,}05; \qquad a_2 = 0{,}037.$$

Определим коэффициенты a_1 и a_0

$$a_1 = 0{,}66 - a_2 \cdot 2{,}069 = 0{,}66 - 0{,}037 \cdot 2{,}069 = 0{,}5835 \approx 0{,}58;$$

$$a_0 = 99{,}9 - a_1 \cdot 132{,}9 - a_2 \cdot 428{,}5 = 99{,}9 - 0{,}58 \cdot 132{,}9 - 0{,}037 \cdot 428{,}5 =$$

$$= 6{,}4985 \approx 6{,}5.$$

Линейное уравнение множественной регрессии имеет следующий вид

$$y_{x_1 x_2} = 6{,}5 + 0{,}6 \cdot x_1 + 0{,}037 \cdot x_2.$$

Теоретические значения уравнения регрессии приведены в столбце 9 табл. 3.8.

Множественные коэффициенты корреляции и детерминации

При статистической оценке тесноты линейной связи между результативным y и двумя факторными признаками x_1, x_2 может быть использована следующая формула множественного коэффициента корреляции

$$r_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2}},$$

где $r_{yx_1}, r_{yx_2}, r_{x_1x_2}$ – парные коэффициенты корреляции.

Парный коэффициент корреляции, характеризующий тесноту связи между результативным признаком y и факторным x_1 (см. задачу 3.1)

$$r_{yx_1} = \frac{n\sum_{i=1}^{n} y_i x_{1i} - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_{1i}}{\sqrt{[n\sum_{i=1}^{n} x_{1i}^2 - (\sum_{i=1}^{n} x_{1i})^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}} = 0{,}935.$$

При расчете парных коэффициентов корреляции, характеризующих тесноту связи между результативным признаком y и факторным x_2, между факторными x_1 и x_2 в формулах записываются суммарные значения показателей из табл.3.8 и $\sum y^2$ =127527 из табл.3.4 (задача 3.1). Для упрощения расчетов все значения в числителях и знаменателях парных коэффициентов делятся на 10^4.

$$r_{yx_2} = \frac{n\sum_{i=1}^{n} y_i x_{2i} - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_{2i}}{\sqrt{[n\sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{2i})^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}} =$$

$$= \frac{10 \cdot 507300 - 999 \cdot 4285}{\sqrt{10 \cdot 2160777 - 4285^2} \cdot \sqrt{10 \cdot 127527 - 999^2}} =$$

$$= \frac{507{,}3 - 999 \cdot 0{,}4285}{\sqrt{2160{,}777 - 42{,}85^2} \cdot \sqrt{127{,}527 - 9{,}99^2}} = 0{,}835;$$

$$r_{x_1x_2} = \frac{n\sum_{i=1}^{n} x_{1i} x_{2i} - \sum_{i=1}^{n} x_{1i} \sum_{i=1}^{n} x_{2i}}{\sqrt{[n\sum_{i=1}^{n} x_{1i}^2 - (\sum_{i=1}^{n} x_{1i})^2][n\sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{2i})^2]}} =$$

$$= \frac{10 \cdot 684669 - 1329 \cdot 4285}{\sqrt{10 \cdot 232295 - 1329^2} \cdot \sqrt{10 \cdot 2160777 - 4285^2}} =$$

$$= \frac{684{,}669 - 1329 \cdot 0{,}4285}{\sqrt{232{,}295 - 13{,}29^2} \cdot \sqrt{2160{,}777 - 42{,}85^2}} = 0{,}857.$$

Между факторами, т.е. среднегодовой стоимостью основных фондов и численностью персонала имеется тесная линейная связь ($r_{x_1x_2} = 0{,}857$), поэтому из модели можно исключить второй фактор и использовать парную регрессию.

Множественный коэффициент корреляции

$$r_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}} =$$

$$= \sqrt{\frac{0{,}935 + 0{,}835^2 - 2 \cdot 0{,}935 \cdot 0{,}835 \cdot 0{,}857}{1 - 0{,}857^2}} = 0{,}9359 \approx 0{,}936.$$

Коэффициент множественной детерминации

$$B_y = r_y^2 = 0{,}876.$$

Связь между стоимостью товарной продукции, среднегодовой стоимостью основных фондов и численностью персонала сильная, а изменение факторных признаков влияет на изменение стоимости продукции на 87,6%.

Частные коэффициенты корреляции и детерминации

При статистической оценке тесноты линейной связи между результативным y и двумя факторными признаками x_1, x_2 частный коэффициент корреляции результативного признака и первого фактора x_1 при элиминации второго фактора x_2 равен

$$r_{y1(2)} = r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}} =$$

$$= \frac{0{,}935 - 0{,}835 \cdot 0{,}857}{\sqrt{(1 - 0{,}835^2) \cdot (1 - 0{,}857^2)}} = 0{,}774.$$

Частный коэффициент корреляции результативного признака и второго фактора x_2 при элиминации первого фактора x_1 равен

$$r_{y2(1)} = r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1x_2}^2)}} =$$

$$= \frac{0{,}835 - 0{,}935 \cdot 0{,}857}{\sqrt{(1 - 0{,}935^2) \cdot (1 - 0{,}857^2)}} = 0{,}184.$$

Частные коэффициенты детерминации

$$B_{y1(2)} = r^2_{y1(2)} = 0{,}774^2 = 0{,}548;$$

$$B_{y2(1)} = r^2_{y2(1)} = 0{,}184^2 = 0{,}034.$$

Задача 3.3

Имеется информация о среднегодовой стоимости основных производственных фондов и объеме товарной продукции по десяти промышленным предприятиям (графы 1 и 2 табл.3.10.).

Определить коэффициент ранговой корреляции Спирмена.

Таблица 3.10

Расчет рангового коэффициента корреляции

Номер предприятия, k	Объем товарной продукции, млн р.	Среднегодовая стоимость основных фондов, млн р.	Ранги		Разность рангов d_k	d_k^2
			объема товарной продукции i_{k1}	среднегодовой стоимости основных фондов i_{k2}		
А	1	2	3	4	5	6
1	56	100	4	5	–1	1
2	141	235	7	9	–2	4
3	154	200	8	8	0	0
4	42	17	2	1	1	1
5	164	175	9	7	2	4
6	66	70	5	3	2	4
7	174	240	10	10	0	0
8	118	156	6	6	0	0
9	47	74	3	4	–1	1
10	37	62	1	2	–1	1
Итого						16

Для расчета коэффициента ранговой корреляции определяются ранги объема товарной продукции i_{k1}, среднегодовой стоимости основных фондов i_{k2} (графы 3,4 табл.3.9) и квадрат разности рангов (графы 5, 6 табл.3.9).

Коэффициента ранговой корреляции Спирмена при отсутствии связанных рангов

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_k^2}{n^3 - n} = 1 - \frac{6\sum_{i=1}^{n} (i_{k1} - i_{k2})^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 16}{10(100 - 1)} = 0{,}903,$$

где d_k – разность рангов k–го объекта;

n – количество объектов;

i_{k1}, i_{k2} – ранги k–го объекта по первому и второму признакам.

Величина коэффициента ранговой корреляции свидетельствует о наличии прямой тесной связи между объемом выпуска продукции и среднегодовой стоимостью основных фондов.

Коэффициент конкордации

Задача 3.4

Имеется информация об объеме товарной продукции, среднегодовой стоимости основных производственных фондов и численности персонала по десяти промышленным предприятиям (графы 1, 2, 3 табл.3.11).

Определить коэффициент конкордации.

Решение

Коэффициент конкордации (согласованности) для выборки объемом n при отсутствии связанных рангов вычисляется по формуле

$$K = \frac{12}{m^2(n^3 - n)} \sum_{k=1}^{n} [\sum_{j=1}^{m} i_{kj} - \frac{m(n+1)}{2}]^2,$$

где i_{kj} – ранг k –го наблюдения j –го признака;

k = 1, . . . , n – номер объекта;

j = 1, . . . , m – номер признака.

Ранги признаков (графы 4, 5 и 6 табл.3.11) определяются в соответствии с возрастающими значениями этих признаков.

Таблица 3.11

Ранги признаков

Номер предприятия	Объем товарной продукции, млн р.	Среднегодовая стоимость основных фондов, млн р.	Численность персонала, , чел.	Ранги		
				объема товарной продукции i_{k1}	среднегодовой стоимости основных фондов i_{k2}	численности персонала i_{k3}
А	1	2	3	4	5	6
1	56	100	357	4	5	4
2	141	235	474	7	9	7
3	154	200	600	8	8	9
4	42	17	27	2	1	1
5	164	175	550	9	7	8
6	66	70	440	5	3	5
7	174	240	734	10	10	10
8	118	156	446	6	6	6
9	47	74	315	3	4	2
10	37	62	339	1	2	3

Для количества предприятий n=10 и числе признаков m=3

$$\frac{m(n+1)}{2} = \frac{3 \cdot (10+1)}{2} = 16{,}5$$

рассчитываются суммарные значения рангов по признакам для каждого объекта и по всем объектам (графы 4, 5 и 6 табл.3.12).

Таблица 3.12

Расчет коэффициента конкордации

Номер предприятия	Ранги			$\sum_{j=1}^{m} i_{kj}$	$\sum_{j=1}^{m} i_{kj} - 16{,}5$	$(\sum_{j=1}^{m} i_{kj} - 16{,}5)^2$
	объема товарной продукции i_{k1}	среднегодовой стоимости основных фондов i_{k2}	численности персонала i_{k3}			
А	1	2	3	4	5	6
1	4	5	4	13	–3,5	12,25
2	7	9	7	23	6,5	42,25
3	8	8	9	25	8,5	72,25
4	2	1	1	4	–12,5	156,25
5	9	7	8	24	7,5	56,25
6	5	3	5	13	–3,5	12,25
7	10	10	10	30	13,5	182,25
8	6	6	6	18	1,5	2,25
9	3	4	2	9	–7,5	56,25
10	1	2	3	6	–10,5	110,25
Итого						702,50

Коэффициент конкордации для выборки объемом n=10, числе признаков m=3 и отсутствии связанных рангов

$$K = \frac{12}{m^2(n^3 - n)} \sum_{k=1}^{n} [\sum_{j=1}^{m} i_{kj} - \frac{m(n+1)}{2}]^2 = \frac{12}{3^2(10^3 - 10)} \cdot 702{,}5 = 0{,}946$$

показывает, что между объемом товарной продукции, среднегодовой стоимостью основных производственных фондов и численностью персонала в данном случае существует тесная прямая связь.