МАТЕМАТИКА

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Методические указания для выполнения расчетных заданий

САНКТ-ПЕТЕРБУРГ
2019

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное
учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра высшей математики

МАТЕМАТИКА

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Методические указания для выполнения расчетных заданий

САНКТ-ПЕТЕРБУРГ
2019

УДК 519.2.06(073)

МАТЕМАТИКА. Элементы математической статистики. Корреляционно-регрессионный анализ: Методические указания для выполнения расчетных заданий / Санкт-Петербургский горный университет. Сост.: *Л.В. Бакеева, Е.В. Пастухова,* СПб, 2019. 42 с.

Методические указания разработаны в соответствии с требованиями государственного образовательного стандарта высшего образования.

Содержат основные понятия математической статистики, в них излагаются основы выборочного метода и корреляционно-регрессионного анализа. Изложение теоретического материала сопровождается разобранными типовыми примерами.

Методические указания могут быть использованы для работы на практических занятиях и для выполнения обучающимися заданий самостоятельной работы в соответствии с программами подготовки специалистов и бакалавров инженерно-технических и экономических направлений подготовки и специальностей по дисциплинам «Математика», «Теория вероятностей и математическая статистика»

Научный редактор проф. *А.П. Господариков*

Рецензент проф. *С.И. Перегудин* (СПбГУ)

1. ВЫБОРКИ И ИХ ХАРАКТЕРИСТИКИ

1.1 ПРЕДМЕТ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика – раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей.

Математическая статистика тесно связана с теорией вероятностей. Обе эти математические дисциплины изучают массовые случайные явления. При этом теория вероятностей выводит из математической модели свойства реального процесса, а математическая статистика устанавливает свойства математической модели, исходя из данных наблюдений (говорят «из статистических данных»).

Предметом математической статистики является изучение случайных величин (или случайных событий, процессов) по результатам наблюдений. Полученные в результате наблюдения (опыта, эксперимента) данные сначала надо каким-либо образом обработать: упорядочить, представить в удобном для анализа виде. Это первая задача. Вторая задача – оценить, хотя бы приблизительно, интересующие исследователя характеристики наблюдаемой случайной величины. Например, дать оценки неизвестной вероятности события, неизвестной функции распределения, математического ожидания, дисперсии случайной величины и параметров распределения, вид которого неизвестен и т.д.

Следующей, назовем ее условно третьей задачей, является проверка статистических гипотез (согласование результатов оценивания с опытными данными). Например, выдвигается гипотеза, что: а) наблюдаемая случайная величина подчиняется нормальному закону; б) математическое ожидание наблюдаемой случайной величины равно нулю и т.д.

Одной из важнейших задач математической статистики является разработка методов, позволяющих по результатам исследования выборки (т.е. части общей совокупности объектов) сделать обоснованный вывод о распределении признака (случайной величины X) изучаемых объектов по всей совокупности.

Результаты исследования статистических данных методами математической статистики используются для принятия решения в задачах планирования, управления, прогнозирования и организации производства, при контроле качества продукции, при выборе оптимального времени настройки и замены действующей аппаратуры и т.д., то есть для научных и практических выводов.

Говорят, что *«математическая статистика – это теория принятия решений в условиях неопределенности».*

1.2 ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТИ

Пусть требуется изучить данную совокупность объектов относительно некоторого признака. Например, рассматривая работу диспетчера, можно исследовать его загруженность, тип клиентов, скорость обслуживания, время поступления заявок и т.д. Каждый такой признак или их комбинации представляет собой случайную величину.

Совокупность подлежащих изучению объектов или возможных результатов наблюдений, производимых в неизменных условиях над одним объектом приводит к понятию *генеральной совокупности*.

Определение. Генеральная совокупность – это случайная величина $X(\omega)$, заданная на пространстве элементарных событий Ω с выделенным в ней классом S подмножеств событий, для которых заданы их вероятности.

Зачастую проводить сплошное исследование, например, перепись населения, трудно или дорого, экономически нецелесообразно, а иногда невозможно. В этих случаях наилучшим способом исследования является выборочное наблюдение: выбирается из генеральной совокупности только часть ее объектов («выборка») для их изучения.

Выборочной совокупностью (выборкой) называется совокупность объектов, отобранных случайным образом из генеральной совокупности.

Определение. Выборка – это последовательность $X_1, X_2, ..., X_n$ независимых одинаково распределенных случайных

величин, распределение каждой из которых совпадает с распределением генеральной случайной величины.

Число объектов (наблюдений) в совокупности, генеральной или выборочной, называется ее *объемом*; обозначается соответственно N или n. Конкретные значения выборки, полученные в результате наблюдений (испытаний), называются *реализацией* выборки и обозначаются строчными буквами $x_1, x_2, ..., x_n$.

Метод статистического исследования, состоящий в том, что на основе изучения выборочной совокупности делается заключение о всей генеральной совокупности, называется *выборочным*.

Для получения удовлетворительных оценок характеристик генеральной совокупности необходимо, чтобы выборка была *репрезентативной* (или *представительной*), т.е. достаточно полно представлять изучаемые признаки генеральной совокупности. Условием обеспечения репрезентативности выборки является соблюдение случайности отбора (закон больших чисел), т.е. все объекты генеральной совокупности должны иметь равные вероятности попасть в выборку.

Различаются выборки с возвращением (*повторные*) и без возвращения (*бесповторные*). В первом случае отобранный объект возвращается в генеральную совокупность перед извлечением следующего; во втором - не возвращается. Заметим, что если объем выборки значительно меньше объема генеральной совокупности, различие между повторной и бесповторной выборками очень мало и его можно не учитывать.

В зависимости от конкретных условий для обеспечения репрезентативности применяют различные способы отбора: *простой*, при котором из генеральной совокупности извлекают по одному объекту; *типический*, при котором генеральную совокупность делят на «типические» части и отбор осуществляется из каждой части; *механический*, при котором отбор производится через определенный интервал; *серийный*, при котором объекты из генеральной совокупности отбираются «сериями» для сплошного их исследования. На практике обычно применяются сочетания вышеупомянутых способов отбора.

1.3 СТАТИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ВЫБОРКИ

Пример 1. Десять абитуриентов проходят тестирование по математике. Каждый из них может набрать от 0 до 5 баллов включительно. Пусть X_k - количество баллов, набранных $k-$м ($k = 1, 2, ..., 10$) абитуриентом. Тогда значения $0, 1, 2, 3, 4, 5$ – возможные количества баллов, набранных одним абитуриентом, – образуют генеральную совокупность. Выборка $X_1, X_2, ..., X_{10}$ – результат тестирования 10 абитуриентов. Реализациями выборки могут быть следующие наборы чисел: $\{5, 3, 0, 1, 4, 2, 5, 4, 1, 5\}$ или $\{4, 4, 5, 3, 3, 1, 5, 2, 2, 5\}$, т.е. все возможные комбинации десяти чисел от 0 до 5.

Пусть изучается некоторая случайная величина X. С этой целью над случайной величиной производится ряд независимых опытов (наблюдений). В каждом из этих опытов величина X принимает то или иное значение.

Пусть она приняла m_1 раз значение x_1, m_2 раз – значение $x_2, \ldots,$ m_k раз - значение x_k. При этом $m_1 + m_2 + ... + m_k = n$ - объем выборки. Значения x_1, $x_2, \ldots,$ x_k называются *вариантами* случайной величины X, а изменение этих значений *варьированием*.

Расположение выборочных наблюдаемых значений случайной величины (признака) в порядке неубывания называется *ранжированием* статистических данных.

Полученная таким образом последовательность $x_{(1)}$, $x_{(2)}, \ldots, x_{(n)}$ значений случайной величины X (где $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ и $x_{(1)} = \min\limits_{1 \le i \le n} X_i, \ldots, x_{(n)} = \max\limits_{1 \le i \le n} X_i$) называется *вариационным рядом*.

Числа m_i, показывающие сколько раз встречаются варианты x_i в ряде наблюдений, называются частотами, а отношение их к объему выборки – *частостями* или *относительными частостями* (обозначают p_i^* или w_i), то есть

$$p_i^* = \frac{m_i}{n}, \text{ где } n = \sum_{i=1}^{k} n_i .$$

Перечень вариантов и соответствующих им частот или частостей называется *статистическим распределением ряда* или *статистическим рядом*.

Различаются дискретные и непрерывные статистические ряды.

Дискретным статистическим рядом называется ранжированная совокупность вариант x_i с соответствующими им частотами. Записывается дискретный ряд в виде таблицы. Первая строка содержит варианты, а вторая их частоты или частости.

Пример 2. В результате тестирования (см. пример 1) группа абитуриентов набрала баллы: 5, 3, 0, 1, 4, 2, 5, 4, 1, 5. Записать полученную выборку в виде статистического ряда.

Решение. Случайная величина X – число набранных баллов является дискретной случайной величиной.

Вначале составим ранжированный вариационный ряд $x_{(1)}$, $x_{(2)}$,…, $x_{(10)}$, то есть расположим числа (баллы) в порядке неубывания их величин:

0, 1, 1, 2, 3, 4, 4, 5, 5, 5.

Вычислив частоту и частость вариантов $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$, получим статистическое распределение выборки (так называемый дискретный статистический ряд):

x_i	0	1	2	3	4	5
m_i	1	2	1	1	2	3

$\left(\sum_{i=1}^{6} n_1 = 10\right)$

или

x_i	0	1	2	3	4	5
p_i^*	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$

$\left(\sum_{i=1}^{6} p_i^* = 1\right)$.

В случае, когда число значений признака (случайной величины X) велико или признак является непрерывным (то есть когда случайная величина X может принять любое значение в некотором интервале), составляются *интервальный* статистический ряд. В первую строку таблицы статистического распределения вписываются

частичные промежутки $(x_0, x_1]$, $(x_1, x_2]$,…, $(x_{k-1}, x_k]$, берутся обычно одинаковыми по длине. Для определения величины интервала h можно использовать формулу Стерджесса:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}32 \lg n},$$

где $x_{\max} - x_{\min} = R$ – размах признака, т.е. разность между наибольшим и наименьшим значениями признака, $1 + 3{,}32 \lg n = m$ – число интервалов. За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{\min} - \frac{h}{2}$. Во второй строке статистического ряда вписываются количество наблюдений m_i $\left(i = \overline{1,k}\right)$, попавших в каждый интервал.

Пример 3. Измерили рост (с точностью до 1 см) 30 наудачу отобранных студентов. Результаты измерений таковы:

178, 160, 154, 183, 155, 153, 167, 186, 163, 155, 157, 175, 170, 160, 159, 173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175, 173, 164, 172.

Построить интервальный статистический ряд.

Решение. Для удобства проранжируем полученные данные:

153, 154, 155, 155, 156, 157, 158, 159, 160, 163, 164, 165, 166, 167, 167, 169, 170, 171, 171, 172, 173, 173, 175, 175, 178, 179, 179, 182, 183, 186.

Очевидно, что рост студентов – непрерывная случайная величина. Для полученной выборки: $x_{\min} = 153$, $x_{\max} = 186$. По формуле Стерджесса, при $n = 30$, находим длину частичного интервала:

$$h = \frac{186 - 153}{1 + 3{,}32 \lg 30} = \frac{33}{1 + 3{,}32 \lg 30} \approx \frac{33}{5{,}907} \approx 5{,}59.$$

Примем $h = 6$, тогда $x_{нач} = 153 - \frac{6}{2} = 150$.

Число интервалов: $m = 1 + 3{,}32 \lg 30 = 5{,}907 \approx 6$.

Исходные данные разбиваем на 6 промежутков: $(150;156]$, $(156;162]$, $(162;168]$, $(168;174]$, $(174;180]$, $(180;186]$.

Подсчитав число студентов m_i, попавших в каждый из полученных промежутков получим интервальный статистический ряд:

$(x_i; x_{i+1}]$	(150;156]	(156;162]	(162;168]	(168;174]	(174;180]	(180;186]
Частота, m_i	4	3	6	7	5	3
Частость, p_i^*	0,13	0,17	0,20	0,23	0,17	0,10

1.4 ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Эмпирической (статистической) функцией распределения называется функция $F^*(x)$, определяющая для каждого значения x относительную частоту события $X < x$. Следовательно, по определению:

$$F^*(x) = p^*\{X < x\}.$$

Для нахождения эмпирической функции распределения удобно $F^*(x)$ записать в виде:

$$F^*(x) = \frac{m_x}{n},$$

где n – объем выборки, m_x – число выборочных значений величины X, меньших x.

Эмпирическую функцию распределения можно задать таблично или графически.

Пример 4. Построить функцию $F^*(x)$, используя данные и результаты примера 2.

Решение. Объем выборки по условию примера $n = 10$. Наименьшая варианта равна 0, значит $m_x = 0$ при $x \le 0$ (наблюдений меньше 0 нет). Тогда $F^*(x) = \frac{0}{10} = 0$. Если $0 < x \le 1$, то неравенство $X < x$ выполняется для варианты $x_1 = 0$, которая встречается 1 раз ($m_x = 1$), поэтому $F^*(x) = \frac{1}{10} = 0{,}1$ и т.д. Окончательно получаем:

$$F^*(x)=\begin{cases}0, & \text{при } x\le 0,\\ 0{,}1, & \text{при } 0<x\le 1,\\ 0{,}3, & \text{при } 1<x\le 2,\\ 0{,}4, & \text{при } 2<x\le 3,\\ 0{,}5, & \text{при } 3<x\le 4,\\ 0{,}6, & \text{при } 4<x\le 5\\ 1, & \text{при } 5<x.\end{cases}$$

График эмпирической функции распределения приведен на рис. 1.

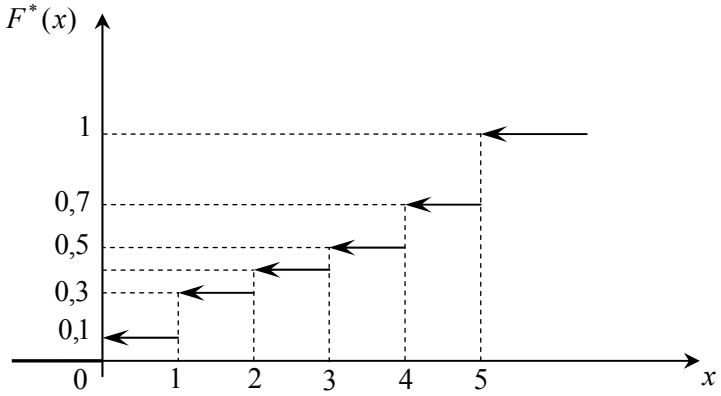

Рис. 1. Эмпирическая функция распределения дискретной случайной величины

В рассматриваемом примере функция $F^*(x)$ есть выборочная функция распределения дискретной случайной величины, построенная по дискретному статистическому ряду.

Если случайная величина непрерывная и ее выборочные значения представлены в виде интервального статистического ряда, то выборочную функцию распределения получают другим способом.

Рассмотрим построение эмпирической функции распределения для интервального статистического ряда на следующем примере.

Пример 5. Построить функцию $F^*(x)$, используя данные и результаты примера 3.

Решение. Очевидно, что для $x \in (-\infty, 150]$ $F^*(x) = 0$, так как $m_x = 0$.

Используя результаты расчетов, представленных в таблице, подсчитаем на концах интервалов значения функции $F^*(x)$ в виде «наращенной относительной частоты»:

Рост	(150;156]	(156;162]	(162;168]	(168;174]	(174;180]	(180;186]
$F^*(x)$	0,13	0,30	0,50	0,73	0,90	1,00

Табличные значения не полностью определяют выборочную функцию распределения непрерывной случайной величины, поэтому при графическом изображении такой функции ее доопределяют, соединив точки графика, соответствующие концам интервала, отрезками прямой (рис.2):

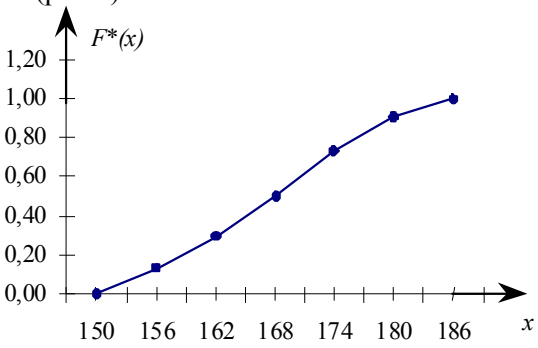

Рис. 2. Эмпирическая функция распределения непрерывной случайной величины

1.5 ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Статистическое распределение изображается графически (для наглядности) в виде так называемых полигона и гистограммы. Полигон, как правило, служит для изображения дискретного стати-

стического ряда (т.е. варианты отличаются на постоянную величину).

Полигоном частот называют ломаную, отрезки которой соединяют на плоскости точки с координатами (x_1, m_1), (x_2, m_2),…, (x_k, m_k); *полигоном частостей* – ломаную, соединяющую точки с координатами $(x_1,\ p_1^*)$, (x_2, p_2^*),…, $(x_k,\ p_k^*)$. Иногда полигон называют *многоугольником распределения*.

Варианты x_i откладываются на оси абсцисс, а частоты и соответственно частости – на оси ординат.

Пример 6. Пусть дана выборка в виде распределения частот:

x_i	0	1	2	3	4	5
m_i	1	2	1	1	2	3

$\left(\sum_{i=1}^{6} n_i = 10\right)$

Построить полигон частостей.

Решение. Статистический вариационный ряд можно записать в виде (см. пример 2):

x_i	0	1	2	3	4	5
p_i^*	0,1	0,2	0,1	0,1	0,2	0,3

$\left(\sum_{i=1}^{6} p_i^* = 1\right)$.

Полигон частостей для данного ряда имеет вид, изображенный на рис. 3:

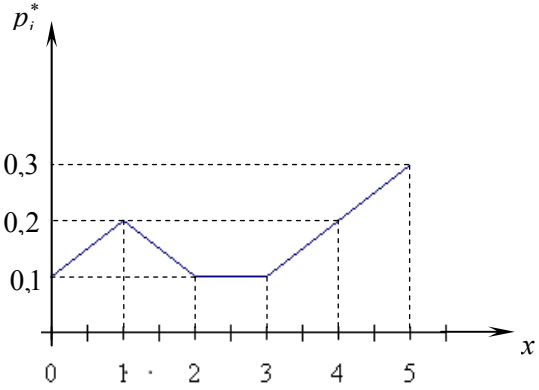

Рис.3. Полигон частостей

Полигон частостей является статистическим аналогом многоугольника распределения дискретной случайной величины.

Для непрерывно распределенного признака (то есть варианты могут отличаться одна от другой на сколь угодно малую величину) можно построить полигон частот, взяв середины интервалов в качестве значений признака $x_1, x_2, ..., x_k$. Однако, чаще распределение непрерывного признака изображают графически в виде так называемой гистограммы.

Гистограммой частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны частотам или частостям соответствующих интервалов. Если соединить середины верхних оснований прямоугольников отрезками прямой, то можно получить полигон того же распределения.

Пример 7. Построить гистограмму частостей по данным группировки промышленных предприятий по средней годовой стоимости основных производственных фондов, приведенным в таблице.

Группы предприятий по стоимости ОПФ, млн.руб.	19,8-23,8	23,8-27,8	27,8-31,8	31,8-35,8	35,8-39,8
Число предприятий, m_i	2	6	9	5	3

Решение. Для построения гистограммы частостей найдем p_i^*. Так как объем выборки $n = 25$, то $p_1^* = \frac{2}{25} = 0{,}08$; $p_2^* = \frac{6}{25} = 0{,}24$; $p_3^* = \frac{9}{25} = 0{,}36$; $p_4^* = \frac{5}{25} = 0{,}2$; $p_5^* = \frac{3}{25} = 0{,}12$.

Гистограмма частостей изображена на рис. 4:

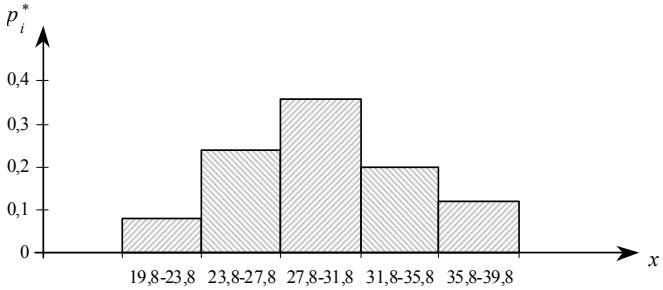

Рис. 4. Гистограмма частостей

Графическое изображение статистических распределений в виде полигона и гистограммы позволяет получить первоначальное представление о закономерностях, имеющих место в совокупности наблюдений.

1.6 ТОЧЕЧНЫЕ ОЦЕНКИ. ВЫБОРОЧНАЯ СРЕДНЯЯ И ВЫБОРОЧНАЯ ДИСПЕРСИЯ

Оценки параметров генеральной совокупности, полученные на основании выборки, называются *статистическими*. Если статистическая оценка характеризуется одним числом, она называется *точечной*. К числу таких оценок относятся выборочная средняя и выборочная дисперсия.

Выборочная средняя определяется как среднее арифметическое полученных по выборке значений:

$$\overline{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i \ ,$$

где x_i – варианта выборки;

n_i – частота варианты;

n – объем выборки.

Выборочную среднюю можно записать и так:

$$\overline{x}_{\text{в}} = \sum_{i=1}^{k} x_i \cdot p_i^* \ , \text{ где } \ p_i^* = \frac{n_i}{n} - \text{частость.}$$

Выборочная средняя может обозначаться и без нижнего индекса: $\bar{x}$.

Отметим, что в случае интервального статистического ряда в качестве варианты x_i берут середины интервалов ряда, а в качестве n_i – частоты соответствующих интервалов.

Выборочной дисперсией называется среднее арифметическое квадратов отклонений значений выборки от выборочной средней $\bar{x}_в$:

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 \cdot n_i \ ,$$

или, что то же самое,

$$D_в = \sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 \cdot p_i^* \ .$$

Для расчетов может быть использована также формула:

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

где $\overline{x^2}$ - выборочная средняя квадратов вариант выборки.

Выборочное *среднее квадратическое отклонение* выборки определяется формулой:

$$\sigma_в = \sqrt{D_в} \ .$$

Особенность выборочного среднего квадратического отклонения состоит в том, что оно измеряется в тех же единицах, что и изучаемый признак.

Статистическая оценка является случайной величиной и меняется в зависимости от выборки. Если математическое ожидание статистической оценки равно оцениваемому параметру генеральной совокупности, то такая оценка называется *несмещенной*, если не равно – то *смещенной*.

Выборочная средняя является оценкой математического ожидания случайной величины и представляет собой несмещенную оценку. Выборочная дисперсия оценивает дисперсию генеральной совокупности и является смещенной оценкой.

Для устранения смещенности выборочной дисперсии ее умножают на $\frac{n}{n-1}$ и получают величину:

$$S_{\text{в}}^{\,2} = \frac{n}{n-1} D_{\text{в}},$$

которая называется несмещенной или *исправленной выборочной дисперсией*.

Величина

$$S_{\text{в}} = \sqrt{S_{\text{в}}^{\,2}}$$

называется *исправленным выборочным средним квадратическим отклонением*.

Пример 8. Имеются данные о выручке в продовольственном магазине «Оазис» соответственно по месяцам (млн. руб.):

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
Выручка	2,2	2,5	2,3	2,2	2,3	2,5	2,2	2,2	2,4	2,3	2,4	2,2

Найти выборочную среднюю и выборочную дисперсию.

Решение. Построим сначала статистический ряд распределения:

Выручка, x_i	2,2	2,3	2,4	2,5
Частота, n_i	5	3	2	2

$\left(\sum_{i=1}^{4} n_1 = 12\right)$

Найдем выборочную среднюю:

$$\bar{x}_{\text{в}} = \frac{1}{12}\sum_{i=1}^{4} x_i \cdot n_i = \frac{2,2 \cdot 5 + 2,3 \cdot 3 + 2,4 \cdot 2 + 2,5 \cdot 2}{12} = 2,31.$$

Для вычисления выборочной дисперсии используем формулу $D_{\text{в}} = \overline{x^2} - (\bar{x}_{\text{в}})^2$. Чтобы воспользоваться данной формулой найдем сначала $\overline{x^2}$:

$$\overline{x^2} = \frac{2,2^2 \cdot 5 + 2,3^2 \cdot 3 + 2,4^2 \cdot 2 + 2,5^2 \cdot 2}{12} = \frac{24,2 + 15,87 + 11,52 + 12,5}{12} = 5,34$$

тогда $D_{\text{в}} = 5,34 - (2,31)^2 = 0,039$.

В качестве описательных характеристик вариационного ряда $x_{(1)}$, $x_{(2)}$,…, $x_{(n)}$ (или полученного из него статистического распределения выборки) используются медиана, мода, размах вариации (выборки).

Размах вариации определяется по формуле:

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

где $x_{\max}$ - наибольшая, $x_{\min}$ - наименьшая варианты ряда.

Модой Mo вариационного ряда называется варианта, имеющая наибольшую частоту.

Медианой Me вариационного ряда называется значение признака (варианта), приходящееся на середину ряда.

Если $n = 2k$ (то есть ряд $x_{(1)}$, $x_{(2)}$,…, $x_{(k)}$, $x_{(k+1)}$, $x_{(k+2)}$,…, $x_{(2k)}$ имеет четное число членов), то $Me = \dfrac{x_{(k)} + x_{(k+1)}}{2}$. Если $n = 2k+1$ (то есть ряд имеет нечетное число членов), то $Me = x_{(k+1)}$.

Пример 9. В результате тестирования (см. пример 2) группа абитуриентов набрала баллы: 5, 3, 0, 1, 4, 2, 5, 4, 1, 5. Найти характеристики выборки.

Решение. Статистическое распределение выборки (так называемый дискретный статистический ряд) имеет вид:

x_i	0	1	2	3	4	5
n_i	1	2	1	1	2	3

$\left(\sum_{i=1}^{6} n_i = 10\right)$

Тогда:

$$\bar{x}_{\text{в}} = \frac{1}{10} \cdot (0 \cdot 1 + 1 \cdot 2 + 2 \cdot 1 + 3 \cdot 1 + 4 \cdot 2 + 5 \cdot 3) = 3,$$

$$D_{\text{в}} = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x}_{\text{в}})^2 \cdot n_i =$$

$$= \frac{1}{10} \cdot \left((0-3)^2 \cdot 1 + (1-3)^2 \cdot 2 + (2-3)^2 \cdot 1 + (3-3)^2 \cdot 1 + (4-3)^2 \cdot 2 + (5-3)^2 \cdot 3\right) = 3,2$$

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{3,2} \approx 1,79,$$

$$S_{\text{в}}{}^2 = \frac{n}{n-1} D_{\text{в}} = \frac{10}{9} \cdot 3{,}2 \approx 3{,}56 ,$$

$$S_{\text{в}} = \sqrt{S_{\text{в}}{}^2} = \sqrt{3{,}56} \approx 1{,}87 ,$$

$$R = x_{\max} - x_{\min} = 5 - 0 = 5 ,$$

$Mo = 5$, так как 5 наиболее часто встречающаяся варианта,

$$Me = \frac{x_{(5)} + x_{(6)}}{2} = \frac{3+4}{2} = 3{,}5 .$$

Для непрерывно распределенного признака формулы для вычисления моды и медианы имеют вид:

$$Mo = x_{Mo} + h \cdot \frac{f_{Mo} - f_{(Mo-1)}}{\left(f_{Mo} - f_{(Mo-1)}\right) + \left(f_{Mo} + f_{(Mo-1)}\right)},$$

где x_{Mo} – начало модального интервального интервала, то есть интервала, имеющего наибольшую частоту,

f_{Mo} – частота модального интервального,

$f_{(Mo-1)}$ – частота интервала, предшествующего модальному,

$f_{(Mo+1)}$ – частота интервала, следующего за модальным,

h – интервал группировки;

$$Me = x_{Me} + h \cdot \frac{\frac{n+1}{2} - S_{(Me-1)}}{f_{Me}},$$

где x_{Me} – начало медианного интервала, то есть интервала содержащего серединные значения вариационного ряда,

$S_{(Me-1)}$ – накопленная частота интервала, предшествующего модальному.

2. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

2.1 ПОНЯТИЕ О КОРРЕЛЯЦИОННОЙ И РЕГРЕССИОННОЙ СВЯЗИ

Проводя исследования, необходимо считаться с взаимосвязью наблюдаемых процессов и явлений. При этом полнота описания, так или иначе, определяется количественными характеристиками

причинно-следственных связей между ними. Оценка наиболее существенных из них, а также воздействия одних факторов на другие является одной из основных задач статистики. Изучение реальных процессов обычно предполагает наблюдение над целым рядом случайных величин. Возникает задача изучения взаимосвязи между случайными величинами. Формы проявления взаимосвязей разнообразны. Различают два вида зависимостей между явлениями: функциональную и корреляционную (статистическую). При *функциональной* зависимости каждому значению независимой переменной X соответствует вполне определенное значение зависимой переменной Y.

В большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определенное, а множество возможных значений другой переменной. Иначе говоря, каждому значению одной переменной соответствует определенное (условное) распределение другой переменной. Такая зависимость называется *статистической* (или *стохастической*, *вероятностной*). Статистическую зависимость называют *корреляционной*, если при изменении значений одной величины меняется среднее значение другой. Если переменные не равноправны, т.е. четко ясно, какая из них причина, какая – следствие, то такая зависимость, при которой одна из переменных служит причиной изменения другой, называется *регрессионной*.

При сравнении функциональных и корреляционных зависимостей следует иметь в виду, что при функциональной зависимости, зная X, можно вычислить величину Y, а при корреляционной зависимости устанавливается лишь тенденция изменения Y при изменении X.

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа. Основными задачами корреляционного анализа являются выявление связи между случайными величинами и оценка тесноты этой связи. К основным задачам, которые решают с помощью регрессионного анализа, относят: установление формы зависимости между переменными (линейная-нелинейная, отрицательная-положительная и т.д.); определение функции регрессии в виде математического уравнения того

или иного типа и установление влияния объясняющих переменных на зависимую переменную; оценка неизвестных значений зависимой переменной (с помощью функции регрессии можно воспроизвести значения зависимой переменной внутри интервала заданных значений независимых переменных – интерполяция, или оценить течение процесса вне заданного интервала – экстраполяция).

2.2 КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Для характеристики корреляционной зависимости между случайными величинами вводится коэффициент корреляции r.

Коэффициент корреляции между двумя случайными величинами X и Y вычисляется по формуле:

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y},$$

где $\sigma_x = \sqrt{\overline{x^2} - (\overline{x}_в)^2}$, $\sigma_y = \sqrt{\overline{y^2} - (\overline{y}_в)^2}$ – средние квадратические отклонения случайных величин X и Y соответственно.

Отметим некоторые *свойства* коэффициента корреляции:

1. Если X и Y независимые случайные величины, то коэффициент корреляции равен нулю.

2. Коэффициент корреляции принимает значения на отрезке $[-1, 1]$, то есть $-1 \le r \le 1$. В зависимости от того, насколько $|r|$ приближается к 1, в математической статистике различают (шкала Шеддока): связи нет ($r < 0{,}2$), связь слабую ($0{,}2 \le r < 0{,}5$), умеренную ($0{,}5 \le r < 0{,}75$), тесную $(0{,}75 \le r \le 0{,}95)$ и очень тесную $(0{,}95 \le r < 1)$.

3. Если $|r| = 1$, то между случайными величинами X и Y имеет место функциональная, а именно линейная зависимость.

4. Коэффициент корреляции указывает на направление связи. Если $r > 0$, то связь прямая; если $r < 0$ отрицателен, что свидетельствует о наличии обратной связи.

Квадрат коэффициента корреляции называется коэффициентом *детерминации*: $\eta = r^2$.

Коэффициент детерминации η показывает, какая часть общей вариации Y обусловлена вариацией X.

Пример 10. С целью анализа влияния заработной платы на текучесть рабочей силы на пяти однотипных предприятиях проведены измерения уровня зарплаты (тыс. руб.) X и числа уволившихся за год рабочих Y:

X	30	40	50	55	60
Y	60	35	20	20	15

Определить степень влияния заработной платы на текучесть рабочей силы.

Решение. Для определения тесноты связи вычислим коэффициент корреляции, составив расчетную таблицу:

i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	**2**	**3**	**4**	**5**	**6**
1	30	60	900	3600	1800
2	40	35	1600	1225	1400
3	50	20	2500	400	1000
4	55	20	3025	400	1100
5	60	15	3600	225	900
$\sum$	**235**	**150**	**11625**	**5850**	**6200**

Так как коэффициент корреляции рассчитывается по формуле

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y},$$

то необходимые величины вычислим следующим образом:

1. Найдем средние значения: $\bar{x}$ (сумма значений второго столбца, деленная на число строк):

$$\bar{x} = \frac{\sum_{i=1}^{5} x}{n} = \frac{235}{5} = 47;$$

среднее значение $\bar{y}$ (сумма значений третьего столбца, деленная на число строк):

$$\bar{y} = \frac{\sum_{i=1}^{5} y}{n} = \frac{150}{5} = 30;$$

среднее значение $\overline{xy}$ (среднее значение шестого столбца):

$$\overline{xy} = \frac{\sum_{i=1}^{5} \overline{xy}}{n} = \frac{6200}{5} = 1240.$$

2. Найдем средние квадратические отклонения σ_x и σ_y:

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x}_в)^2} = \sqrt{\frac{\sum_{i=1}^{5} x^2}{n} - \left(\frac{\sum_{i=1}^{5} x}{n}\right)^2} = \sqrt{\frac{11625}{5} - \left(\frac{235}{5}\right)^2} =$$

$$\sqrt{2325 - (47)^2} = \sqrt{116} \approx 10{,}77.$$

Величина $\overline{x^2}$ рассчитывается как среднее значение четвертого столбца.

Аналогично $\sigma_y = \sqrt{\overline{y^2} - (\bar{y}_в)^2} = \sqrt{1170 - 30^2} = \sqrt{270} \approx 16{,}432$,

где $\overline{y^2}$ – среднее значение пятого столбца.

3. Подставляя найденные значения в формулу коэффициента корреляции, получим

$$r = \frac{1240 - 47 \cdot 30}{16{,}432 \cdot 10{,}77} = \frac{-170}{176{,}97} = -0{,}96.$$

Таким образом, можно сделать вывод, что связь между заработной платой и текучестью рабочей силы очень тесная и обратная, так как полученный коэффициент корреляции отрицательный. Это говорит о том, что чем меньше заработная плата (X), тем больше число уволившихся.

Выясним, какая часть вариации Y обусловлена вариацией X. Вычислим коэффициент детерминации:

$$\eta = r^2 = (-0{,}96)^2 = 0{,}92.$$

То есть вариации текучести рабочей силы (Y) на 92 % обусловлена вариацией заработной платы (X).

2.3 ЛИНЕЙНАЯ ПАРНАЯ РЕГРЕССИЯ

После того, как с помощью корреляционного анализа выявлено наличие статистических связей между переменными и оценена степень тесноты, обычно переходят к математическому описанию вида зависимостей с использованием регрессионного анализа. Если коэффициент корреляции $r < 0{,}2$, то согласно шкале Шеддока связи между переменными нет, и, следовательно, не имеет смысла описывать модель связи.

Регрессионная модель представляет собой математическое выражение, связывающее случайные величины X и Y. *Уравнение регрессии* – это зависимость величины Y от X.

Часто встречающейся моделью зависимости является *линейная парная корреляция*. Вообще говоря, уравнение регрессии может описывать взаимосвязь не двух, а более переменных (то есть быть не парной, а множественной). Кроме того, связь между переменными далеко не всегда линейна.

В общем случае уравнение регрессии имеет вид:

$$Y = \varphi(X, \beta) + \varepsilon,$$

где β – параметры модели, ε – ошибка наблюдений.

Уравнение парной линейной регрессии выглядит следующим образом:

$$\hat{y} = a + bx,$$

где a и b - параметры уравнения линейной регрессии.

Для нахождения параметром применяют *метод наименьших квадратов*, согласно которому неизвестные a и b выбираются таким образом, чтобы сумма квадратов отклонений эмпирических средних значений от значений, найденных по уравнению регрессии была минимальной:

$$\sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}_i\right)^2 \to \min.$$

Откуда получим систему нормальных уравнений для нахождения искомых параметров:

$$\begin{cases} a\,n + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i; \\ a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}$$

Разделив обе части уравнений на n, получим систему нормальных уравнений в виде:

$$\begin{cases} a + b\,\overline{x} = \overline{y}, \\ a\,\overline{x} + b\overline{x^2} = \overline{xy}. \end{cases}$$

Решая систему уравнений, найдем:

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}.$$

Зная, что $\overline{x^2} - \overline{x}^2 = \sigma_x^2$ и формулу для вычисления коэффициента корреляции можем записать:

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}.$$

Коэффициент b называется *коэффициентом регрессии*. Он показывает, на сколько единиц в среднем изменяется переменная Y при изменении X на одну единицу.

Замечание 1. Знак коэффициента регрессии указывает на направление связи: если $b > 0$, связь прямая, если $b < 0$ - обратная. Очевидно, что знаки коэффициентов корреляции и регрессии должны совпадать.

Решая систему относительно параметра a, получим:

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - b\overline{x}\,.$$

Для установления влияния на зависимую переменную независимой переменной, то есть для интерпретации модели используется коэффициент эластичности:

$$Э_x = b\frac{\overline{x}}{\overline{y}}.$$

Коэффициент эластичности показывает, на сколько процентов изменится Y при изменении X на 1 %.

Пример 11. В условиях предыдущей задачи найти уравнение линейной регрессии, выражающее зависимость между заработной платой рабочих и числом уволившихся.

Решение.

1. Для определения параметров a и b линии регрессии $\hat{y} = a\,x + b$ составим систему нормальных уравнений:

$$\begin{cases} a + b\,\overline{x} = \overline{y}, \\ a\,\overline{x} + b\overline{x^2} = \overline{xy}. \end{cases}$$

2. Подставляя найденные в предыдущей задаче средние значения $\overline{x} = 47$, $\overline{y} = 30$, $\overline{x^2} = 2325$, $\overline{xy} = 1240$, получим:

$$\begin{cases} a + 47b = 30, \\ 47a + 2325b = 1240. \end{cases}$$

3. Решая эту систему, найдем $b = -1{,}46$; $a = 98{,}85$. Тогда уравнение регрессии:

$$\hat{y} = 98{,}85 - 1{,}46x.$$

Отрицательный коэффициент регрессии подтверждает то, что связь между заработной платой рабочих и текучестью кадров обратная. Вычислим коэффициент эластичности:

$$Э_x = b\frac{\overline{x}}{\overline{y}} = -1{,}46 \cdot \frac{47}{30} \approx -2{,}3\%.$$

Полученный коэффициент свидетельствует о том, что при увеличении заработной платы на 1%, число увольняющихся в среднем сократится на 2,3 %.

Замечание 2. Выборочную линию регрессии можно задать также при помощи линейного уравнения $\hat{x} = c + dy$, параметры c и d для которого находятся из системы:

$$\begin{cases} c\,n + d\sum\limits_{i=1}^{n} y_i = \sum\limits_{i=1}^{n} x_i; \\ c\sum\limits_{i=1}^{n} y_i + b\sum\limits_{i=1}^{n} y_i^2 = \sum\limits_{i=1}^{n} x_i y_i \end{cases} \quad \text{или} \quad \begin{cases} c + d\,\overline{y} = \overline{x}, \\ c\,\overline{y} + d\,\overline{y^2} = \overline{xy}. \end{cases}$$

Коэффициент d называется *коэффициентом регрессии*. Он показывает, на сколько единиц в среднем изменяется переменная X при изменении Y на одну единицу.

$$d = r \cdot \frac{\sigma_x}{\sigma_y}.$$

Замечание 3. Следует иметь в виду, что $\hat{y} = a + bx$ и $\hat{x} = c + dy$ различные прямые (рис. 5). Первая прямая получается в результате решения задачи о минимизации суммы квадратов отклонений по вертикали, а вторая – при решении задачи о минимизации суммы квадратов отклонений по горизонтали. Линии регрессии, заданные этими уравнениями пересекаются в точке $M(\overline{x}, \overline{y})$ с координатами, соответствующими средним значениям корреляционно связанных между собой переменных.

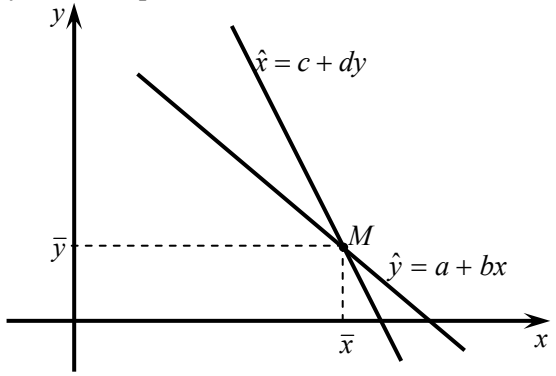

Рис. 5. Линии регрессии

Замечание 4. Если известны коэффициенты регрессии b и d, выборочные средние $\overline{x}$ и $\overline{y}$, то уравнения линейной регрессии могут быть найдены по формулам:

Y на X: $\hat{y} - \overline{y} = b(x - \overline{x})$ $\qquad$ X на Y: $\hat{x} - \overline{x} = d(y - \overline{y})$

3. ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Задание 1. Для выборок **а)**, **б)** и **в)** определить размах R, моду Mo, медиану Me, выборочное среднее $\bar{x}$, выборочную дисперсию $D_в$, «исправленную» выборочную дисперсию $S_в^2$. Для **а)** составить вариационный и статистический ряды; для **б)** найти эмпирическую функцию распределения $F_n^*(x)$; для **в)** построить гистограмму и полигон, эмпирическую функцию распределения $F_n^*(x)$.

1.1. **а)** 7, 3, 3, 6, 4, 5, 1, 2, 1, 3.

б)

x_i	11	13	15	17	19	21	23
n_i	2	4	8	12	16	10	3

в)

x_i	[0; 4)	[4; 8)	[8; 12)	[12; 16)	[16; 20)	[20; 24)
n_i	1	1	3	2	1	1

1.2. **а)** 6, 1, 4, 8, 5, 7, 2, 5, 7, 6.

б)

x_i	12	14	16	18	20	22	23
n_i	3	5	9	10	8	7	4

в)

x_i	[0; 3)	[3; 6)	[6; 9)	[9; 12)	[12; 15)	[15; 18)
n_i	4	1	2	3	3	1

1.3. **а)** 4, 9, 5, 2, 6, 9, 3, 3, 4, 9.

б)

x_i	10	12	14	16	18	20	22
n_i	6	4	1	5	7	8	10

в)

x_i	[0; 6)	[6; 12)	[12; 18)	[18; 24)	[24; 30)
n_i	2	3	1	1	2

1.4. **а)** 3, 7, 6, 4, 7, 1, 4, 2, 1, 2.

б)

x_i	9	11	13	15	17	19	21
n_i	9	5	4	7	8	10	6

в)

x_i	[0; 5)	[5; 10)	[10; 15)	[15; 20)	[20; 25)
n_i	1	3	4	2	1

1.5. **а)** 2, 5, 7, 6, 8, 3, 1, 5, 7, 5.

б)

x_i	8	10	12	14	16	18	20
n_i	2	5	4	6	10	6	7

в)

x_i	[0; 4)	[4; 8)	[8; 12)	[12; 16)	[16; 20)

n_i	2	4	1	3	4

1.6. **а)** 1, 3, 8, 8, 9, 5, 2, 3, 4, 8.

б)

x_i	7	10	13	16	19	22	23
n_i	6	1	7	10	6	4	2

в)

x_i	[0; 3)	[3; 6)	[6; 9)	[9; 12)	[12; 15)
n_i	2	4	4	1	3

1.7. **а)** 7, 1, 9, 2, 8, 7, 3, 2, 1, 1.

б)

x_i	6	9	12	15	18	21	22
n_i	6	8	10	4	5	7	9

в)

x_i	[0; 6)	[6; 12)	[12; 18)	[18; 24)	[24; 30)	[30; 36)
n_i	3	4	2	4	1	3

1.8. **а)** 6, 8, 1, 4, 7, 9, 4, 6, 7, 4.

б)

x_i	5	8	11	14	17	20	21
n_i	1	8	10	3	4	1	9

в)

x_i	[0; 5)	[5; 10)	[10; 15)	[15; 20)	[20; 25)	[25; 30)
n_i	2	2	1	3	3	1

1.9. **а)** 5, 6, 2, 6, 6, 1, 1, 4, 4, 7.

б)

x_i	4	7	10	13	16	19	20
n_i	3	10	8	1	6	4	6

в)

x_i	[0; 2)	[2; 4)	[4; 6)	[6; 8)	[8; 10)	[10; 12)
n_i	4	4	1	3	3	2

1.10. **а)** 4, 4, 3, 8, 5, 3, 2, 2, 1, 1.

б)

x_i	3	7	11	15	19	22	23
n_i	10	8	1	2	7	9	1

в)

x_i	[0; 4)	[4; 8)	[8; 12)	[12; 16)	[16; 20)	[20; 24)
n_i	2	3	3	4	2	1

1.11. **а)** 3, 2, 4, 2, 4, 5, 3, 6, 7, 3.

б)

x_i	1	5	9	13	17	21	22
n_i	6	4	5	8	2	4	4

в)

x_i	[0; 7)	[7; 14)	[14; 21)	[21; 28)
n_i	3	1	1	2

1.12. **а)** 2, 9, 5, 4, 3, 7, 4, 4, 4, 6.

б)

x_i	11	13	15	17	19	21	22
n_i	6	1	5	7	9	10	4

в)

x_i	[0; 3)	[3; 6)	[6; 9)	[9; 12)	[12; 15)	[15; 18)
n_i	2	4	3	2	2	1

1.13. **а)** 1, 7, 6, 6, 2, 9, 1, 2, 1, 9.

б)

x_i	12	14	16	18	20	22	23
n_i	6	1	8	4	10	8	7

в)

x_i	[0; 6)	[6; 12)	[12; 18)	[18; 24)	[20; 24)	[24; 30)
n_i	2	1	3	2	1	1

1.14. **а)** 7, 5, 7, 8, 1, 1, 2, 5, 7, 2.

б)

x_i	13	15	17	19	21	23	24
n_i	10	8	4	1	6	4	7

в)

x_i	[0; 4)	[4; 8)	[8; 12)	[12; 16)	[20; 24)
n_i	1	2	3	3	1

1.15. **а)** 6, 3, 8, 2, 2, 3, 3, 3, 4, 5.

б)

x_i	10	12	14	16	18	20	21
n_i	5	1	4	9	7	3	10

в)

x_i	[0; 2)	[2; 4)	[4; 6)	[6; 8)	[8; 10)	[10; 12)
n_i	4	1	3	2	4	1

1.16. **а)** 5, 1, 9, 4, 3, 5, 4, 2, 1, 8.

б)

x_i	9	11	13	15	17	19	20
n_i	8	4	5	6	10	6	8

в)

x_i	[0; 5)	[5; 10)	[10; 15)	[15; 20)	[20; 25)
n_i	2	1	3	2	4

1.17. **а)** 4, 8, 1, 6, 4, 7, 1, 5, 7, 1.

б)

x_i	8	10	12	14	16	18	19
n_i	8	1	2	5	7	2	1

в)

x_i	[0; 3)	[3; 6)	[6; 9)	[9; 12)	[12; 15)
n_i	1	4	3	2	4

1.18. **а)** 3, 6, 2, 8, 5, 9, 2, 3, 4, 4.

б)

x_i	7	9	11	13	15	17	20
n_i	8	1	5	2	7	8	5

в)

x_i	[0; 7)	[7; 14)	[14; 21)	[21; 28)	[28; 35)
n_i	1	3	1	1	2

1.19. **а)** 2, 4, 3, 2, 6, 1, 3, 2, 1, 7.

б)

x_i	6	8	10	12	14	16	19
n_i	4	9	2	5	1	7	10

в)

x_i	[0; 4)	[4; 8)	[8; 12)	[12; 16)
n_i	2	4	3	1

1.20. **а)** 1, 2, 4, 4, 7, 3, 4, 6, 7, 3.

б)

x_i	5	7	9	11	13	15	18
n_i	10	6	1	7	8	91	2

в)

x_i	[0; 6)	[6; 12)	[12; 18)	[18; 24)
n_i	2	3	4	2

1.21. **а)** 7, 9, 5, 6, 8, 5, 1, 4, 4, 6.

б)

x_i	4	6	8	10	12	14	19
n_i	9	5	1	7	8	6	4

в)

x_i	[0; 5)	[5; 10)	[10; 15)	[15; 20)	[20; 25)	[25; 30)
n_i	1	1	3	2	2	4

1.22. **а)** 6, 7, 6, 8, 4, 7, 1, 2, 1, 9.

б)

x_i	3	6	9	12	15	18	20
n_i	8	4	8	1	7	8	9

в)

x_i	[0; 4)	[4; 8)	[8; 12)	[12; 16)	[16; 20)	[20; 24)
n_i	4	1	3	4	2	2

1.23. **а)** 5, 5, 7, 2, 4, 9, 2, 6, 7, 2.

б)

x_i	2	5	8	11	14	17	22
n_i	9	4	7	10	9	10	2

в)

x_i	[0; 2)	[2; 4)	[4; 6)	[6; 8)	[8; 10)	[10; 12)
n_i	3	4	4	2	1	3

1.24. **а)** 4, 3, 8, 4, 4, 1, 3, 4, 4, 5.

б)

x_i	1	5	9	13	17	21	22
n_i	1	8	6	4	5	1	7

в)

x_i	[0; 7)	[7; 14)	[14; 21)	[21; 28)
n_i	1	4	3	2

1.25. **а)** 2, 1, 9, 6, 4, 3, 4, 2, 1, 8.

б)

x_i	9	10	11	12	13	14	15
n_i	4	10	8	1	3	4	9

в)

x_i	[0; 6)	[6; 12)	[12; 18)	[18; 24)	[24; 30)
n_i	2	1	3	2	1

1.26. **а)** 7, 8, 6, 5, 9, 4, 2, 3, 5, 5.

б)

x_i	4	7	10	13	16	19	22
n_i	8	6	1	9	6	8	2

в)

x_i	[0; 3)	[3; 6)	[6; 9)	[9; 12)	[12; 15)	[15; 18)
n_i	2	2	4	3	2	1

1.27. **а)** 5, 9, 5, 9, 3, 8, 1, 3, 1, 8.

б)

x_i	3	5	7	9	11	13	15
n_i	7	5	7	2	6	9	8

в)

x_i	[0; 1)	[1; 2)	[2; 3)	[3; 4)	[4; 5)	[5; 6)	[6; 7)
n_i	4	2	3	1	3	2	2

1.28. **а)** 6, 4, 8, 1, 5, 8, 3, 5, 8, 1.

б)

x_i	2	6	10	14	18	22	26
n_i	8	5	6	10	8	10	1

в)

x_i	[0; 5)	[5; 10)	[10; 15)	[15; 20)	[20; 25)	[25; 30)
n_i	3	2	4	1	2	3

1.29. **а)** 5, 2, 9, 3, 5, 1, 4, 3, 5, 4.

б)

x_i	1	6	11	16	21	26	31
n_i	2	7	7	3	6	1	8

в)

x_i	[0; 6)	[6; 12)	[12; 18)	[18; 24)	[24; 30)	[30; 36)
n_i	4	2	1	3	1	2

1.30. **а)** 2, 4, 7, 8, 2, 5, 2, 4, 1, 6.

б)

x_i	7	8	9	10	11	12	13
n_i	5	9	7	2	1	5	8

в)

x_i	[0; 7)	[7; 14)	[14; 21)	[21; 28)	[28; 35)
n_i	3	1	2	3	1

Задание 2. Для каждой из приведенных ниже выборок (см. по вариантам) (предполагается, что между признаками существует линейная зависимость):

1. Вычислить выборочный коэффициент линейной корреляции $r_в$ и оценить степень зависимости между переменными;

2. Найти уравнения прямых линий регрессии Y на X и X на Y, построить их графики;

3. Построить корреляционное поле, линии регрессии;

4. Интерпретировать полученную модель, сделать выводы и прогноз.

2.1. В таблице приведены данные о расходе топлива (y, л на 100 км) автомобиля с двигателем объемом 2 литра с автоматической трансмиссией в зависимости от скорости движения (x, км/ч).

x_i	10	30	40	70	90	110	130	140	150	160
y_i	4,5	4,8	5,1	6	7,5	8,1	9	9,8	11,3	14

Получить прогноз о расходе топлива при скорости 175 км/ч.

2.2. В таблице приведены данные о сроке службы колеса вагона в годах (x) и износа толщины обода колеса, (y, мм).

x_i	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
y_i	0,4	0,7	1,2	1,7	1,9	2,2	2,6	3	3,5	3,8

Получить прогноз об износе толщины обода колеса через 5,5 лет.

2.3. Показатели стоимости основных производственных фондов (x, млн. руб.) и среднесуточной производительности (y, тонны) приведены в таблице.

x_i	2,1	2,3	2,4	2,9	4,1	4,7	5,5	7,2	10,2	14,3
y_i	27	29	30	35	36	44	47	55	63	73

Получить прогноз среднесуточной производительности при стоимости основных производственных фондов 16 млн. руб.

2.4. В таблице приведены данные об остаточной величине глубины протектора передних колес автомобиля в мм (y) в зависимости от величины пробега (x, тыс. км).

x_i	0	5	10	15	20	30	40	50	60	70
y_i	9,0	8,5	7,9	7,5	7,0	6,1	5,0	4,1	3	2,0

Получить прогноз об износе протектора колеса через 42 тыс. км.

2.5. В таблице приведены данные о расходе топлива (y, л/100 км) автомобиля с дизельным двигателем объемом 2,2 литра с механической трансмиссией в зависимости от скорости движения (x, км/ч).

x_i	10	20	40	60	90	110	120	130	140	150
y_i	1,5	1,8	3	3,9	4,8	5,5	5,7	7	8,1	9,4

Получить прогноз о расходе топлива при скорости 160 км/ч.

2.6. В таблице приведены данные об остаточной величине глубины протектора задних колес автомобиля в мм (y) в зависимости от величины пробега (x, тыс. км).

x_i	0	10	20	30	40	50	60	70	80	90
y_i	9,0	8,2	7,4	6,6	5,8	4,9	4,1	3,3	2,5	1,8

Получить прогноз о предельно допустимом пробеге колес автомобиля при минимально допустимой глубине протектора 1,6 мм.

2.7. В таблице приведены данные о зависимости теплопроводности легких бетонов (y, Вт/(м· С^{o}) от плотности (x, кг/м^3).

x_i	800	900	1000	1100	1200	1300	1400	1500	1600	1700
y_i	0,2	0,22	0,24	0,28	0,33	0,38	0,4	0,42	0,44	0,47

Получить прогноз теплопроводности при плотности 1800 кг/м3.

2.8. В таблице приведены данные о количестве пропусков занятий (x, ч) студентом в течение учебного семестра и результатах (y, %) написания экзаменационного теста.

x_i,	2	4	8	12	14	20	24	26	30	34
y_i	85	75	70	60	50	40	20	15	10	5

Получить прогноз результатов теста при пропуске в 18 ч.

2.9. В таблице приведены данные о зависимости прочности портландцемента (y, МПа) от его удельной поверхности (x, см^2/г).

$x_i \cdot 10^3$	3	3,5	4	4,5	5	5,5	6	6,5	7	7,5
y_i	25	28	30	32	36	39	41	44	46	47

Получить прогноз о прочности при удельной поверхности 6,2 10^3 $см^2/г$.

2.10. В таблице приведены результаты измерений положения y (м) материальной точки в зависимости от времени t (сек).

t	1	2	3	4	5	6	7	8	9	10
y	5,1	6,9	9,1	10,8	13,2	14,9	17,2	18,8	21,2	22,9

Получить прогноз о возможном положении точки через 12 сек.

2.11. Для исследования износа рабочей части резца в зависимости от времени работы взяли 10 новых резцов и каждый день измеряли толщину рабочей части. Результаты сведены в таблицу, где y (мм) – толщина рабочей части резца, x – продолжительность работы в днях:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	0,1	0,15	0,3	0,4	0,45	0,55	0,65	0,75	0,9	1

Получить прогноз об износе толщины рабочей части резца за 12 дней.

2.12. В таблице приведены данные о растворимости (y) натриевой селитры $NaNO_3$ на 100 $г$ воды в зависимости от температуры (t, ^{0}C).

t_i	0	2	10	16	21	30	35	51	63	67
y_i	66,7	69,2	76,3	81,6	85,7	94,7	99,4	113,6	119,8	123

Получить прогноз о растворимости при температуре 60^0C.

2.13. За изменением реакции разложения аммиака следили по изменению давления (P, мм ртутного столба) в различные моменты времени (t, сек). Результаты наблюдений приведены в таблице.

t	100	200	300	400	500	600	700	800	1000
P	11	22,1	33,2	44	55,2	66,3	77,5	87,9	110

Получить прогноз о возможном давлении при t = 900.

2.14. В таблице приведены результаты измерений сопротивления проводника (R, Ом) в зависимости от температуры (t,^{0}C).

t	100	200	300	400	500	600	700	800	900	1000
R	15	19	23	27	31	34	37	39	42	45

Получить прогноз о возможном сопротивлении проводника при температуре 60^0C.

2.15. В таблице приведены результаты измерений положения y (м) материальной точки в зависимости от времени t (сек).

t	1	2	3	4	5	6	7	8	9	10
y	6,3	9,9	14,1	18,2	21,9	26,1	29,8	33,8	37,9	41,9

Получить прогноз о возможном положении точки через 11 сек.

2.16. В таблице приведена динамика валового выпуска (y, у.е.) за последние 10 лет (x – год).

x_i	1	2	3	4	5	6	7	8	9	10
y_i	178	182	190	199	200	213	220	231	235	242

Получить прогноз валового выпуска на следующий год.

2.17. Показатели стоимости основных производственных фондов (x, млн. руб.) и среднесуточной производительности (y, тонны) приведены в таблице.

x_i	2,1	2,3	2,4	2,9	4,1	4,7	5,5	7,2	10,2	14,3
y_i	27	29	30	35	36	44	47	55	63	73

Получить прогноз среднесуточной производительности при стоимости основных производственных фондов 16 млн. руб.

2.18. В таблице приведены данные об объемах производства (x, у.е.) некоторой компании в течение 10 месяцев и соответствующей операционной прибылью (y,тыс. руб.).

x_i	500	520	523	530	550	555	560	562	565	570
y_i	61	66,8	67	69	74	76,7	78	79	79,3	81

Получить прогноз о возможной месячной прибыли, если объем производства достигнет 600 у.е.

2.19. В таблице приведены данные об уровне безработицы (x) и уровне преступности (y) в некотором населенном пункте.

x_i	0,6	1,3	2,2	3,3	4,2	5,3	6,0	6,3	6,4	6,5
y_i	4,2	4,27	4,32	4,47	4,53	4,68	4,85	5,01	5,15	5,22

Получить прогноз уровня преступности в случае, когда безработица отсутствует.

2.20. В таблице приведены данные численности занятого населения (x, млн.) и валового выпуска продукции (y, у.е.).

x_i	70	73	74	75	76	77	79	80	81	83
y_i	219	241	250	264	265	272	281	291	309	320

Получить прогноз валового выпуска продукции в случае, если занятое население увеличится на 10% по сравнению с начальными данными (80 млн.)

2.21. В таблице приведены данные об объеме спроса (y, у.е.) на некоторую продукцию и цены на эту продукцию (x, тыс. руб.).

x_i	10	10,6	11	12	12,5	12,8	13	13,2	13,3	13,7
y_i	68	64	59	52	45	42	38	37	35	34

Получить прогноз объема спроса в случае, если цена на продукцию достигнет 14 тыс. руб.

2.22. В таблице приведены данные о времени работы (t) некоторого алгоритма в зависимости от количества его элементов (x).

x_i	9	12	14	16	18	20	21	23	24	25
t_i	200	280	320	380	460	510	600	690	750	820

Получить прогноз о времени работы алгоритма, состоящего из 30 элементов.

2.23. При моделировании распространения сетей беспроводного доступа были получены следующие данные о стоимости подключения потенциального абонента (y, у.е.) в зависимости от радиуса обслуживания базовой станции (x, км.) при плотности населения $\rho = 10$ чел./км^2.

x_i	1	1,5	2	2,5	3	3,5	4	4,5	5	6
y_i	3202	2897	2510	2130	1805	1300	1249	1001	820	615

Получить прогноз о стоимости подключения потенциального абонента в случае, если радиус обслуживания базовой станции составит 6,5 км.

2.24. В таблице приведены данные о длине диагонали экрана (x, дюйм) и качестве изображения (y, %) при нахождении на фиксированном расстоянии от экрана.

x_i	14	15	17	19	20	21	22	24	27	32
y_i	70	69	68,5	67	66,5	65,5	65	63	60	53

Получить прогноз о том, каким может быть качество изображения при диагонали экрана 40 дюймов.

2.25. В таблице приведены данные о показателях конкуренции (x) и средневзвешенные по частоте упоминания количества патентов (y)

x_i	0,87	0,88	0,89	0,9	0,91	0,92	0,93	0,94	0,95	0,96
y_i	3,35	3,62	4,21	4,5	4,9	5,3	5,8	6,11	6,3	6,1

Получить прогноз о количестве патентов, в случае, если показатель конкуренции равен 0,86.

2.26. В таблице приведены данные по группе грузовых автотранспортных предприятий города за отчетный год о показателях грузооборота (x, млн.т км) и суммам затрат на перевозки (y, тыс. руб.):

x_i	62	40	38	25	15	52	27	47	24	18
y_i	1550	1080	1033	750	472	1310	804	1245	724	579

Получить прогноз об объеме грузооборота, если сумма затрат на перевозки составила 1359 тыс. руб.

2.27. В таблице приведены данные по годовым отчетам промышленных предприятий города о среднесписочном числе работников (x, чел.) и объеме выпущенной продукции (y, млн. руб.):

x_i	700	1100	1285	705	1300	1450	800	1380	1425	1208
y_i	402	792	1116	435	1281	1756	510	1392	1756	1014

Получить прогноз о среднесписочном числе работников, если объем продукции за год составила 1579 млн. руб.

2.28. Торговое предприятие имеет сеть, состоящую из 10 магазинов, информация о деятельности которых: годовой товарооборот (у, млн. руб.) и торговая площадь (х, тыс. м^2) представлена в таблице:

x_i	0,25	0,42	0,57	0,59	0,79	0,95	0,99	1,23	1,29	1,33
y_i	21,9	40,1	43,2	44,3	58,3	70,6	77,2	91,2	93,2	93,4

Получить прогноз о годовом товарообороте в случае, если торговая площадь составит ровно 1 тыс. м^2.

2.29. В таблице приведены данные по основные показатели деятельности коммерческих банков региона: кредитные вложения (x, млн.руб.) и прибыль (y, млн. руб.):

x_i	50,2	0,5	89,8	88,3	21,0	59,1	156,0	136,4	150,8	99,9
y_i	25,1	,01	2,0	5,3	22,1	0,2	5,9	3,9	0,4	13,4

Получить прогноз о возможной прибыли от деятельности коммерческого банка, если кредитные вложения составили 73,5 млн. руб.

2. 30. В таблице приведены данные о расходе топлива (y, л на 100 км) автомобиля с двигателем объемом 1,5 литра с автоматической трансмиссией в зависимости от скорости движения (x, км/ч).

x_i	10	20	40	60	90	110	130	140	150	160
y_i	3,8	4	4,2	4,8	5,5	6	7	8,1	10	12

Получить прогноз о расходе топлива при скорости 170 км/ч.

ПРИМЕР ВЫПОЛНЕНИЯ ЗАДАНИЯ 2

Задание. Результаты 10 измерений зависимости выхода продукта Y (в кг/ч) от температуры реакции X (в °С), полученных в химическом производстве, представлены в следующей таблице:

X	30	35	40	45	50	55	60	65	70	75
Y	5	20	16	50	58	52	58	90	95	100

Для выборки:

1. Вычислить выборочный коэффициент корреляции $r_в$ и оценить степень зависимости между переменными;

2. Найти выборочные уравнения прямых линий регрессии Y на X и X на Y и построить их графики;

3. Построить корреляционное поле, линии регрессии;

4. Интерпретировать полученную модель, сделать выводы и прогноз.

Решение.

1. Вычислим выборочный коэффициент регрессии. Для этого составим расчетную таблицу:

i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	**2**	**3**	**4**	**5**	**6**
1	30	5	900	25	150
2	35	20	1225	400	700
3	40	16	1600	256	640
4	45	50	2025	2500	2250
5	50	58	2500	3364	2900
6	55	52	3025	2704	2860
7	60	58	3600	3364	3480
8	65	90	4225	8100	5850
9	70	95	4900	9025	6650
10	75	100	5625	10000	7500
Σ	$\sum x_i =$ $= 525$	$\sum y_i =$ $= 544$	$\sum x_i^2 =$ $= 29625$	$\sum y_i^2 =$ $= 39738$	$\sum x_i \cdot y_i =$ $= 32980$

Коэффициент корреляции рассчитывается по формуле:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}.$$

Найдем средние значения:

$\bar{x}$ (сумма значений второго столбца, деленная на число строк):

$$\bar{x} = \frac{\sum_{i=1}^{10} x}{n} = \frac{525}{10} = 52{,}5;$$

$\bar{y}$ (сумма значений третьего столбца, деленная на число строк):

$$\bar{y} = \frac{\sum_{i=1}^{10} y}{n} = \frac{544}{10} = 54{,}4;$$

$\overline{xy}$ (среднее значение шестого столбца):

$$\overline{xy} = \frac{\sum_{i=1}^{10} \overline{xy}}{n} = \frac{32980}{10} = 3298.$$

Найдем средние квадратические отклонения σ_x и σ_y:

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x}_в)^2} == \sqrt{\frac{29625}{10} - (52{,}5)^2} = \sqrt{206{,}25} = 14{,}36,$$

где $\overline{x^2}$ рассчитывается как среднее значение четвертого столбца.

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y}_в)^2} = \sqrt{\frac{39738}{10} - (54{,}4)^2} = \sqrt{1014{,}4} \approx 31{,}85,$$

Где $\overline{y^2}$ – среднее значение пятого столбца.

Подставляя найденные значения в формулу коэффициента корреляции, получим:

$$r = \frac{3298 - 52{,}5 \cdot 54{,}4}{14{,}66 \cdot 31{,}85} = 0{,}97.$$

2. Составим уравнения линейной регрессии.

Y и X.

Для определения параметров a и b линии регрессии $\hat{y} = a + bx$ составим систему нормальных уравнений:

$$\begin{cases} a + b\,\bar{x} = \bar{y}, \\ a\,\bar{x} + b\,\overline{x^2} = \overline{xy}. \end{cases}$$

Подставляя найденные в пункте 1 задачи средние значения $\bar{x} = 52{,}5$, $\bar{y} = 54{,}4$, $\overline{x^2} = 2962{,}5$, $\overline{xy} = 3298$, получим:

$$\begin{cases} a + 52{,}5b = 54{,}5; \\ 52{,}5a + 2962{,}5b = 3298. \end{cases}$$

Решая эту систему, найдем $b = 2{,}15$ и $a = -58{,}1$. Тогда уравнение регрессии Y и X имеет вид:

$$\hat{y} = -58{,}1 + 2{,}15\,x.$$

X и Y.

Составим уравнения линейной регрессии X и Y используя формулы: $d = r \cdot \dfrac{\sigma_x}{\sigma_y}$ и $\hat{x} - \bar{x} = d\left(y - \bar{y}\right)$.

$$d = r \cdot \frac{\sigma_x}{\sigma_y} = 0{,}97 \cdot \frac{14{,}36}{31{,}85} = 0{,}44$$

$$\bar{x}_y - 52{,}5 = 0{,}44\left(y - 54{,}4\right) \text{ или } \bar{x}_y = 0{,}44\,y + 28{,}56.$$

3. Построим корреляционное поле и графики прямых линий регрессии Y на X и X на Y (рис. 10). Из чертежа видно, что полученные уравнения хорошо согласуются с исходными данными.

4. По вычисленным коэффициентам, можно сделать вывод, что связь между температурой реакции и выходом продукта прямая и очень тесная, так как полученный коэффициент корреляции ($r = 0{,}97$) положительный и очень близок к единице. Это говорит о том, что чем больше температура реакции (X), тем больше выход продукта (Y).

Выясним, какая часть вариации Y обусловлена вариацией X, для этого вычислим коэффициент детерминации:

$$\eta = r^2 = (0{,}97)^2 = 0{,}94.$$

То есть вариация выхода продукта (Y) на 94% обусловлена вариацией температурой реакции (X).

Положительный коэффициент регрессии $b = 2{,}15$ подтверждает то, что связь между температурой реакции и выходом продукта прямая. Вычислим коэффициент эластичности (регрессии):

$$Э_x = b\frac{\overline{x}}{\overline{y}} = 2{,}15 \cdot \frac{52{,}5}{54{,}4} \approx 2{,}1\% .$$

Полученный коэффициент свидетельствует о том, что при увеличении температуры реакции на 1%, выход продукта в среднем увеличится на 2,1 %.

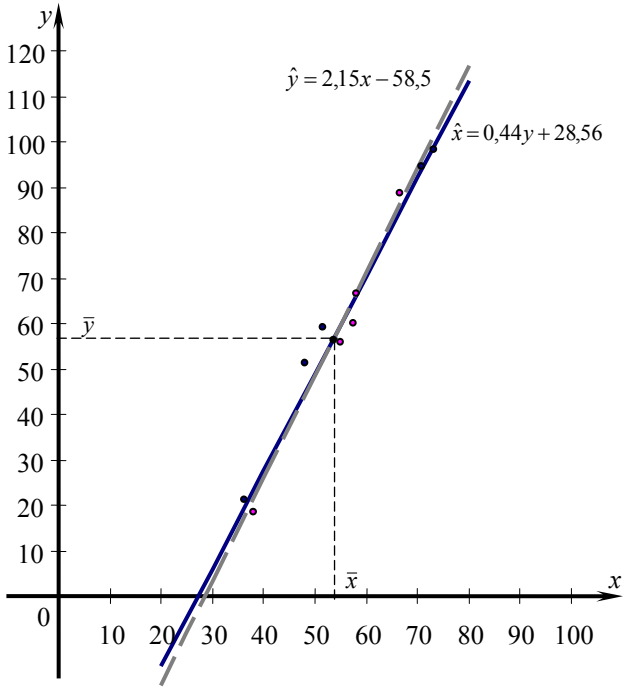

Рис. 6. Корреляционное поле. Линии регрессии

Спрогнозируем выход продукта при $x = 79°\,\text{C}$. Так как при увеличении температуры реакции на 1%, выход продукта в среднем увеличится на 2,1 %, то увеличение температуры реакции до $x = 79°\,\text{C}$ (т.е. примерно на 5,3 %) должно привести к увеличению выхода продукта примерно на 11,1 % (т.е примерно на 11,1 кг/ч).

Подставляя в уравнение регрессии $\hat{y} = -58{,}1 + 2{,}15x$ значение $x = 79$, получим $\hat{y} = 111{,}75$, т.е. при температуре реакции $x = 79°\,\text{C}$ получим выход продукта 111,75 кг/ч, т.е на 11,75 кг/ч больше.

Рекомендательный библиографический список

Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1979.-400 с.

Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением EXCEL: Учебное пособие для вузов.– Ростов н/Д, Феникс, 2005.-112 с.

Господариков А.П., Ивакин В.В., Лебедев И.А., Зацепин М.А. Высшая математика. Теория вероятностей и основы математической статистики. Учебное пособие.- Горный университет, 2013.-52 с.

Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебное пособие для вузов – М.: Юнити-Дана, 2000.-543 с.

Палий И.А. Прикладная статистика: Учебное пособие для вузов – М.: Высшая школа, 2004.–176 с.

СОДЕРЖАНИЕ

МАТЕМАТИКА

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Методические указания для выполнения расчетных заданий

Сост.: *Л.В. Бакеева, Е.В. Пастухова*

Печатается с оригинал-макета, подготовленного кафедрой высшей математики

Ответственный за выпуск *Е.В. Пастухова*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 10.06.2019. Формат 60×84/16.
Усл. печ. л. 2,4. Усл.кр.-отт. 2,4. Уч.-изд.л. 2,0. Тираж 100 экз. Заказ 541. С 197.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2