ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Методические указания к самостоятельным работам для студентов бакалавриата направлений 12.03.01, 15.03.04, 23.03.01 и 27.03.03

САНКТ-ПЕТЕРБУРГ
2021

Министерство науки и высшего образования Российской Федерации

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

Кафедра высшей математики

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

*Методические указания к самостоятельным работам
для студентов бакалавриата направлений
12.03.01, 15.03.04, 23.03.01 и 27.03.03*

САНКТ-ПЕТЕРБУРГ
2021

УДК 517.1+517.2(073)

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ: Методические указания к самостоятельным работам / Санкт-Петербургский горный университет. Сост.: *Л.В. Бакеева*, *Е.В. Пастухова, Е.Г. Булдакова*. СПб, 2021. 52 с.

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

Методические указания содержат основные теоретические сведения корреляционно-регрессионого анализа и некоторые критерии проверки статистической значимости его результатов, типовой пример, иллюстрирующий основные этапы обработки экспериментальных данных, задания для самостоятельной работы.

Могут быть использованы для самостоятельной работы в соответствии с программами подготовки бакалавров по направлениям 12.03.01 «Приборостроение», 15.03.04 «Машиностроение», 23.03.01 «Технология транспортных процессов» и 27.03.03 «Системный анализ и управление» по дисциплине «Теория вероятностей и математическая статистика».

Научный редактор проф. *А.П. Господариков*

Рецензент д.ф-м.н. *С.И. Перегудин* (СПбГУ)

ВВЕДЕНИЕ

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

Методические указания содержат основные понятия математической статистики, в них излагаются основы и методы корреляционно-регрессионного анализа. Изложение теоретического материала сопровождается разобранными типовыми примерами.

Цель предлагаемых методических указаний – помочь студенту приобрести навыки применения математической статистики к решению прикладных задач.

Методические указания могут быть использованы для выполнения обучающимися заданий самостоятельной работы в соответствии с программами подготовки бакалавров по направлениям подготовки 12.03.01 «Приборостроение», 15.03.04 «Машиностроение», 23.03.01 «Технология транспортных процессов» и 27.03.03 «Системный анализ и управление» по дисциплине «Теория вероятностей и математическая статистика».

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.

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.

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*.

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 ис-

следования 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.

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. Привести пример выборки – результаты тестирования 10 абитуриентов.

Решение. 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

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

Перечень вариант 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 $x_{(1)}$, $x_{(2)},\ldots, x_{(10)}$, 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, табл. 1 и 2):

Таблица 1

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

Таблица 2

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 записываются 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 (табл. 3):

Таблица 3

$(x_i; x_{i+1}]$	$(150;156]$	$(156;162]$	$(162;168]$	$(168;174]$	$(174;180]$	$(180;186$
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.

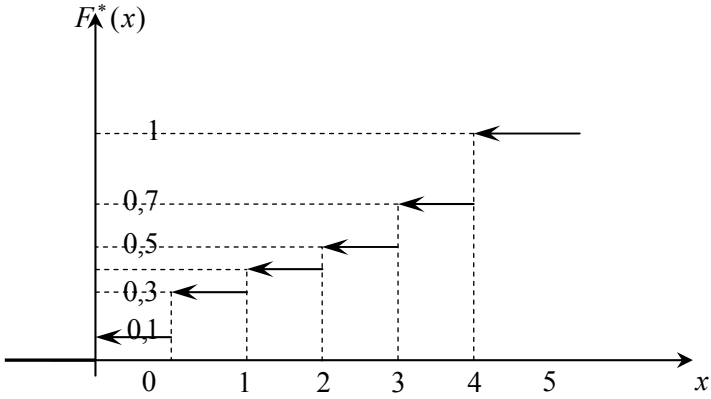

Рис. 1. Эмпирическая функция распределения дискретной случайной величины

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» (табл. 4):

Таблица 4

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):

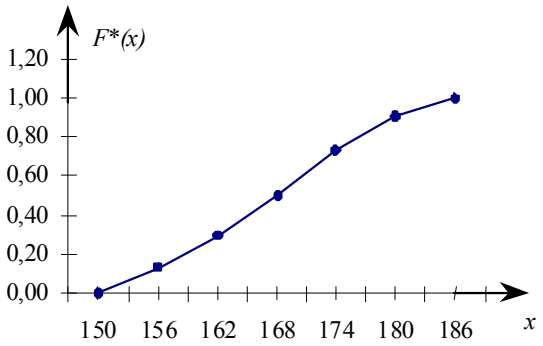

Рис. 2. Эмпирическая функция распределения непрерывной случайной величины

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 статистического 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 со-

ответственно частости – на оси ординат.

Пример 6. Пусть дана выборка в виде распределения частот (табл.5):

Таблица 5

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 (табл. 6) (см. пример 2):

Таблица 6

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:

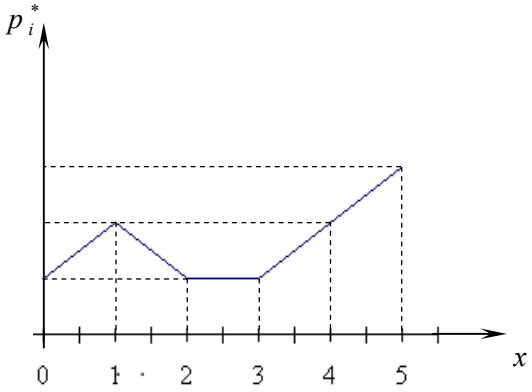

Рис.3. Полигон частостей

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 7.

Таблица 7

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:

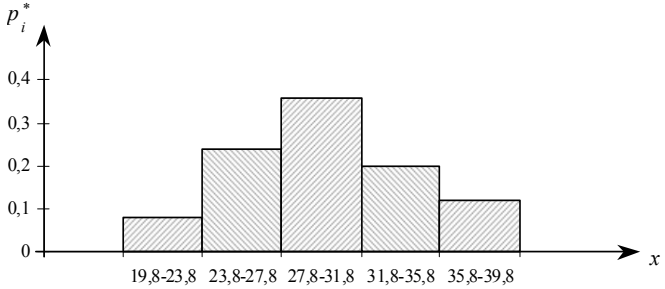

Рис. 4. Гистограмма частостей

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:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i , \qquad (1.1)$$

где x_i – варианта выборки; n_i – частота варианты; n – объем выборки.

Выборочную среднюю можно записать и так:

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

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 $\bar{x}_в$:

$$D = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 \cdot n_i , \qquad (1.2)$$

или, что то же самое,

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

Для расчетов может быть использована также формула:

$$D = \overline{x^2} - (\bar{x})^2, \qquad (1.3)$$

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

Выборочное *среднее квадратическое отклонение* выборки определяется формулой:

$$\sigma = \sqrt{D} \qquad (1.4)$$

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.

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 (табл. 8):

Таблица 8

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 по формуле (1.1):

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

Для вычисления выборочной дисперсии используем формулу (1.3). 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 $D = 5{,}34 - (2{,}31)^2 = 0{,}039$.

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 Мо вариационного ряда называется варианта, имеющая наибольшую частоту.

Медианой Ме 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 (табл. 9):

Таблица 9

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 по формулам 1.1-1.3:

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

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

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

$$\sigma = \sqrt{D} = \sqrt{3,2} \approx 1,79,$$

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

$Mo = 5$, 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 ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ. ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ И ЕГО ПАРАМЕТРОВ

Парная (простая) линейная регрессия представляет собой математическую модель, описывающую среднее значение зависимой (объясняемой) переменной как функцию одной независимой (объясняющей) переменной x, т.е. это модель вида:

$$\hat{y} = f(x),$$

где $\hat{y}$ – зависимая переменная (результативный признак); x – независимая, объясняющая переменная (признак-фактор). Знак «^» означает, что между переменными x и y нет в общем случае строгой функциональной зависимости. Практически в каждом отдельном случае величина y представляет собой сумму двух слагаемых:

$$y = \hat{y} + \varepsilon,$$

где y – фактическое значение результативного признака; $\hat{y}$ – теоретическое значение результативного признака, найденное исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонение реальных значений результативного признака от теоретических, найденных по уравнению регрессии.

Случайная величина ε включает влияние неучтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

Различают линейные и нелинейные регрессии.

Линейная регрессия: $y = a + bx + \varepsilon$.

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам.

Регрессии, *нелинейные по объясняющим переменным*:

1. Полиномы разных степеней:

$$y = a + b_1 x + b_2 x^2 + .. + b_n x^n + \varepsilon .$$

2.Равносторонняя гипербола: $y = a + \dfrac{b}{x} + \varepsilon$.

Регрессии, *нелинейные по оцениваемым параметрам*:

1. Степенная $y = ax^b + \varepsilon$;

2. Показательная $y = ab^x + \varepsilon$;

3. Экспоненциальная $y = e^{a+bx} + \varepsilon$.

Построение линейной регрессии сводится к оценке ее параметров a и b. Подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических $\hat{y}$ минимальна:

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

Для линейных и нелинейных уравнений, приводимых к линейным, решается система относительно параметров a и b:

$$\begin{cases} a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i{}^2=\sum_{i=1}^{n}x_iy_i, \\ an+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i. \end{cases} \tag{2.1}$$

Решая систему уравнений, найдем искомые оценки параметров a и b. Можно воспользоваться следующими известными формулами, следующие непосредственно из решения системы:

$$a=\bar{y}-b\bar{x},\ b=\frac{\operatorname{cov}(x,y)}{\sigma_x^2}, \tag{2.2}$$

где $\operatorname{cov}(x,y)=\overline{xy}-\bar{x}\cdot\bar{y}$ – ковариация признаков x и y; $\sigma_x^2=\overline{x^2}-\bar{x}^2$ - дисперсия признака x;

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\ ;\ \ \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i\ ;\ \ \overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i\ ;\ \ \overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2\ . \tag{2.3}$$

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. *Дисперсия* – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. *Математическое ожидание* – сумма произведений значений случайной величины на соответствующие вероятности.

Тесноту связи изучаемых явлений оценивает *линейный коэффициент парной корреляции* r_{xy}. Для линейной регрессии $(-1 \le r_{xy} \le 1)$ он равен:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}; \qquad (2.4)$$

индекс корреляции ρ_{xy} для нелинейной регрессии $(0 \le \rho_{xy} \le 1)$:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \qquad (2.5)$$

где $\sigma^2_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$ – общая дисперсия результативного признака *y*;

$\sigma^2_{ост} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – остаточная дисперсия, определяемая из уравнения регрессии $\hat{y} = f(x)$.

Оценку качества построенной модели дают *коэффициент детерминации* и *средняя ошибка аппроксимации*.

Коэффициент детерминации:

1) для линейной регрессии равен r^2_{xy}, где r_{xy} коэффициент корреляции и вычисляется по формуле 2.4;

2) для нелинейной регрессии равен ρ^2_{xy}, где ρ_{xy} индекс корреляции и вычисляется по формуле 2.5.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических, определяется по формуле:

$$\overline{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%. \qquad (2.6)$$

Значение средней ошибки аппроксимации до 15% свидетельствует о хорошо подобранной математической модели.

После того как найдено уравнение линейной регрессии, проводится *оценка значимости* как уравнения в целом, так и отдельных его параметров. Уравнение регрессии строится не по генеральной совокупности, которая неизвестна, а по выборке из нее. Точки из генеральной совокупности попадают в выборку случайным образом, поэтому в соответствии с теорией вероятностей возможен вариант, когда выборка из «широкой» генеральной совокупности окажется «узкой». В этом случае:

а) уравнение регрессии, построенное по выборке, может значительно отличаться от уравнения регрессии для генеральной совокупности, что приводит к ошибке прогноза;

б) коэффициент детерминации и другие характеристики точности окажутся неоправданно высокими, вводящими в заблуждение о прогнозных качествах уравнения регрессии.

В условиях отсутствия информации обо всех точках генеральной совокупности единственный способ уменьшить ошибки в первом случае заключается в использовании при оценке коэффициентов уравнения регрессии метода, обеспечивающего их несмещенность и эффективность. Вероятность наступления второго случая может быть значительно снижена благодаря априори известному свойству генеральной совокупности с двумя независимыми друг от друга переменными – в ней отсутствует именно эта связь. Достигается такое снижение за счет проверки *статистической значимости* полученного уравнения регрессии.

Оценка значимости уравнения регрессии в целом проводится на основе *F-критерия Фишера*, базирующегося на применении методов дисперсионного анализа. Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения $\overline{y}$ раскладывается на две части – *«объясненную»* и *«необъясненную»*:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 ,$$

где $\sum_{i=1}^{n}(y_i-\bar{y})^2$ – общая сумма квадратов отклонений; $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ – сумма квадратов отклонений, обьясненная регрессией (или факторная сумма квадратов отклонений); $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 10;

Таблица 10

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	$S^2_{общ}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
Факторная	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	m	$S^2_{факт}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{m}$
Остаточная	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-m-1$	$S^2_{ост}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-m-1}$

Вычисление компонент дисперсии на одну степень свободы приводит дисперсии к сравнимому виду (степени свободы – это числа, показывающие количество элементов варьирования, которые могут принимать произвольные значения, не имеющие заданных характеристик). Сопоставляя факторную и остаточную компоненты дисперсии в расчете на одну степень свободы, получим величину *F*-критерия Фишера:

$$F = \frac{S^2_{факт}}{S^2_{ост}}. \qquad (2.7)$$

Фактическое значение *F*-критерия Фишера сравнивается с табличным значением $F_{табл}(\alpha; k_1; k_2)$ при уровне значимости α и степенях свободы $k_1 = m, k_2 = n - m - 1$. При этом, если фактическое значение *F*-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии $m = 1$, поэтому имеем:

$$F = \frac{S^2_{факт}}{S^2_{ост}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} (n-2). \qquad (2.8)$$

Величина *F*-критерия Фишера связана с коэффициентом детерминации r^2_{xy}, и ее можно рассчитать по формуле:

$$F = \frac{r^2_{xy}}{1 - r^2_{xy}} (n-2). \qquad (2.9)$$

Для оценки *статистической значимости параметров регрессии и корреляции* рассчитываются *t-критерий Стьюдента* и *доверительные интервалы* каждого из показателей. Оценка значимости коэффициентов регрессии и корреляции с помощью *t*-критерия Стьюдента проводится путем сопоставления значений этих коэффициентов с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; t_a = \frac{a}{m_a}; t_r = \frac{r_{xy}}{m_r}. \qquad (2.10)$$

Стандартные ошибки параметров линейной регрессии и коэффициента корреляции определяются по формулам:

$$m_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{S_{ост}^2}{n\sigma_x^2}}; \qquad (2.11)$$

$$m_a = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \cdot \frac{\sum_{i=1}^{n} x_i^{\,2}}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{S_{ост}^2 \frac{\sum_{i=1}^{n} x_i^{\,2}}{n^2\sigma_x^2}}; \qquad (2.12)$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}}. \qquad (2.13)$$

Сравнивая фактическое $t_{факт}$ и критическое (табличное) $t_{табл}$ значения t-критерия Стьюдента, делаем вывод о значимости параметров регрессии и корреляции. Если $t_{табл} < t_{факт}$, то a, b, r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x. Если $t_{табл} > t_{факт}$, то признается случайная природа формирования параметров a, b или r_{xy}.

Для расчета доверительного интервала определяем *предельную ошибку* Δ для каждого показателя:

$$\Delta_a = t_{табл} m_a,\ \Delta_b = t_{табл} m_b.$$

Формулы для расчета *доверительных интервалов* имеют следующий вид:

$$\gamma_a = a \pm \Delta_a,\ \gamma_{a_{\min}} = a - \Delta_a,\ \gamma_{a_{\max}} = a + \Delta_a,$$
$$\gamma_b = b \pm \Delta_b,\ \gamma_{b_{\min}} = b - \Delta_b,\ \gamma_{b_{\max}} = b + \Delta_b. \qquad (2.14)$$

Если в границы доверительного интервала попадает нуль, (нижняя граница отрицательна, а верхняя положительна), то оцени-

ваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Связь между *F*-критерием Фишера и *t*-критерием Стьюдента выражается равенством:

$$|t_r| = |t_b| = \sqrt{F} .$$

Важной предпосылкой построения качественной регрессионной модели по МНК является независимость значений случайных отклонений от значений отклонений во всех других наблюдениях. Это гарантирует отсутствие *коррелированности между любыми отклонениями* и, в частности, между соседними. *Автокорреляция (последовательная корреляция)* определяется как корреляция между наблюдаемыми показателями, упорядоченными во времени. Существует два метода обнаружения автокорреляции.

а. Графический метод.

Задан ряд вариантов графического определения автокорреляции. Один из вариантов увязывает отклонения e_i с моментами их получения i. При этом по оси абсцисс откладывают либо время получения статистических данных, либо порядковый номер наблюдения, а по оси ординат – значения отклонений e_i (либо оценки отклонений).

Предполагая, что если имеется определенная связь между отклонениями, то автокорреляция имеет место. Отсутствие зависимости скоре всего будет свидетельствовать об отсутствии автокорреляции. Автокорреляция становится более наглядной, если построить график зависимости e_i от e_{i-1}.

б. Критерий Дарбина-Уотсона.

Этот критерий является наиболее известным для обнаружения автокорреляции. При статистическом анализе уравнения регрессии на начальном этапе часто проверяют выполнимость одной предпосылки: условия статистической независимости отклонений между собой. При этом проверяется некоррелированность соседних величин.

Для анализа коррелированности отклонений используют статистику Дарбина-Уотсона:

$$DW = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (2.15)$$

Критические значения d_1 и d_2 определяются на основе специальных таблиц для требуемого уровня значимости α, числа наблюдений n и количества объясняющих переменных $m = 1$.

Автокорреляция отсутствует, если выполняется следующие условия:

$$\begin{cases} d_1 < DW, \\ DW < 4 - d_2. \end{cases}$$

Близость DW к нулю говорит о положительной автокорреляции, а к 4 – об отрицательной автокорреляции.

Не обращаясь к таблицам, можно пользоваться приближенным правилом: считать, что автокорреляция остатков отсутствует, если $1{,}5 < DW < 2{,}5$. Для более надежного вывода целесообразно обращаться к табличным значениям.

Объяснение наличия автокорреляции в остатках может быть в том, что не учтен некий фактор, имеющий сильное влияние на объясняемую переменную. Поскольку он не учтен, его влияние остается в необъясненной части – остатках, что и становится заметным при оценке автокорреляции остатков.

2.2. Линеаризация данных

Физические величины, определяющие результаты эксперимента, выступают в роли переменных и параметров некоторой функциональной зависимости, теоретически получаемой в рамках модели. После экспериментальной регистрации зависимости ее сравнивают с теоретической. Методом сравнения можно не только численно определить, т.е. измерить, значения физических величин,

не измеряемых другим способом, но и сделать заключение об адекватности принятой модели эксперимента.

Проще всего проверить линейную зависимость. Если зависимость нелинейная, в некоторых случаях ее можно преобразовать в линейную, т.е. выполнить линеаризацию данных (табл. 11).

Таблица 11

Вид нелинейной зависимости	Вид получаемой линейной зависимости	y	x	a	b
$v = k \cdot u^z$	$\ln v = z \ln u + \ln k$	$\ln v$	$\ln u$	z	$\ln k$
$v = k \cdot e^{zu}$	$\ln v = zu + \ln k$	$\ln v$	u	z	$\ln k$
$v = k \cdot e^{\frac{z}{u}}$	$\ln v = \frac{z}{u} + \ln k$	$\ln v$	$\frac{1}{u}$	z	$\ln k$
$v = \frac{u}{k + zu}$	$\frac{1}{v} = \frac{k}{u} + z$	$\frac{1}{v}$	$\frac{1}{u}$	k	z

3. РЕШЕНИЕ ТИПОВЫХ ЗАДАЧ

Задача. Аналитическое агенство интересует зависимость кадастровой стоимости жилья y от инженерно-геологических условий строительства и степени подверженности территории разрушительным воздействиям природы x. Для выяснения характера этой связи было отобрано 15 городов. Приведены данные (табл. 12) о кадастровой стоимостьи жилья (тыс. руб./м^2) и инженерно-геологических условиях строительства и степени подверженности территории разрушительным воздействиям природы (баллы, из 100 возможных). Для приведенных данных:

1. Построить линейное уравнение парной регрессии y по x.

2. Рассчитать линейный коэффициент парной корреляции, коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F-критерия Фишера и t-критерия Стьюдента.

4. Провести анализ коррелированности отклонений.

5. Построить степенную и показательную модели парной регрессии y по x и сравнить их с линейной моделью.

Таблица 12

x	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	13	16	15	20	19	21	26	24	30	32	30	35	34	40	39

Решение.

1. Построим вспомогательную таблицу для расчета параметров уравнения линейной регрессии (табл. 13).

Таблица 13

№	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	A_i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	6	13	78	36	169	12,76	0,058	0,018
2	7	16	112	49	256	14,69	1,716	0,081
3	8	15	120	64	225	16,62	2,624	0,108
4	9	20	180	81	400	18,55	2,103	0,072
5	10	19	190	100	361	20,48	2,190	0,077
6	11	21	231	121	441	22,41	1,988	0,067
7	12	26	312	144	676	24,34	2,756	0,063
8	13	24	312	169	576	26,27	5,153	0,094

№	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	A_i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
9	14	30	420	196	900	28,2	3,24	0,06
10	15	32	480	225	1024	30,13	3,497	0,058
11	16	30	480	256	900	32,06	4,244	0,068
12	17	35	595	289	1225	33,99	1,020	0,028
13	18	34	612	324	1156	35,92	3,686	0,056
14	19	40	760	361	1600	37,85	4,623	0,053
15	20	39	780	400	1521	39,78	0,608	0,02
Итого	195	394	5662	2815	11430	394,05	39,51	0,930
Сред. знач.	13	26,267	377,4	187,67	762	26,27	2,63	0,062
σ	2,94	8,49						
σ^2	18,67	72,04						

По формулам (2.2) находим параметры регрессии:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{377,47 - 13 \cdot 26,267}{187,67 - 13^2} = 1,93;$$

$$a = \bar{y} - b\bar{x} = 26,267 - 1,93 \cdot 13 = 1,18.$$

Получим уравнение регрессии: $\hat{y} = 1,93x + 1,18.$

С помощью уравнения регрессии заполним столбцы 7–9 табл. 4.

2. Вычислим коэффициент корреляции:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 1,93 \cdot \frac{4,32}{8,49} = 0,98.$$

Т.к. значение коэффициента корреляции больше 0,7, то это говорит о наличии тесной линейной связи между признаками.

Коэффициент детерминации: $r_{xy}^2 = 0,96$. Это означает, что 96% вариации кадастровой стоимости жилья *y* объясняется вариацией фактора *x* - инженерно-геологическими условиями строительства и степенью подверженности территории разрушительным воздействиям природы.

Качество модели определяет средняя ошибка аппроксимации:

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} A_i \cdot 100\% = 0{,}062 \cdot 100\% = 6{,}2\% .$$

Качество построенной модели оценивается как хорошее, так как $\overline{A}$ не превышает 10%.

3. Проведем оценку статистической значимости уравнения регрессии в целом с помощью *F*-критерия Фишера. Определим фактическое значение *F*-критерия:

$$F_{факт} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2) = \frac{0{,}96}{1-0{,}96}(15-2) = 312 .$$

Табличное значение критерия при пятипроцентном уровне значимости и степенях свободы $k_1 = 1$, $k_2 = 13$ составляет $F_{табл} = 4{,}67$. Так как $F_{факт} > F_{табл}$, то уравнение регрессии признается статистически значимым.

Проведем оценку статистической значимости параметров регрессии и корреляции с помощью *t*-критерия Стьюдента и определением доверительного интервала каждого из параметров.

Табличное значение *t*-критерия Стьюдента для числа степеней $df = n - 2 = 15 - 2 = 13$ и уровня значимости $\alpha = 0{,}05$ составит $t_{табл} = 2{,}16$.

С учетом значения остаточной дисперсии

$$S_{ост}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - m - 1} = \frac{39{,}51}{13} = 3{,}039 ,$$

по формулам (2.11–2.13) определим стандартные ошибки m_a, m_b, $m_{r_{xy}}$ на одну степень свободы:

$$m_a = \sqrt{S_{ост}^2 \frac{\sum_{i=1}^{n} x_i^2}{n^2 \sigma_x^2}} = \sqrt{3{,}039 \cdot \frac{2815}{15^2 \cdot 8{,}67}} = 2{,}094;$$

$$m_b = \sqrt{\frac{S_{ост}^2}{n\sigma_x^2}} = \sqrt{\frac{3{,}039}{15 \cdot 8{,}67}} = 0{,}15;$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}} = \sqrt{\frac{1 - 0{,}96}{15-2}} = 0{,}055.$$

Далее вычислим значения *t*-критериев Стьюдента:

$$t_a = \frac{a}{m_a} = \frac{1{,}18}{2{,}094} = 0{,}56,$$

$$t_b = \frac{b}{m_b} = \frac{1{,}93}{0{,}15} = 12{,}87,$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0{,}98}{0{,}055} = 17{,}818.$$

Фактические значения *t*-критериев для t_b и $t_{r_{xy}}$ превосходят табличные значения:

$$t_b = 12{,}87 > t_{табл} = 2{,}16;\ t_{r_{xy}} = 7{,}818 > t_{табл} = 2{,}16.$$

Таким образом, параметры *b* и r_{xy} неслучайно отличаются от нуля, т.е. статистически значимы.

Фактическое значение *t*-критерия для t_a не превосходит табличного значения:

$$t_a = 0{,}564 < t_{табл} = 2{,}16.$$

Следовательно, коэффициент a незначим. Отсутствие значимости коэффициента в модели описания говорит о целесообразности исключения соответствующего слагаемого из уравнения.

Замечание. Незначимые коэффициенты должны быть оценены, исходя из существа изучаемого метода обработки, так как незначимость коэффициента может определяться малым интервалом варьирования изучаемого фактора, но не отсутствием его влияния на функцию отклика.

Определим доверительные интервалы для параметров регрессии a и b по формулам (2.14). Для этого вычислим предельную ошибку для каждого показателя:

$$\Delta_a = t_{табл} \cdot m_a = 2{,}16 \cdot 2{,}094 = 4{,}52 ;$$

$$\Delta_b = t_{табл} \cdot m_b = 2{,}16 \cdot 0{,}15 = 0{,}324 .$$

Доверительные интервалы:

$$\gamma_a = 1{,}18 \pm 4{,}48; \ -3.3 < \gamma_a < 5{,}66;$$

$$\gamma_b = 1{,}93 \pm 0{,}324; \ 1{,}606 < \gamma_b < 2{,}254 .$$

Анализ верхней и нижней границ доверительных интервалов приводит к выводу о том, что с вероятностью $p = 1 - \alpha = 0{,}95$ параметр b, находясь в указанных границах, не принимает нулевых значений, т.е. является статистически значимым и существенно отличается от нуля.

Параметр a, находясь в указанных границах, принимает нулевое значение, т.е. является статистически незначимым, что и было установлено выше.

Уравнение регрессии окончательно имеет вид:

$$\hat{y} = 1{,}93x.$$

4. Для анализа коррелированности отклонений используем статистику Дарбина-Уотсона (2.15):

Составим вспомогательную таблицу (табл. 14).

Таблица 14

№	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	$e_i^2 = (y_i - \hat{y}_i)^2$	$(e_i - e_{i-1})^2$
1	**2**	**3**	**4**	**5**	**6**
1	13	11,58	1,42	2,0164	0
2	16	13,51	2,49	6,2001	1,1449
3	15	15,44	-0,44	0,1936	8,5849
4	20	17,37	2,63	6,9169	9,4249
5	19	19,3	-0,3	0,09	8,5849
6	21	21,23	-0,23	0,0529	0,0049
7	26	23,16	2,84	8,0656	9,4249
8	24	25,09	-1,09	1,1881	15,4449
9	30	27,02	2,98	8,8804	16,5649
10	32	28,95	3,05	9,3025	0,0049
11	30	30,88	-0,88	0,7744	15,4449
12	35	32,81	2,19	4,7961	9,4249
13	34	34,74	-0,74	0,5476	8,5849
14	40	36,67	3,33	11,0889	16,5649
15	39	38,6	0,4	0,16	8,5849
Итого	394			60,2735	127,788

$$DW = \frac{127,788}{60,2735} = 2,12 .$$

Воспользуемся приближенным правилом. Т.к. $1,5 < DW = 2,12 < 2,5$, то автокорреляции остатков нет.

Построим средствами MS Excel по столбцу 4 график остатков (рис. 5)

Используя график остатков (рис.5), можно говорить об отсутствии зависимости между e_i и x, и об отсутствии автокорреляции.

Рис.5 График остатков

5. Уравнение *степенной модели* имеет вид:

$$y = ax^b \qquad (3.1)$$

Для принятой модели необходимо произвести линеаризацию переменных. Для этого произведем логарифмирование обеих частей уравнения (3.1):

$$\ln y = \ln a + b \ln x .$$

Обозначим $A = \ln a$, данные для вычисления параметров приведены в таблице (табл. 15).

$$b = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{8{,}168 - 2{,}503 \cdot 3{,}211}{6{,}398 - 2{,}503^2} = 0{,}985 ;$$

$$A = \overline{Y} - b\overline{X} = 3{,}211 - 0{,}985 \cdot 2{,}503 = 0{,}746 .$$

Уравнение регрессии будет иметь вид:

$$Y = 0{,}985X + 0{,}746 .$$

Перейдем к исходным переменным x и y, выполнив потенцирование последнего уравнения:

$$\hat{y} = e^{0{,}746} \cdot x^{0{,}985} .$$

Окончательно получим уравнение степенной модели регрессии (3.1):

$$\hat{y} = 2{,}109 \cdot x^{0{,}985}.$$

Таблица 15

№	x_i	y_i	$Y_i = \ln y_i$	$X_i = \ln x_i$	$X_i Y_i$	X_i^2	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	6	13	2,564	1,791	4,596	3,210	12,318	0,465
2	7	16	2,772	1,946	5,395	3,787	14,338	2,761
3	8	15	2,708	2,079	5,631	4,324	16,353	1,833
4	9	20	2,995	2,197	6,582	4,827	18,365	2,671
5	10	19	2,944	2,303	6,779	5,302	20,374	1,888
6	11	21	3,044	2,398	7,300	5,750	22,379	1,903
7	12	26	3,258	2,485	8,096	6,175	24,382	2,618
8	13	24	3,178	2,565	8,152	6,579	26,382	5,675
9	14	30	3,401	2,639	8,976	6,965	28,380	2,624
10	15	32	3,465	2,708	9,385	7,334	30,375	2,638
11	16	30	3,401	2,773	9,430	7,687	32,369	5,614
12	17	35	3,555	2,833	10,073	8,027	34,361	0,408
13	18	34	3,526	2,890	10,192	8,354	36,351	5,529
14	19	40	3,688	2,944	10,861	8,669	38,339	2,757
15	20	39	3,663	2,996	10,975	8,974	40,326	1,758
Ито-го	195	394	48,169	37,548	122,426	95,965		41,143
Сред. знач.	13	26,26	3,211	2,503	8,162	6,398		
σ	2,94	8,49						
σ^2	18,67	72,04						

Определим индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{1 - \frac{41{,}143}{1080{,}933}} = 0{,}98.$$

Тогда коэффициент детерминации равен $\rho_{xy}^2 = 0{,}96$.

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\% = \frac{1}{15} \cdot 0{,}974 \cdot 100\% = 6{,}49\% .$$

Качество построенной модели оценивается как хорошее, так как $\overline{A}$ не превышает 10%.

6. Уравнение *показательной модели* имеет вид:

$$y = a \cdot b^x . \qquad (3.2)$$

Для построения этой модели необходимо произвести линеаризацию переменных. Для этого произведем логарифмирование обеих частей уравнения (3.2):

$$\ln y = \ln a + x \ln b .$$

Данные для вычисления параметров уравнения приведены в таблице (табл. 16).

Таблица 16

№	x_i	y_i	$Y_i = \ln y_i$	$x_i Y_i$	x_i^2	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	**2**	**3**	**4**	**5**	**6**	**8**	**9**
1	6	13	2,564	15,384	36	14,234	1,523
2	7	16	2,772	19,404	49	15,373	0,393
3	8	15	2,708	21,664	64	16,603	2,569
4	9	20	2,995	26,955	81	17,931	4,280
5	10	19	2,944	29,44	100	19,366	0,134
6	11	21	3,044	33,484	121	20,915	0,007
7	12	26	3,258	39,096	144	22,588	11,642
8	13	24	3,178	41,314	169	24,395	0,156
9	14	30	3,401	47,614	196	26,347	13,347
10	15	32	3,465	51,975	225	28,454	12,572
11	16	30	3,401	54,416	256	30,731	0,534
12	17	35	3,555	60,435	289	33,189	3,279
13	18	34	3,526	63,468	324	35,844	3,401
14	19	40	3,688	70,072	361	38,712	1,659
15	20	39	3,663	73,26	400	41,809	7,889
Итого	195	394	48,169	647,981	2815		63,387
Сред. знач.	13	26,267	3,211	43,199	187,67		
σ	2,94	8,49					
σ^2	18,67	72,04					

$$\ln b = \frac{\overline{xY} - \bar{x} \cdot \bar{Y}}{\overline{X^2} - \bar{X}^2} = \frac{43{,}199 - 13 \cdot 3{,}211}{187{,}67 - 13^2} = 0{,}078;$$

$$\ln a = \bar{Y} - \ln b \cdot \bar{x} = 3{,}211 - 0{,}078 \cdot 13 = 2{,}194;$$

$$\ln y = 2{,}194 + x \cdot 0{,}078.$$

Уравнение показательной модели регрессии (3.2) имеет вид:

$$\hat{y} = 8{,}97 \cdot e^{0{,}078x}; \ \hat{y} = 8{,}97 \cdot 1{,}08^x.$$

Индекс корреляции равен:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \sqrt{1 - \frac{63{,}387}{1080{,}933}} = 0{,}94.$$

Тогда коэффициент детерминации равен $\rho_{xy}^2 = 0{,}88$;

$$\overline{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% = \frac{1}{15} \cdot 0{,}982 \cdot 100\% = 6{,}55\%.$$

Качество построенной модели хорошее ($\overline{A} < 10\%$).

4. 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. 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:

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

2. Построить линейное уравнение парной регрессии y по x.

3. Найти линейный коэффициент парной корреляции, коэффициент детерминации и среднюю ошибку аппроксимации.

4. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью *F*-критерия Фишера и *t*-критерия Стьюдента.

5. Провести анализ коррелированности отклонений.

6. Построить степенную и показательную модели парной регрессии y по x и сравнить их с линейной моделью.

Вариант 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

Вариант 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*

Вариант 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

Вариант 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

Вариант 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

Вариант 6.

В таблице приведены данные о показателях конкуренции компаний, выполняющих работы в сфере инженерно-геодезических изысканий и землеустроительных (кадастровых) работ x и средневзвешенные по частоте упоминания количества выполненных проектов 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

Вариант 7.

В таблице приведены данные с 2008 г. по 2017 г. по площадям рекультивированных земель в регионе (x, тыс. га) и площадям земель, использованных при строительных работах (y, тыс. га).

x_i	62	40	38	25	15	52	27	47	24	18
y_i	15,5	10,8	10,3	7,5	4,7	13,1	8,0	24,5	7,2	7,9

Вариант 8.

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

Вариант 9.

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

Вариант 10.

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

Вариант 11.

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

Вариант 12.

В таблице приведены данные по 10 микрорайонам города о кадастровой стоимостьи жилья (тыс. руб./м^2) и обеспеченности централизованным инженерным оборудованием и благоустройство территории и застройки (баллы, из 100 возможных).

x_i	45,8	45,0	44,7	56,7	41,7	59,1	50,4	38,3	46,1	63,3
y_i	65,1	64,5	54,6	56,1	45,3	64,5	65,0	58,0	63,5	55,2

Вариант 13.

В таблице приведены данные с 2010 г. по 2019 г. по городу о кадастровой стоимостьи жилья (тыс. руб./м^2) и исторической ценности застройки, эстетической и ландшафтной ценности территории (баллы, из 10 возможных).

x_i	65,1	60,5	57,6	56,1	54,3	54,5	55,0	57,0	58,2	58,2
y_i	3	2,5	2	2,5	3	3,5	4	4,5	5	5,5

Вариант 14.

В таблице приведены данные по 10 микрорайонам города о кадастровой стоимостьи жилья (тыс. руб./м^2) и состоянии окружающей среды, санитарных и микроклиматических условия (баллы, из 100 возможных).

x_i	70	69	68,5	67	66,5	65,5	65	63	60	53
y_i	84,0	77,0	75,0	69,2	60,8	51,5	52,2	51,5	50,0	47,3

Вариант 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	5,1	6,9	9,1	10,8	13,2	14,9	17,2	18,8	21,2	22,9

Вариант 16.

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

Вариант 17.

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

Вариант 18.

В таблице приведены данные по 10 микрорайонам города о кадастровой стоимостьи жилья (тыс. руб./м^2) и уровне развития сферы социального культурно-бытового обслуживания населения микрорайонного значения (баллы, из 100 возможных).

x_i	70	69	68,5	67	66,5	65,5	65	63	60	53
y_i	54,0	47,0	45,0	49,2	50,8	51,5	52,2	51,5	50,0	47,3

Вариант 19.

Компания, выполняющая работы в сфере инженерно-геодезических изысканий и землеустроительных (кадастровых) работ, имеет 10 филиалов. В таблице приведены показатели о годовом обороте (*x*, млн. руб.) и количестве выполненных проектов (*y*, га).

x_i	0,25	0,42	0,57	0,59	0,79	0,95	0,99	1,23	1,29	1,33
y_i	1,9	10,1	1,2	4,3	8,3	7,6	7,2	5,2	6,2	19,4

Вариант 20.

В таблице приведены данные по площадям рекультивированных земель 10 регионов (*x*, тыс. га) и площадям земель, импользованных для разработки месторождений полезных ископаемых (включая общераспространенные полезные ископаемые) (*y*, тыс. га).

x_i	50,2	2,5	19,8	28,3	21,0	9,1	26,0	36,4	10,8	39,9
y_i	25,1	1,01	12,0	15,3	14,1	4,2	15,9	23,9	4,4	13,4

Вариант 21.

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

Вариант 22.

В таблице приведены результаты 10 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

Вариант 23.

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

Вариант 24.

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

Вариант 25.

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

Вариант 26.

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

Вариант 27.

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

Вариант 28.

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

Вариант 29.

В таблице приведены данные по 10 населенным пунктам муниципального района о кадастровой стоимостьи жилья (тыс. руб./м^2) и доступности населения к центру населенного пункта, объектам культуры и быта (баллы, из 100 возможных).

x_i	45,8	45,0	44,7	56,7	41,7	59,1	50,4	38,3	46,1	63,3
y_i	76,1	64,6	64,6	76,1	43,3	69,9	69,0	38,0	69,1	65,2

Вариант 30.

В таблице приведены данные с 2008 г. по 2017 г. по городу о кадастровой стоимостьи жилья (тыс. руб./м^2) и доступности населения к центру, объектам культуры и быта (баллы, из 100 возможных).

x_i	68,1	64,5	59,7	52,2	45,9	47,8	58,2	57,6	65,4	54,9
y_i	75,1	64,5	54,6	56,1	45,3	69,5	65,0	58,0	69,5	55,2

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Айзек М.П.* Вычисления, графики и анализ данных в Excel 2013. Самоучитель / М.П. Айзек. СПб.: Наука и техника. 2015. 416 с.

2. *Бакеева Л.В.* Математика. Элементы математической статистики. Корреляционно-регрессионный анализ. Методические указания для выполнения расчетных заданий / Л.В. Бакеева, Е.В. Пастухова. Горный университет. 2019. 42 с.

3. . *Бакеева Л.В.* Прикладная математика. Этапы статистической обработки массивов экспериментальных данных. Методические указания для самостоятельной работы / Л.В. Бакеева, Е.Г. Булдакова. Горный университет, 2020. 32 с.

4. *Гмурман В.Е.* Теория вероятностей и математическая статистика: Учебник для прикладного бакалавриата / В.Е. Гмурман. Люберцы: Юрайт. 2016. 479 с.

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

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

7. *Кремер Н.Ш.* Теория вероятностей и математическая статистика: Учебник и практикум для академического бакалавриата / Н.Ш. Кремер. Люберцы: Юрайт. 2016. 514 с.

8. *Палий И.А.* Прикладная статистика: Учебное пособие для вузов / И.А. Палий. М.: Высшая школа. 2004. 176 с.

9. *Яворский В.А.* Планирование научного эксперимента и обработка экспериментальных данных: Методические указания к лабораторным работам / В.А. Яворский. М.: МФТИ. 2011. 45 с.

СОДЕРЖАНИЕ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Методические указания к самостоятельным работам для студентов бакалавриата направлений 12.03.01, 15.03.04, 23.03.01 и 27.03.03

Сост. *Л.В. Бакеева, Е.В. Пастухова, Е.Г. Булдакова*

Печатается с оригинал-макета, подготовленного кафедрой высшей математики

Ответственный за выпуск *Л.В. Бакеева*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 02.09.2021. Формат 60×84/16.
Усл. печ. л. 3,0. Усл.кр.-отт. 3,0. Уч.-изд.л. 2,7. Тираж 75 экз. Заказ 779.

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