МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Воронежский государственный технический университет»

Кафедра автоматизированных и вычислительных систем

54-2019

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к выполнению лабораторных работ
и контрольной работы
для студентов направления 09.03.01
«Информатика и вычислительная техника» (профиль
«Вычислительные машины, комплексы, системы и сети»)
очной и заочной форм обучения

Воронеж 2019

УДК 681.32
ББК 32.971.9

Составители: *канд. техн. наук Т.И. Сергеева,*
канд. техн. наук Г.В. Петрухнова

Обработка экспериментальных данных: к выполнению лабораторных работ и контрольной работы для студентов направления 09.03.01 «Информатика и вычислительная техника» (профиль «Вычислительные машины, комплексы, системы и сети») / ФГБОУ ВО «Воронежский государственный технический университет»; сост. Т. И. Сергеева, Г. В. Петрухнова.

Воронеж: Изд-во ВГТУ, 2019. 51 с.

Методические указания содержат теоретические и практические сведения, необходимые для проведения статистической обработки данных, представленных одной или несколькими переменными.

МУ подготовлены в электронном виде и содержатся в файле Методичка ОЭД.pdf.

Табл. 6. Ил. 35. Библиогр.: 3 назв.

Предназначены для студентов третьего курса очной формы обучения и студентов четвертого курса заочной формы обучения.

.

Рецензент д-р техн. наук, проф. Барабанов В.Ф.

Издается по решению учебно-методического совета Воронежского государственного технического университета

ВВЕДЕНИЕ

Методические указания содержат теоретические и практические сведения, необходимые для освоения технологий обработки экспериментальных данных, выраженных одной или несколькими переменными.

1. АНАЛИЗ ОДНОМЕРНЫХ ЧИСЛОВЫХ ДАННЫХ В EXCEL

Excel включает в себя **два инструмента анализа**, полезных **для оценки данных, выражающихся одной переменной.**

Инструмент анализа «**Описательная статистика**» позволяет измерить общую направленность, изменчивость и асимметрию данных.

Инструмент анализа «**Гистограмма**» позволяет составить таблицу распределения частот, интегральных частот и построить саму гистограмму.

Для Excel 2003, если команда **Анализ данных** отсутствует в меню **Сервис**, выбирают в меню **Сервис** пункт **Надстройки**, в списке надстроек ставят флажок у параметра **Анализ данных.**

Для Excel 2007, если на панели **Данные** отсутствует кнопка **Анализ данных**, то на названии панели **Данные** вызывают контекстное меню (щелкают правой кнопкой мыши), выбирают **Настройка панели быстрого доступа**, в открывшемся диалоговом окне слева на панели щелкают по слову **Надстройки**, внизу диалогового окна щелкают по кнопке **Перейти…,** в списке **Доступные надстройки** ставят флажок у параметра **Пакет анализа**, кнопка **ОК**.

1.1. Инструмент анализа: описательная статистика

Описательная статистика позволяет анализировать данные, выраженные одной переменной. Например, расход

топлива на изготовление 100 кг экспериментального сырья; средний объем продаж этого сырья различным компаниям, относящимся к одной группе; объем прибыли сети аптек и т.д. Описательная статистика измеряет общую направленность, изменчивость и асимметрию данных.

Инструмент анализа «**Описательная статистика**» **реализуется** следующим образом:

– на новом листе ввести данные в столбец А (например, о расходе топлива); можно в верхней ячейке А1 ввести название данных (например, Расход топлива на 100 кг сырья);

– выбрать Анализ данных (для Excel 2003 – Сервис, Анализ данных; для Excel 2007 – панель Данные, кнопка Анализ данных);

– в появившемся диалоговом окне из списка «Инструменты анализа» выбрать «Описательная статистика», ОК;

– появится диалоговое окно «Описательная статистика», представленное на рис. 1.

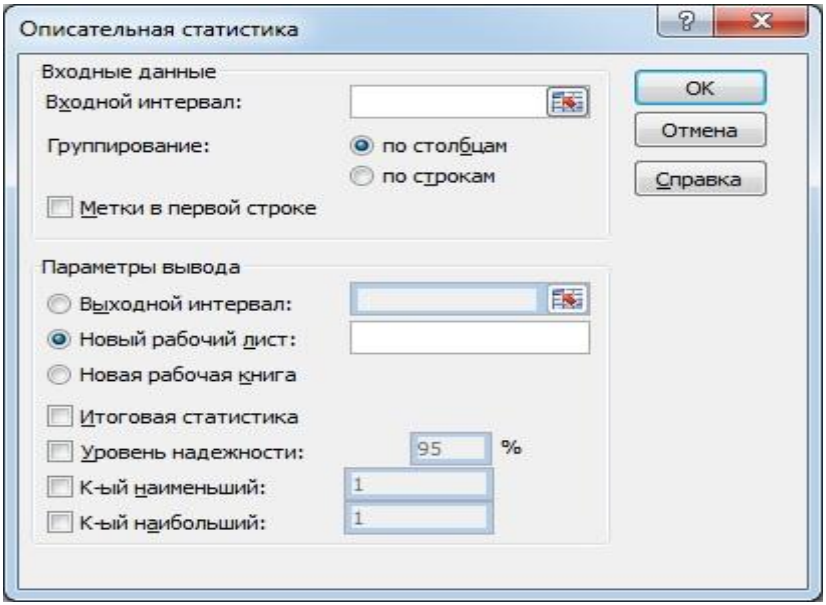

Рис. 1. Окно «Описательная статистика»

В окне «**Описательная статистика**» необходимо осуществить следующие действия:

– в поле «**Входной интервал**» указывают промежуток ячеек с исходными данными, не включая название к ним (если оно есть); выбор интервала лучше сделать с помощью мыши;

– **группирование** может осуществляться по столбцам или строкам в зависимости от расположения данных, обычно Excel сам правильно определяет параметр «Группирование»;

– параметр «**Метки в первой строке**» (или «**Метки в первом столбце**», если данные были расположены по строкам) активируют (выставляют флажок), если есть подпись у данных; если подписи нет, то флаг не выставляют;

– параметр «**Выходной интервал**» выбирают (щелчком), в адресной строке указывают левую верхнюю ячейку области листа для выходных данных; также есть возможность отправить выходные данные на новый лист в текущей книге или на новый лист новой книги;

– параметр «**Итоговая статистика**» следует включить (выставить флаг), иначе не будет вывода результатов;

– параметр «**Уровень надежности**»; включение данной опции позволяет вычислить половину длины доверительного интервала для среднего значения с заданной значимостью в процентах;

– параметр «**К-й наименьший**»; если отметить (выставить флаг) данный параметр, то будет определено k-е наименьшее значение данных, для этого также надо ввести значение k;

– параметр «**К-й наибольший**»; если отметить (выставить флаг) данный параметр, то будет определено k-е наибольшее значение данных, для этого также надо ввести значение k;

– после ввода всех значений нажимают ОК; Excel вычислит характеристики, относящиеся к описательной статистике, и разместит их в области вывода.

Форматирование таблиц с выходными данными. Для улучшения внешнего вида таблицы с результатами можно

осуществить следующее: увеличение ширины столбца, чтобы все данные были видны; уменьшение количества цифр после запятой до двух; осуществить обрисовку ячеек таблицы.

Форматированная таблица с исходными данными и результатами может иметь вид, представленный на рис. 2.

Расход топлива на 100 кг
21
23
24
20
21
23
24
20
19
25
26
27
21
22
25
26
27
28
21
21
22

Расход топлива на 100 кг	
Среднее	23,14
Стандартная ошибка	0,58
Медиана	23
Мода	21
Стандартное отклонение	2,65
Дисперсия выборки	7,03
Эксцесс	-1,11
Асимметричность	0,31
Интервал	9
Минимум	19
Максимум	28
Сумма	486
Счет	21
Наибольший(5)	26
Наименьший(5)	21
Уровень надежности(95,0%)	1,21

Рис. 2. Окно с выходными данными

Интерпретация выходных результатов. Выходные данные содержат три показателя общей направленности: среднее, медиана, мода.

Среднее определяет средний расход топлива на 100 кг – 23,14. Вычисляется как результат деления Суммы (486) на Счет (21).

Медиана – среднее из множества чисел. То есть половина всех значений в исследуемом ряду будет меньше медианы, а другая половина – больше ее. Медиана равна 23. Таким образом, примерно половина машин имеет показатель расхода топлива на 100 кг, больший 23, и примерно половина меньший показатель. Если все значения были отсортированы и ранги от 1 до 21 соответствуют отсортированным значениям, то медиана – это одиннадцатое значение, 23 л/кг. Имеется 10 значений ниже этого показателя и 10 значений выше. Для нечетного числа данных n медианой является значение с рангом (n+1)/2, а для четного – медиана находится ровно между двумя значениями с рангами n/2 и n/2+1. Ранг – это порядковый номер значения в отсортированном списке.

Мода – наиболее часто встречающееся значение, равна 21 л/кг. Если каждое значение встречается только один раз и модой является каждое значение, то Excel выводит «#Н/Д». В таком случае надо построить распределение частот и найти значение с наибольшей частотой; полученные значения называются модальным интервалом.

В таблице выходных данных имеется **несколько характеристик дисперсии.**

Интервал (9 л/кг) определяется как разность Максимума (28 л/кг) и Минимума (19 л/кг). Таким образом, интервал – это разность между максимальным и минимальным значениями в выборке.

Стандартное отклонение (2,65 л/кг) – это квадратный корень из дисперсии. Это характеристика среднего отклонения данных от среднего значения. Определяется следующим образом: вычисляется отклонение между каждым значением и средним, затем отклонения возводятся в квадрат и суммиру-

ются, сумма квадратов отклонений делится на счет минус один (то есть на n-1). В результате получается **дисперсия выборки** (7,03). Квадратный корень из дисперсии является стандартным отклонением.

То, что в выходной таблице называется стандартным отклонением и дисперсией, на самом деле является выборочным стандартным отклонением и выборочной дисперсией, вычисляемыми с (n-1) в знаменателе. Выборочное стандартное отклонение можно вычислить самостоятельно, используя статистическую функцию =СТАНДОТКЛОН(A2:A22); выборочную дисперсию можно вычислить, используя функцию =ДИСП(A2:A22). Для нахождения стандартного отклонения и дисперсии по генеральной совокупности, вычисленными с n в знаменателе, используют функции =СТАНДОТКЛОНП(A2:A22) и =ДИСПР(A2:A22).

Значения **Наибольший(1)** и **Наименьший(1)** являются соответственно первым наибольшим и первым наименьшим значениями расхода топлива. Если в качестве k задать, например, 5 как на рис. 1, то можно будет определить пятое наибольшее и пятое наименьшее значения. Данные значения соответствуют примерно75-му персентилю (третий квартиль) и 25-му персентилю (первый квартиль) во множестве из 21 значения.

Стандартная ошибка (0,58) равняется выборочному стандартному отклонению, деленному на корень квадратный из размера выборки. Стандартная ошибка является характеристикой достоверности среднего и используется в статистических выводах (доверительные интервалы и проверка гипотез).

Уровень надежности (95%) (1,21) равняется половине длины 95% доверительного интервала для среднего. Левая граница девяностопятипроцентного доверительного интервала равна среднему минус половина ширины, а правая – среднему плюс половина ширины, т.е. 23,14- 1, 21 = 21,94 и 23,14+1.21 = 24,35 соответственно. Таким образом, доверительный интервал, куда попадает 95% значений расхода топлива, составляет от 21,94 до 24,35.

Эксцесс является показателем островершинности симметричных распределений. Если распределение более плоское, чем нормальное, то эксцесс будет положительным. Если же распределение имеет более выраженный острый пик, чем нормальное, то эксцесс отрицательный. В примере эксцесс равен -1.11, то есть распределение имеет более выраженный острый пик, чем нормальное распределение.

Асимметричность показывает степень симметрии распределения. Говорят, что распределение имеет положительное отклонение, или скошено вправо, если большинство значений расположены в положительном направлении. Если же большинство значений расположено в отрицательном направлении, то распределение имеет отрицательное отклонение, или скошено влево. В примере асимметричность положительна (+0.31), значит, значения скошены вправо от среднего значения.

Существует другой показатель асимметричности (коэффициент асимметричности Пирсона), являющийся альтернативой характеристики асимметричности Excel. Показатель определяется как 3*(среднее-медиана)/стандартное отклонение. В рассматриваемом примере коэффициент Пирсона равен 0.16.

Среднее зависит от экстремальных значений данных. Экстремальные значения в положительной части оси увеличивают среднее, оно становится больше медианы, и коэффициент симметричности получается положительным. Экстремальные значения в отрицательном направлении уменьшают среднее, и среднее становится меньше медианы, в этом случае коэффициент отрицателен.

Значения коэффициента, который вычисляет Excel, можно интерпретировать следующим образом:

K Excel < -1 – скошено влево;

$-1 \leq$ K Excel ≤ 1 – приблизительно симметрично;

K Excel > 1 – скошено вправо.

Значения коэффициента асимметричности Пирсона можно интерпретировать следующим образом:

K < -0.5 – скошено влево;

$-0.5 \leq K \leq 0.5$ – приблизительно симметрично;

$K > 0.5$ – скошено вправо.

Анализ данных говорит, что данные приблизительно симметричны с небольшим положительным отклонением вправо.

Для больших наборов данных оба коэффициента обычно дают одинаковые результаты.

1.2. Инструмент анализа: гистограмма

Инструмент анализа Гистограмма строит таблицу распределения частот данных и на ее основе создает диаграмму. В результаты кроме отдельных частот могут быть включены интегральные (накопительные) проценты.

Перед использованием инструмента следует определить **отрезки разбиения**. В противном случае Excel использует равные интервалы. **Их количество приближенно равно квадратному корню из числа значений данных. Обычно достаточно от 5 до 15 интервалов.** При этом интервалы начинаются от минимального и заканчиваются максимальным значением. В случае явного задания интервалов лучше использовать числа, кратные двойке, пятерке или десятке.

Чтобы определить интервалы, сначала определяют минимальное и максимальное значения данных с помощью инструмента «Описательная статистика» или с помощью статистических функций МАКС и МИН.

Для данных о расходе топлива минимум равняется 19, а максимум 28. Гистограмма может начинаться первым интервалом со значением 18, длиной 2, и заканчиваться седьмым интервалом на значении 30. Использованный подход создает пустые интервалы на каждом из концов; слева это интервал «18 или меньше» и справа «более, чем 30».

Максимальное значение для каждого интервала в Excel называется **карманом.** Так, первый карман равняется 18, а интервал содержит значения, равные или меньшие 18. Следующий интервал будет больше 18 до 20 включительно. Инстру-

мент Гистограмма автоматически добавляет интервал, называемый **Еще**. В примере последний карман равен 30, а последний интервал (Еще) будет содержать значения большие, чем 30.

Порядок создания Гистограммы следующий:

– ввести название данных в ячейку, например в А1; ввести данные в столбец А, начиная с ячейки А2;

– ввести слово «Карман», например, в ячейку С1; ввести возможные интервалы в столбец С, начиная с ячейки С2;

– выбрать команду Анализ данных (для Excel 2003 – Сервис, Анализ данных; для Excel 2007 – панель Данные, кнопка Анализ данных);

– выбрать в диалоговом окне Гистограмма в списке доступных инструментов;

– в поле **Входной интервал** ввести ссылки на ячейки с данными, включая метку (А1:А18);

– в поле **Интервал карманов** ввести ссылки на ячейки со значениями, разделяющими интервалы, включая метку (С1:С8); данные точки разбиения, или карманы, должны быть расположены в возрастающем порядке;

– активировать (выставить) флажок Метки, чтобы указать, что метки (названия данных и интервала карманов) включены во входные данные;

– активировать переключатель **Выходной интервал** и в поле ввести ссылку для левого верхнего угла области, где будет располагаться таблица с выходными данными (Е1);

– флажок **Парето** (отсортированная диаграмма); не активированный флажок обеспечивает вывод обычной диаграммы; активированный флажок означает, что интервалы сортируются в соответствии с частотами перед отображением диаграммы;

– флажок **Интегральный процент**; обеспечивает вывод или нет на диаграмме дополнительного графика интегрального процента;

– флажок **Вывод графика** в активном состоянии обеспечивает вывод диаграммы на листе. Дополнительно выводится

таблица распределения частот (в приведенном примере, значения расхода топлива на 100 кг и количества появлений этих значений в выборке).

Диалоговое окно «Гистограмма» представлено на рис. 3.

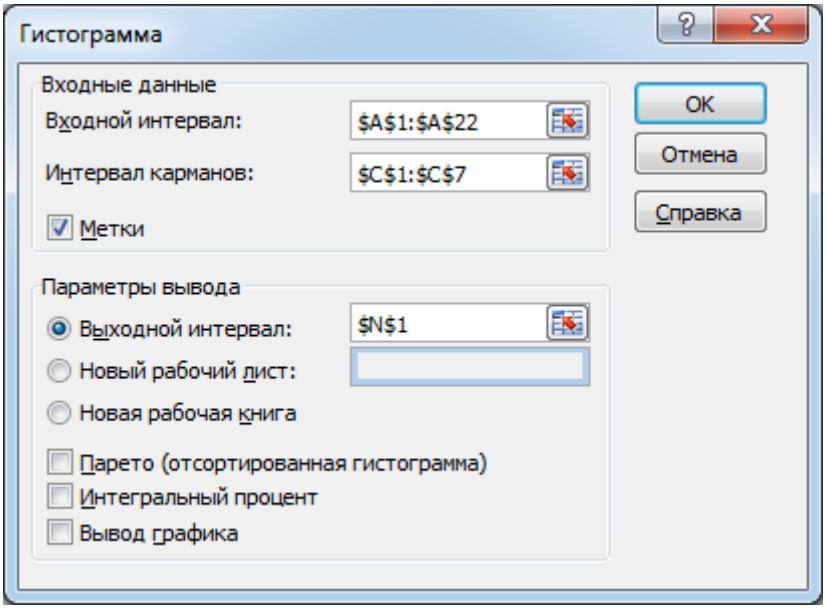

Рис. 3. Диалоговое окно «Гистограмма»

Результат построения гистограммы представлен на рис. 4. **Усовершенствование гистограммы** можно осуществить следующим образом:

– **легенда**, может быть удалена (щелчок по легенде (Частота) и нажатие клавиши Delete);

– **подписи по оси X**, можно изменить, щелкнув по тексту;

– **название диаграммы**, можно изменить, щелкнув по тексту;

– **ширина столбцов**, в обычных диаграммах столбцы не разделены, а находятся рядом. Дважды щелкнуть по одному из

столбцов, в окне Формат ряда данных на вкладке Параметры изменить ширину зазора с 150% на 0%, нажать ОК;

– **цвет столбцов**, щелкнуть правой кнопкой в центре одного из столбцов, выбрать Формат рядов данных, в диалоговом окне выбрать обычную границу и прозрачную заливку.

Измененная диаграмма представлена на рис. 5.

Карман	*Частота*
20	3
22	7
24	4
26	4
28	3
30	0
Еще	0

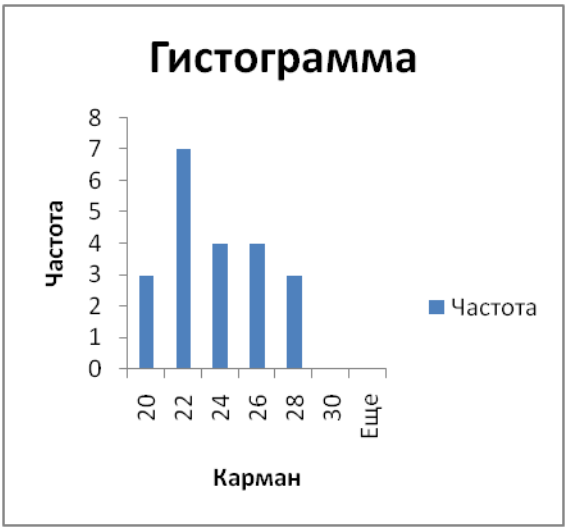

Рис. 4. Окно с построенной гистограммой

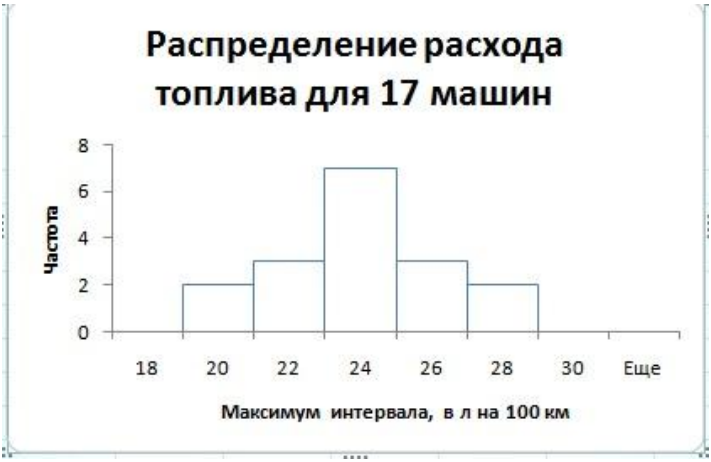

Рис. 5. Гистограмма после форматирования

1.3. Лабораторная работа № 1

Провести анализ данных, используя инструмент «Описательная статистика».

Построить гистограмму.

Сделать выводы по результатам анализа.

Проанализировать результаты по следующим позициям:

– среднее значение оцениваемого показателя;

– срединное значение, количество значений меньше и больше данного;

– наиболее часто встречаемое значение показателя, сколько раз встречается;

– максимальное и минимальное значение показателя;

– стандартное отклонение (среднее отклонение данных от среднего значения);

– определение интервала, в который попадают 95 % значений показателя;

– островершинность распределения показателя;

– степень симметрии распределения.

Задание № 1

Оценить работу сети торговых точек, продающих экспериментальное сырье. Количество торговых точек равно 15. Оценка производится по размеру месячной прибыли. Прибыль нужно сформировать с помощью генератора случайных чисел. Для этого с помощью пакета «Анализ данных» выбираем «Генерацию случайных чисел».

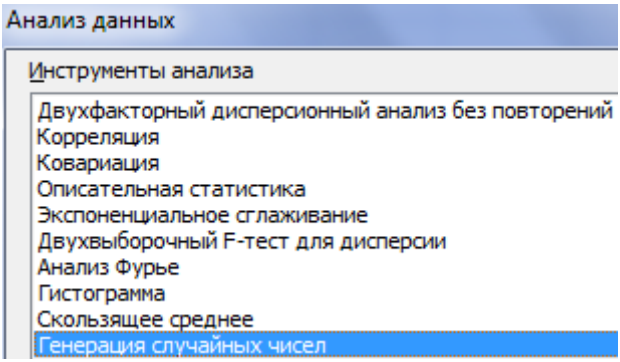

Рис. 6.– Диалоговое окно «Анализ данных»

Заполняем параметры для генерации в соответствии с заданием. Распределение – «нормальное». Жмем ОК. Получаем набор случайных чисел.

При помощи параметра **Случайное рассеивание** можно фиксировать последовательность выводимых случайных чисел. При повторных запусках генератора можно использовать это значение для получения тех же самых случайных чисел.

Число переменных – это число столбцов таблицы, куда будут размещены случайные числа (данные для анализа). Число случайных чисел – число случайных данных. Если Вы укажите число переменных – 1, а число случайных чисел – 15, то получите столбец из 15 чисел.

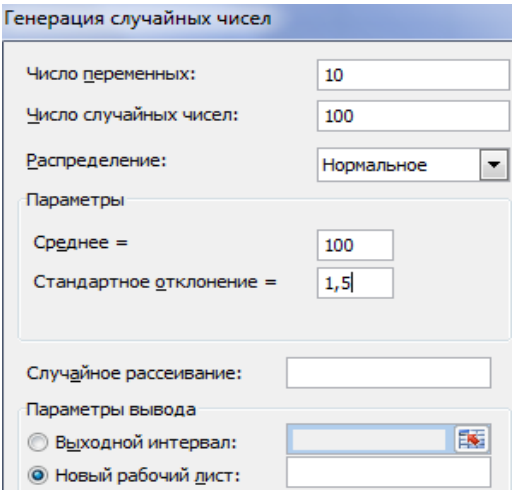

Рис. 7 – Диалоговое окно «Генерация случайных чисел»

Среднее и стандартное отклонение указать в соответствии с вариантом задания. После того, как данные будут сформированы, округлите их, выбрав через контекстное меню **Формат ячеек – числовой**.

Вариант	Среднее	Стандартное отклонение	Вариант	Среднее	Стандартное отклонение
1	100000	10000	15	250000	10000
2	110000	11000	16	260000	11000
3	120000	12000	17	270000	12000
4	130000	13000	18	280000	13000
5	140000	14000	19	290000	14000
6	150000	15000	20	300000	15000
7	160000	16000	21	310000	16000
8	170000	17000	22	320000	17000
9	180000	18000	23	330000	18000
10	190000	19000	24	340000	19000
11	200000	20000	25	350000	20000
12	210000	21000	26	360000	21000
13	230000	22000	27	370000	22000
14	240000	23000	28	380000	23000

Задание № 2

Оценить продажу экспериментального сырья в сети торговых точек. Количество точек продаж - 21. Оценка производится по объему реализации продукции в литрах. Объем продаж сформировать с помощью генератора случайных чисел по аналогии с заданием 1. Среднее соответствует номеру Вашего варианта, увеличенному в 10 раз, стандартное отклонение у всех одинаковое – 10.

Контрольные вопросы

1. Дать определение случайной величины.
2. Дать определение вероятности.
3. Дискретная случайная величина.
4. Непрерывная случайная величина.
5. Закон распределения случайной величины.
6. Математическое ожидание.
7. Медиана.
8. Мода.
9. Дисперсия.
10. Стандартное отклонение (стандарт).
11. Размах.
12. Доверительный интервал.
13. Асимметрия.
14. Эксцесс.

2. АНАЛИЗ КАТЕГОРИЙНЫХ ДАННЫХ В EXCEL

Характеристики данных могут описываться не только числами, но и с помощью категорий. Например, пол (мужской или женский), преподавательская должность (ассистент, доцент, профессор), автомобильные марки (Ниссан, Форд, Тойота, Шевроле и т.д.). Категорийные данные могут быть описаны с помощью частот, получаемых подсчетом числа значений данных, попавших в каждую из категорий.

Для выполнения простого подсчета частот в случае одной переменной и получения перекрестной таблицы частот для двух переменных в Excel используются Сводные таблицы. В статистическом анализе столбцы называются случайными величинами, а строки – случаями или наблюдениями. В ячейке содержится значение случайной величины для выпавшего элементарного события.

2.1. Подсчет частот в случае одной переменной

Рассмотрим подсчет частот в случае одной переменной. Пусть имеются данные об оценке программного продукта пользователями. Пользователи имеют различную организационно-правовую форму. Предположим, оценки программного продукта находятся в диапазоне от «Отлично» до «Очень плохо», а организационно-правовая форма может принимать значения – 1 (частное лицо), 2 (вуз), 3 (корпорация). Закодированные результаты наблюдений представлены на рис. 8.

Подсчет частот реализуется следующим образом.

1. Поместить на лист книги данные, показанные на рис. 8.

2. Выделить любую ячейку в исходной таблице. Выбрать в меню Данные, Сводная таблица (Excel 2003) или Вставка, Сводная таблица (Excel 2007).

3. Построить сводную таблицу на существующем листе ниже исходной таблицы: Вставка – Сводная таблица. Для сводной таблицы выбрать: в область Строка поместить поле Оценка, в область Данные поместить поле Номер. Автоматически для номера будет выбрана функция сумма, сумму необходимо заменить на количество, используя контекстное меню и пункт Параметры поля. Затем выйти из режима построения сводной таблицы.

Номер	Оценка	Пользо-ватель
1	5	3
2	4	1
3	4	1
4	2	2
5	1	3
6	3	1
7	3	3
8	4	2
9	5	3
10	4	2
11	3	1
12	1	1
13	1	2
14	5	3
15	1	1
16	2	1
17	3	3
18	4	3

Оценка	
Отлично	5
Хорошо	4
Удовлетворительно	3
Плохо	2
Очень плохо	1

Форма	
Частное лицо	1
Вуз	2
Корпорация	3

Рис. 8. Оценка программного продукта

Результирующая сводная таблица представлена на рис. 9.

Количество по полю Номер	
Оценка	Итог
1	4
2	2
3	4
4	5
5	3
Общий итог	18

Рис. 9. Сводная таблица, созданная мастером сводных таблиц

Результаты будут более наглядными, если вместо количества отобразить проценты (рис. 10). Для замены абсолютных чисел на проценты выделяют ячейку с данными в столбце Итог, щелкают правой кнопкой мыши, выбирают Параметры поля, щелкают по кнопке Дополнительно, в списке Дополнительные вычисления выбирают Доля от общей суммы, ОК. Уменьшают разрядность полей в графе Итог до целого числа.

Количество по полю Номер	
Оценка	Итог
1	22%
2	11%
3	22%
4	28%
5	17%
Общий итог	100%

Рис. 10. Сводная таблица с итогами в процентах

Для наглядности можно числовые оценки заменить на категории и добавить заголовок к результатам (рис 11).

Процент появления оценки	
Оценка	**Итог**
Отлично	22%
Хорошо	11%
Удовлетворительно	22%
Плохо	28%
Очень плохо	17%
Общий итог	100%

Рис. 11. Итоговая сводная таблица оценки программного продукта

Для этого нужно сводную таблицу скопировать в буфер, перейти в новое место, выбрать Правка, Специальная вставка и активизировать переключатель Значения и форматы чисел. Произвести добавление заголовка, изменение оценок.

Таким образом, **анализ одной категорийной переменной** может осуществляться следующим образом:

– закодировать категорийные данные (каждой категории поставить в соответствие число); пронумеровать данные;

– осуществить подсчет частот для выбранной категорийной переменной, используя мастера сводных таблиц; построить сводную таблицу, используя два поля – категорийную переменную и номер;

– для удобства анализа преобразовать сводную таблицу, перейдя от количественных данных к процентным данным;

– закодированные данные снова заменить на категории; для этого предварительно провести копирование полученных результатов и специальную вставку (значений и форматов чисел); добавить поясняющие подписи.

2.2. Установка зависимостей между двумя категорийными переменными

Для установки зависимостей между двумя категорийными переменными строят перекрестные таблицы. Они показывают для каждой комбинации категорий, сколько раз она встречается. Например, перекрестная таблица может быть полезна при определении зависимости перспектив использования программного продукта от организационно-правовой формы пользователя.

Построение перекрестной таблицы для двух категорийных переменных осуществляется с помощью построения сводной таблицы следующим образом.

1. Поместить на новый лист книги данные, показанные на рис. 8.

2. Выделить любую ячейку в исходной таблице. Выбрать в меню Данные, Сводная таблица (Excel 2003) или Вставка, Сводная таблица (Excel 2007).

3. Построить сводную таблицу на существующем листе ниже исходной таблицы. Для сводной таблицы выбрать: в область Строка поместить поле Пользователь, в область Столбец поместить поле Оценка, в область Данные поместить поле Номер. Автоматически для номера будет выбрана функция сумма, сумму надо заменить на количество через контекстное меню. Затем выйти из режима построения сводной таблицы.

Созданная сводная таблица представлена на рис. 12.

Количество по полю Номер	Оценка					
Пользователь	1	2	3	4	5	Общий итог
1	2	1	2	2		7
2	1	1		2		4
3	1		2	1	3	7
Общий итог	4	2	4	5	3	18

Рис. 12. Сводная таблица с двумя переменными

Для облегчения сравнений перспектив использования программного продукта в зависимости от организационно-правовой формы пользователя количественные показатели следует выразить в процентах. Для этого выделяют любую ячейку с данными на пересечении строк и столбцов, вызывают контекстное меню, выбирают пункт Параметры поля, щелкают по кнопке Дополнительно, в списке Дополнительные вычисления выбирают Доля от суммы по строке. Результаты данных действий отображены на рис. 13.

Количество по полю Номер	Оценка					
Пользователь	1	2	3	4	5	Общий итог
1	28,57%	14,29%	28,57%	28,57%	0,00%	100%
2	25,00%	25,00%	0,00%	50,00%	0,00%	100%
3	14,29%	0,00%	28,57%	14,29%	42,86%	100%
Общий итог	22,22%	11,11%	22,22%	27,78%	16,67%	100%

Рис. 13. Сводная таблица с данными в процентах

Таблица будет выглядеть лучше, если убрать Общий итог по столбцам. Для этого необходимо щелкнуть по произвольной ячейке в таблице и в контекстном меню выбрать Параметры таблицы (Параметры сводной таблицы). В окне Параметры сводной таблицы убрать отметку с пункта Общая сумма по столбцам (вкладка Итоги и фильтры, пункт Показывать общие итоги по столбцам), ОК. Вместо «Количество по полю Номер» для большей наглядности надо написать «Доля по строкам». В результате таблица процентов будет выглядеть, как показано на рис. 14.

Доля по строкам	Оценка					
Пользователь	1	2	3	4	5	Общий итог
1	28,57%	14,29%	28,57%	28,57%	0,00%	100%
2	25,00%	25,00%	0,00%	50,00%	0,00%	100%
3	14,29%	0,00%	28,57%	14,29%	42,86%	100%

Рис. 14. Сводная таблица долей по строкам

Пять категорий оценки перспектив использования программного продукта могут быть избыточными для восприятия. Целесообразно оставить только две категории, объединив оценки 1, 2 и 3 в первую группу, а данные с оценками 4 и 5 – во вторую группу. Для группировки необходимо:

– выделить ячейки, содержащие цифры 1, 2, 3; в контекстном меню выбрать Группа и структура, затем Группировать (Группировать);

– выделить ячейку с именем Группа1, выбрать в контекстном меню Группа и структура, Скрыть детали (Развернуть/Свернуть, Свернуть).

Свернутая таблица показана на рис. 15.

Доля по строкам	Оценка2	Оценка	
	Группа1	Группа2	Общий итог
Пользователь			
1	71,43%	28,57%	100,00%
2	50,00%	50,00%	100,00%
3	42,86%	57,14%	100,00%

Рис. 15. Свернутая таблица

Отформатированная версия таблицы показана на рис. 16.

Оценка программного продукта пользователями

Доля по строкам	**Оценка**		Общий итог
Организационно-правовая форма	Неблагоприятная	Благоприятная	
Частное лицо	71%	29%	100%
Вуз	50%	50%	100%
Корпорация	43%	57%	100%

Рис. 16. Отформатированная таблица с двумя параметрами

Форматирование таблицы проведем следующим образом:

– исходная сводная таблица остается на прежнем месте, скопируем ее в буфер;

– перейдем в новое место, выберем в контекстном меню Специальная вставка, активизируем переключатель Значения и форматы чисел;

– произведет форматирование скопированной таблицы.

Таким образом, **установка зависимостей между категорийными переменными** может осуществляться следующим образом:

– закодировать категорийные данные (каждой категории поставить в соответствие число); пронумеровать строки данных;

– построить сводную таблицу для двух категорийных данных; использовать также поле Номер;

– для наглядности преобразовать сводную таблицу, перейдя от количественных данных к процентным данным;

– произвести группировку категорий для большей наглядности и анализа данных;

– осуществить копирование полученных результатов и специальную вставку (значений и форматов чисел), закодированные данные снова заменить на категории; добавить поясняющие подписи.

2.3. Лабораторная работа № 2

Задание № 1

Имеются данные с результатами опроса 20 студентов по пяти вопросам. Вопросы были следующие:

1) Ваш пол?; возможных ответов два: 0 – мужской, 1 – женский;

2) Важность (значимость) изучения курса анализа данных;

3) Я обеспокоен тем, усвою ли я этот курс;

4) Думаю, что статистика – очень скучный предмет;

5) Курсы лекций для большого числа студентов не позволяют использовать индивидуальный подход к каждому.

Ответы на вопросы с 2 по 5 соответствуют следующим числам:

– 1 – совершенно согласен;

– 2 – согласен;

– 3 – не знаю;

– 4 – не согласен;

– 5 – совершенно не согласен.

Результаты опроса представлены в табл. 1.

Таблица 1

Результаты опроса студентов

	1	2	3	4	5
Номер Студента	Пол	Применение	Обеспокоенность	Скучность	Индивидуальный подход
1	0	4	1	2	2
2	1	3	4	2	2
3	0	4	1	2	2
4	1	2	2	1	1
5	1	5	4	3	2
6	0	5	4	3	3
7	0	4	1	4	1
8	0	3	4	1	1
9	0	2	2	2	2
10	0	4	5	4	4
11	0	4	2	2	1
12	0	4	4	2	3
13	1	2	1	3	1
14	0	3	2	1	3
15	0	5	1	1	1
16	0	4	5	1	3
17	1	3	2	1	1
18	0	4	5	4	3
19	1	1	3	1	1
20	0	3	1	1	1

Числа в одном из столбцов таблицы (столбцы 2, 3, 4, 5 по заданию от преподавателя) сформируйте с помощью генератора случайных чисел. Среднее – 3, стандартное отклонение – 2. Отформатируйте числа так, чтобы они были целыми (дробная часть отсутствовала).

Провести следующий анализ категорийных данных.

1. Создать таблицу размерами 20x3 (номер студента, пол, применение). Создать сводную таблицу (в строках поместить применение, в столбцах поместить пол, на пересечение – количество номеров). Вычислить доли ответов от общей суммы на второй вопрос. Согласуются ли ответы мужчин и женщин? Итоговая таблица может иметь вид, представленный на рис.17. При выполнении задания сохранить все промежуточные таблицы.

Ответ на вопрос: Важность изучения курса анализа данных

Доля от общей суммы	**Пол**		
Применение	Мужской	Женский	Общий итог
Совершенно согласен	0%	5%	5%
Согласен	5%	10%	15%
Не знаю	15%	10%	25%
Не согласен	40%	0%	40%
Совершенно не согласен	10%	5%	15%
Общий итог	70%	30%	100%

Рис.17. Итоговая таблица по анализу важности изучения курса

2. Создать 2 таблицы 20x3 на одном листе. В одну таблицу поместить номер студента, применение, во вторую таблицу – номер студента, обеспокоенность. Распределите ответы на вопросы 2 и 3 по 2 категориям. Ответы 1, 2 и 3 объедините в

категорию «Может быть», а 4 и 5 – в категорию «Нет». Возможные виды итоговых таблиц приведены на рис. 18. При выполнении задания сохраните промежуточные таблицы.

Доля от общей суммы	
Применение	Итог
Может быть	45%
Нет	55%
Общий итог	100%

Доля от общей суммы	
Обеспокоенность	Итог
Может быть	60%
Нет	40%
Общий итог	100%

Рис.18. Возможные виды итоговых таблиц по анализу ответов о важности курса и обеспокоенности студентов

3. Создать таблицу размерами 20x3 (номер студента, пол, скучность). Вычислить доли ответов на четвертый вопрос. Согласуются ли ответы мужчин и женщин? Возможный вид итоговой таблицы приведен на рис. 19. При выполнении задания сохранить промежуточные таблицы.

Доля от общей суммы	Пол		
Скучность	**Мужской**	**Женский**	Общий итог
Совершенно согласен	25%	15%	40%
Согласен	25%	5%	30%
Не знаю	5%	10%	15%
Не согласен	15%	0%	15%

Рис.19. Итоговая таблица по анализу важности изучения курса для мужчин и женщин

4. Создать таблицу 20х3 (номер студента, применение, индивидуальный подход).

В сводной таблице в строках поместить применение, в столбцах - индивидуальный подход, на пересечении – количество по полю номер студента. Распределить ответы на вопросы 2 и 5 по двум категориям. Ответы 1, 2 и 3 объединить в новую категорию под названием «Может быть», а 4 и 5 объеденить в категорию «Нет». Вид итоговой таблицы представлен на рис. 20. При выполнении задания сохранить промежуточные таблицы.

Доля от суммы	**Индивидуальный подход**		
Применение	Может быть	Нет	Общий итог
Может быть	45%	0%	45%
Нет	50%	5%	55%
Общий итог	95%	5%	100%

Рис.20. Итоговая таблица по анализу ответов о возможности использования индивидуального подхода к студентам в процессе изучения курса

Задание № 2

Имеются данные о специализации и поле 25 студентов, посещающих курсы анализа данных. Данные представлены в табл. 2.

Необходимо провести следующий анализ категорийных данных.

1. Выполнить подсчет специализаций (частоты появления или количества) с помощью мастера сводных таблиц. Возможный вид итоговой таблицы приведен на рис. 21.

Таблица 2

Специализация и пол студента

Номер студента	Специализация	Пол
1	Информатика и ВТ	Мужской
2	Информационные системы	Мужской
3	Электронный бизнес	Женский
4	Информатика и ВТ	Женский
5	Информатика и ВТ	Мужской
6	Информационные системы	Мужской
7	Информационные системы	Женский
8	Информатика и ВТ	Мужской
9	Электронный бизнес	Мужской
10	Информатика и ВТ	Женский
11	Электронный бизнес	Мужской
12	Информационные системы	Женский
13	Электронный бизнес	Мужской
14	Электронный бизнес	Женский
15	Информатика и ВТ	Мужской
16	Информатика и ВТ	Мужской
17	Электронный бизнес	Женский
18	Электронный бизнес	Мужской
19	Информационные системы	Мужской
20	Информационные системы	Женский
21	Электронный бизнес	Женский
22	Электронный бизнес	Мужской
23	Электронный бизнес	Женский
24	Информационные системы	Женский
25	Информационные системы	Женский

Специализация	**Кол-во студентов**
Информатика и ВТ	7
Информационные системы	8
Электронный бизнес	10

Рис. 21. Возможный вид итоговой таблицы по подсчету специализаций

2. Создать сводную таблицу, в которой по строкам отразить специализацию, по столбцам – пол, на пересечении – количество по полю Номер студента. При переходе от количества к проценту выбрать долю от суммы по строке. Возможный вид итоговой таблицы приведен на рис. 22.

Количество по полю Номер студента	**Пол**		
Специализация	Женский	Мужской	Общий итог
Информатика и ВТ	29%	71%	100%
Информационные системы	63%	38%	100%
Электронный бизнес	50%	50%	100%
Общий итог	48%	52%	100%

Рис. 22. Возможный вид итоговой таблицы по подсчету процентной доли специализаций среди мужчин и женщин

Контрольные вопросы

1. Как провести анализ одной категорийной переменной?
2. Как проанализировать зависимости между категорийными переменными?

3. АНАЛИЗ ЗАВИСИМОСТЕЙ ЧИСЛОВЫХ ДАННЫХ

Для изучения зависимости между двумя числовыми переменными используются графики рассеяния. В Excel данный вид графиков называют точечной диаграммой, диаграммой рассеяния или XY-графиком. Такое графическое представление часто является первым шагом в процедуре приближения данных кривой с помощью регрессионной модели.

3.1. Построение диаграммы рассеяния

Пример. Имеются данные о стоимости недвижимого имущества: для проданных объектов недвижимости известны жилая площадь объекта и цена объекта.

Цена зависит от площади объекта. Таким образом, цена становится зависимой переменной (называют откликом или Y-переменной), а площадь – казуальной переменной. Аналогично, казуальную переменную называют независимой переменной или X-переменной. Первоначальная цель – визуально исследовать зависимость между размером жилой площади и ценой объекта. Затем вычислить корреляцию.

Порядок построения диаграммы следующий.

1. Первоначально необходимо задать исходные данные так, как это показано на рис. 23.

Жилая площадь	Цена в тыс. руб.
13	1500
13,5	1650
14	1650
15	1490
16,5	1590
19	1550
21	1500
27	2100
28	2300
30	2500
31	2500
33	2650
35	2700
37	2750

Рис.23. Исходные данные о жилой площади

2. Затем необходимо построить точечную диаграмму. При построении диаграммы названия столбцов не выделяют, выделяют только данные. Выбирают тип диаграммы – точечный и вид – первый. Диаграмму назовем – Объекты недвижимости. Легенду уберем. Подпись по оси X – Жилая площадь, в кв. м. Подпись по оси Y – Цена продажи, в тыс. руб. Примерный вид диаграммы представлен на рис. 24. Данные об объектах демонстрируют общую определенную зависимость – в среднем, чем больше жилая площадь, тем выше цена продажи.

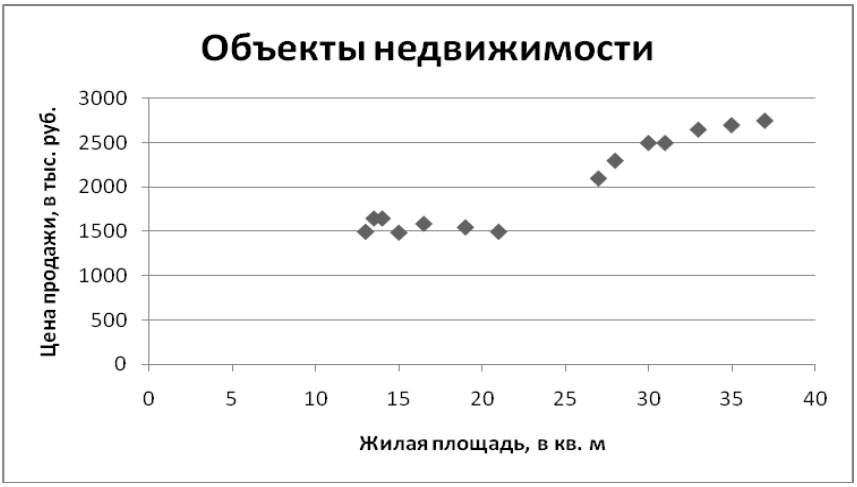

Рис. 24. Исходная точечная диаграмма

3. Диаграмму можно усовершенствовать.

Можно изменить ось X, чтобы она отображала значения от 10 до 40 кв. м. Для этого выделяют ось X, щелкают правой кнопкой, выбирают в контекстном меню Формат оси, щелкают по закладке Масштаб, в строке Минимум вводят 10, в строке Максимум вводят 40, в строке Основная единица – 5.

Можно изменить ось Y, чтобы она отображала значения от 1000 до 3000 тысяч рублей. Для этого выделяют ось Y, щелкают правой кнопкой, выбирают Формат оси, вкладку Шкала, вводят 1000, 3000, 200 в строки Минимальное значе-

ние, Максимальное значение, Цена основных делений соответственно.

Возможный вид диаграммы представлен на рис. 25.

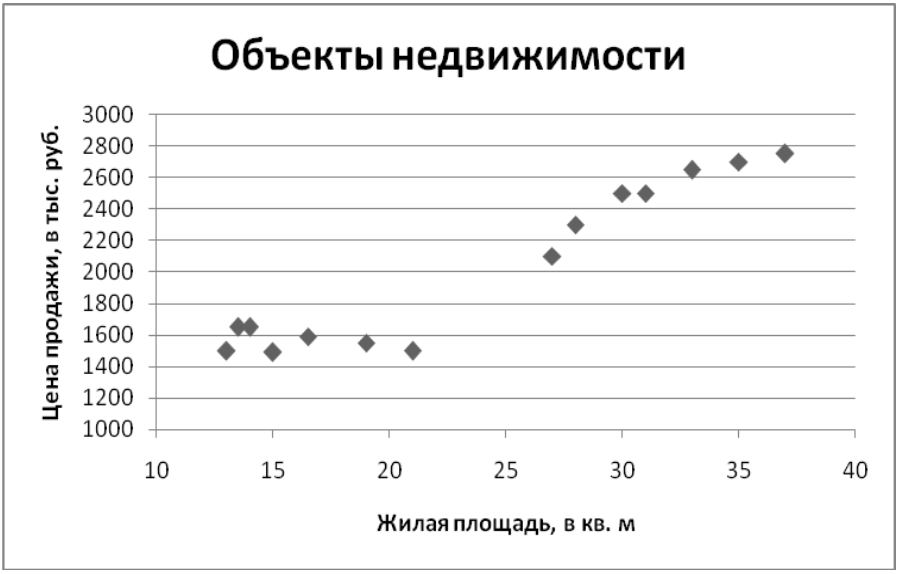

Рис. 25. Окончательная точечная диаграмма

3.2. Инструмент анализа: корреляция

Коэффициент корреляции – это общая характеристика двумерных данных, отражающая существующую между ними **линейную зависимость.**

Возможные значения для коэффициента корреляции лежат в диапазоне от -1 (минимальная отрицательная корреляция, все точки расположены на прямой с отрицательным углом наклона) до +1 (максимальная положительная корреляция, все точки лежат на прямой с положительным углом наклона). 0 соответствует отсутствию линейной зависимости.

Коэффициент корреляции характеризует только линейную зависимость; в случае строго нелинейной зависимости коэффициент корреляции может быть близким к нулю.

Следующие шаги описывают **вычисление коэффициента корреляции** с помощью средств анализа данных.

1. Ввести данные о жилой площади и цене (рис. 23).

2. В ячейку, например D1, внести фразу «Инструмент анализа: Корреляция».

3. Выбрать пункты меню Сервис, Анализ данных (Excel 2007: вкладка Данные, кнопка Анализ данных). В диалоговом окне в списке выбрать Корреляция и нажать ОК.

4. В диалоговом окне «**Корреляция**» указать:

– в строке Входной интервал (отметить мышкой) – интервал исходных данных вместе с названиями граф (метками);

– в переключателе **Метки в первой строке** выставить флажок;

– активировать переключатель **Выходной интервал** и выставить в строке ввода (мышкой), например D2;

– нажать ОК.

Возможный вид диалогового окна представлен на рис. 26.

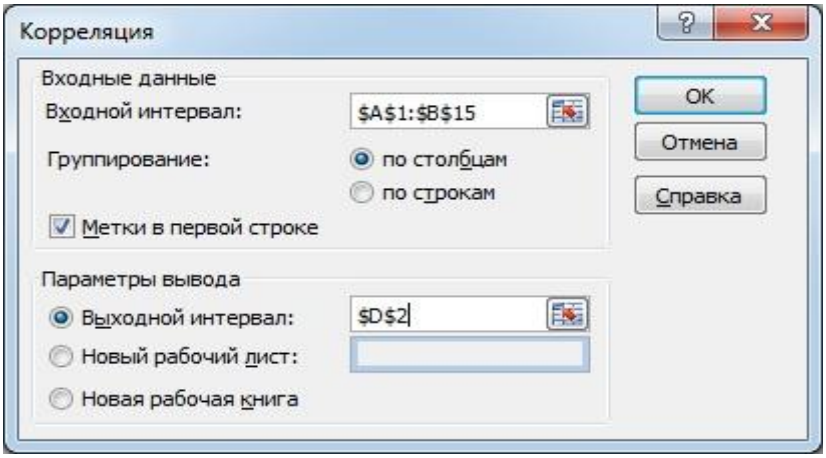

Рис. 26. Диалоговое окно «Корреляция»

Результатом выполнения корреляции является матрица попарных корреляций (рис. 27). На диагонали расположены единицы, показывающие, что каждая переменная положительно коррелированна с собой. Значение 0,954524599 – корреляция Цены и Площади. Полученный коэффициент корреляции говорит о том, что точки лежат почти на прямой с положительным углом наклона. Правая верхняя часть пустая, так как ее значение совпадает со значением из левой нижней части.

Инструмент анализа: Корреляция		
	Жилая площадь	*Цена в тыс. руб.*
Жилая площадь	1	
Цена в тыс. руб.	0,954524599	1

Рис. 27. Результат проведения корреляции

3.3. Корреляция нескольких переменных

Инструмент анализа Корреляция наиболее полезен при определении попарных корреляций трех и более переменных для последующего использования во множественной регрессионной модели.

Вычисление корреляции нескольких переменных реализуется следующим образом.

1. Ввести данные о площади, оценке и цене (рис. 28).

2. В ячейку, например E1, внести фразу «Инструмент анализа: Корреляция».

3. Выбрать пункты меню Сервис, Анализ данных (Excel 2007: вкладка Данные, кнопка Анализ данных). В диалоговом окне в списке выбрать Корреляцию и нажать ОК.

4. В диалоговом окне «Корреляция» указать:

– в строке **Входной интервал** (отметить мышкой) – интервал исходных данных вместе с названиями граф (метками);

– в переключателе **Метки в первой строке** выставить флажок;

– активировать переключатель **Выходной интервал** и выставить в строке ввода (мышкой), например E2; - нажать ОК.

Жилая площадь	**Оценка в тыс. руб.**	**Цена в тыс. руб.**
13	980	1500
13,5	1000	1650
14	1050	1650
15	950	1490
16,5	1100	1590
19	1050	1550
21	950	1500
27	1300	2100
28	1100	2300
30	1900	2500
31	1900	2500
33	2100	2650
35	2200	2700
37	2300	2750

Рис. 28. Исходные данные для вычисления корреляции

Результаты вычислений представлены на рис. 29.

Инструмент анализа Корреляция			
	Жилая площадь	*Оценка в тыс. руб.*	*Цена в тыс. руб.*
Жилая площадь	1		
Оценка в тыс. руб.	0,899	1	
Цена в тыс. руб.	0,955	0,934	1

Рис. 29. Результаты вычисления корреляции трех переменных

Выходные данные представляют собой матрицу трех попарных корреляций. Наибольшая корреляция 0,955 – между Площадью и Ценой. Корреляция между Оценкой и Ценой 0,934 – меньше и означает меньшую линейную зависимость между этими двумя переменными. Наименьшая корреляция 0,899 – между Площадью и Оценкой. Все три коэффициента корреляции близки к единице, что говорит о высокой степени зависимости между показателями, а также о том, что все точки лежат почти на прямой линии с положительным углом наклона. Зависимость между площадью, оценкой и ценой можно также оценить с помощью точечной диаграммы (рис. 30).

Рис. 30. Точечная диаграмма для площади, оценки и цены

3.4. Лабораторная работа № 3

Задание № 1

Выполнить анализ двух (площадь, цена) и трех (площадь, оценка, цена) переменных, описанный в теоретической части. Исходные данные взять из теоретической части.

1. Один из столбцов данных таблицы рис. 26 сформировать случайным образом. Среднее и стандартное отклонение – по заданию от преподавателя.

2. Построить точечную диаграмму для площади и цены.

3. Вычислить коэффициент корреляции и ковариации для площади и цены.

4. Вычислить корреляцию для площади, оценки и цены.

5. Построить точечную диаграмму для площади, оценки и цены.

Задание № 2

Имеются исходные данные о наличии нежилых помещений и месячной стоимости их аренды (табл. 3).

Таблица 3

Площадь нежилых помещений и стоимость их аренды

Площадь нежилого помещения, в кв. м	Месячная стоимость аренды в руб.
100	50000
20	10000
70	45000
80	47000
40	30000
110	45000
80	40000
60	30000
30	32500
50	27500

Выполнить следующие исследования.

1. Сформировать один из столбцов табл. 3 с помощью генератора случайных чисел. Среднее и стандартное отклонение – по заданию от преподавателя.

2. Создать точечную диаграмму. Ответить на вопрос: имеется ли положительная или отрицательная зависимость между площадью нежилого помещения и стоимостью аренды?

2. Вычислить коэффициент корреляции. Прокомментировать направление и величину линейной зависимости.

Задание № 3

Имеются данные о тестировании студентов: полученных баллах, времени на подготовку к тестированию. Есть также данные об оценках, полученных на экзамене по предмету, по которому было тестирование. Данные приведены в табл. 4.

Таблица 4

Затраты времени на подготовку и полученные баллы и оценки

Студент	Время на подготовку к тестированию в час.	Балл	Оценка на экзамене
1	5	54	3
2	10	56	4
3	4	63	3
4	8	64	3
5	12	62	4
6	9	61	4
7	10	63	3
8	12	73	4
9	15	78	5
10	12	72	5
11	12	74	5
12	20	78	4
13	16	83	5
14	14	86	5
15	22	83	5
16	18	81	4
17	30	88	5
18	21	87	5
19	28	89	4
20	24	93	5

Выполнить следующие исследования.

1. Сформировать один из столбцов табл. 3 с помощью генератора случайных чисел. Среднее и стандартное отклонение – по заданию от преподавателя.

2. Создать точечную диаграмму. Ответить на вопрос: имеется ли положительная или отрицательная зависимость между временем на подготовку и баллом и оценкой?

3. Вычислить коэффициенты корреляции между временем на подготовку и баллом и оценкой. Прокомментировать направление и величину линейных зависимостей.

Контрольные вопросы

1. Как построить диаграмму рассеяния для 2 переменных?

2. Как вычислить коэффициент корреляции для двух переменных? Что показывает коэффициент корреляции?

3. Как вычислить коэффициент ковариации?

4. Как определить попарные корреляции трех и более переменных?

4. ПРОСТАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ В EXCEL

Простая линейная регрессия используется для определения линейного уравнения, описывающего линейную зависимость между двумя переменными. В Excel существует 2 метода проведения линейной регрессии: команда Добавить линию тренда, инструмент анализа Регрессия. Перед тем как аппроксимировать данные прямой, нужно изучить график рассеяния. Если точки на графике лежат примерно на одной прямой, то можно применить предложенные выше методы. Если же точки не лежат на прямой, то следует использовать нелинейные методы аппроксимации (или приближенного представления).

4.1. Добавление линейного тренда

Пример. Имеются данные, необходимые для изучения стоимости недвижимого имущества: для проданных объектов недвижимости известны жилая площадь и цена объекта.

Цена зависит от площади объекта. Таким образом, цена становится зависимой переменной, а площадь казуальной переменной. Иногда зависимую переменную называют откликом или Y-переменной. Аналогично, казуальную переменную называют независимой переменной или X-переменной.

Первоначальная цель – визуально исследовать зависимость между размером жилой площади и ценой объекта на графике. Общий подход состоит в расположении данных таким образом, чтобы переменная X горизонтальной оси была в столбце слева, а переменная Y вертикальной оси находилась в столбце справа. Затем нужно построить точечную диаграмму. Порядок построения данной диаграммы рассматривался в лабораторной работе № 3. Для примера, рассмотренного в лабораторной работе № 3, был построен следующий график (рис. 31).

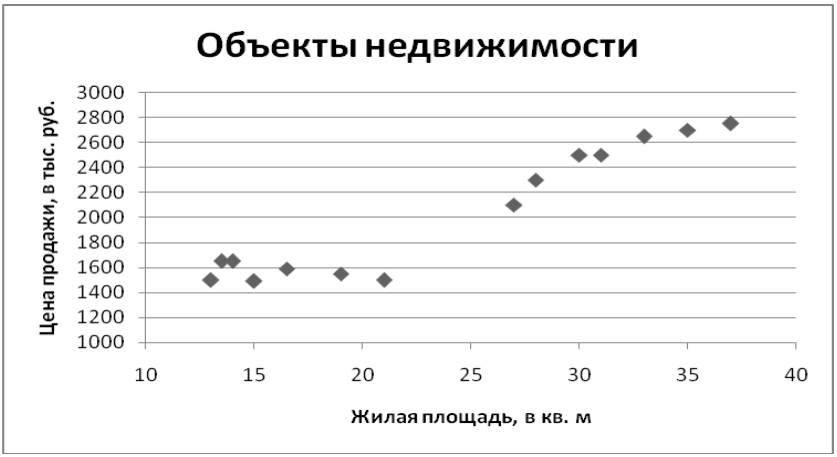

Рис. 31. График перед добавлением линии тренда

Добавление линии тренда на график позволяет получить аппроксимирующую прямую (заменяющую исходное множество точек прямой линией); линейное уравнение, описывающее зависимость между переменными X и Y; R^2 (характери-

стику приближения полученных результатов (прямой и уравнения) к истинной зависимости между переменными X и Y).

Следующие шаги описывают **добавление линейного тренда на график**:

– выделяют ряды данных, щелкнув по любой точке данных. Точки должны стать подсвеченными, в панели формул появляется строка, показывающая, что выделен ряд;

– щелкают правой кнопкой по ряду данных и выбирают в контекстном меню пункт Добавить линию тренда;

– для Excel 2007 в окне Параметры линии тренда активируют переключатель Линейная, переключатель Автоматическое, выставляют флажки Показывать уравнение на диаграмме, Поместить на диаграмму величину достоверности аппроксимации; для Excel 2003 в диалоговом окне Линия тренда выбирают вкладку Тип и щелкают по пиктограмме Линейная, в вкладке Параметры выбирают Автоматическое название аппроксимирующей (сглаживающей) прямой, включают опции Показывать уравнение на диаграмме и Поместить на диаграмму величину достоверности аппроксимации; ОК.Диаграмма с полученной линией тренда представлена на рис. 32.

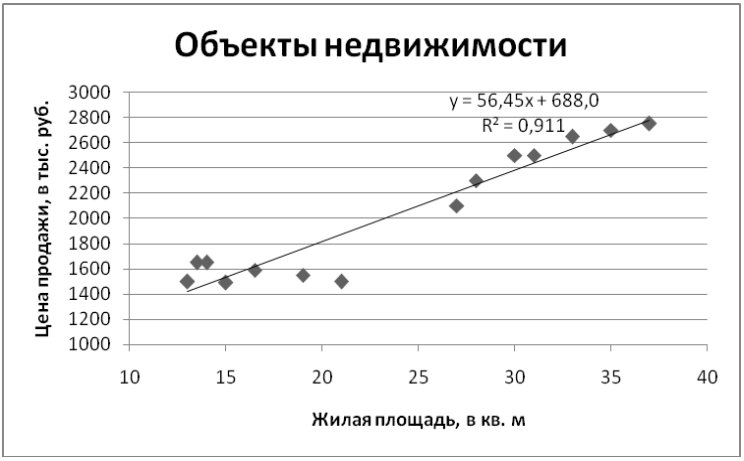

Рис. 32. Диаграмма с линией тренда

Если уравнение и значение R^2 выделены, то их можно переместить, также можно использовать стандартные возможности редактирования текста (изменить начертание, размер шрифта и т.д.).

Интерпретация линии тренда

Во-первых, линия тренда содержит уравнение аппроксимации (y=56,45x+688,0), которое определяет среднее соотношение (зависимость) между x и y. Уравнение можно переписать в следующем виде:

Предсказанная цена продажи = 688,0 + 56,45 * Жилая площадь.

Смещение по Y равно 688,0. Таким образом, недвижимость без жилой площади имеет цену 688 тысяч рублей.

Наклон или коэффициент регрессии 56,45 показывает среднее изменение переменной Y при единичном изменении переменной X. В данном примере единица измерения равна 56,45 тысячам рублей за квадратный метр. Таким образом, если два объекта отличаются на 50 квадратных метров жилой площади, то можно ожидать, что их стоимость будет отличаться на 56,45 * 50 = 2822,5 тысяч рублей.

Во-вторых, можно оценить степень приближения полученного уравнения к реальной зависимости между X и Y. Для этого исследуют значение R^2 (величину достоверности аппроксимации). Значение R^2 равно 0,911. Это достаточно высокая достоверность, которая означает, что примерно 91% колебаний цены продажи может быть выражено линейной функцией от жилой площади.

4.2. Инструмент анализа: регрессия

Для получения дополнительной информации о зависимости двух переменных необходимо использовать инструмент анализа – регрессия.

Следующие шаги описывают **реализацию регрессии** с помощью средств анализа данных.

1. Ввести данные так, чтобы переменная X располагалась слева, переменная Y - справа. Для результатов регрессионного анализа справа от данных необходимо иметь 16 пустых столбцов.

2. Выбрать пункты меню Сервис, Анализ данных (Excel 2007: вкладка Данные, кнопка Анализ данных). В диалоговом окне в списке выбрать Регрессия и нажать ОК.

3. В диалоговом окне «Регрессия» указать:

– в строке Входной интервал Y (отметить мышкой) – интервал значений зависимой переменной вместе с названием графы (меткой);

– в строке Входной интервал X (отметить мышкой) - интервал значений независимой переменной вместе с названием графы (меткой);

– в переключателе Метки выставить флажок;

– константа – ноль, данную опцию включают только в том случае, если хотят, чтобы прямая регрессии проходила через начало координат (0,0); не заполнять;

– уровень надежности; Excel автоматически выводит 95% доверительный интервал для коэффициентов регрессии;

– активировать переключатель Выходной интервал и выставить в строке ввода (мышкой) адрес левого верхнего угла области результатов; можно поместить результаты на новый рабочий лист или новую книгу;

– остатки, включают эту опцию для получения подобранных значений (предсказанных Y) и остатков;

– график остатков, отмечают этот пункт для получения диаграммы остатков для каждого значения переменной X;

– стандартизированные остатки; отмечают этот пункт для получения нормированных остатков (каждый из остатков делится на стандартное отклонение остатков);

– график подбора; отмечают этот пункт для получения точечной диаграммы входных значений Y и подобранных значений Y относительно переменной X;

– нажать ОК.

Возможный вид диалогового окна «Регрессия» представлен на рис. 33.

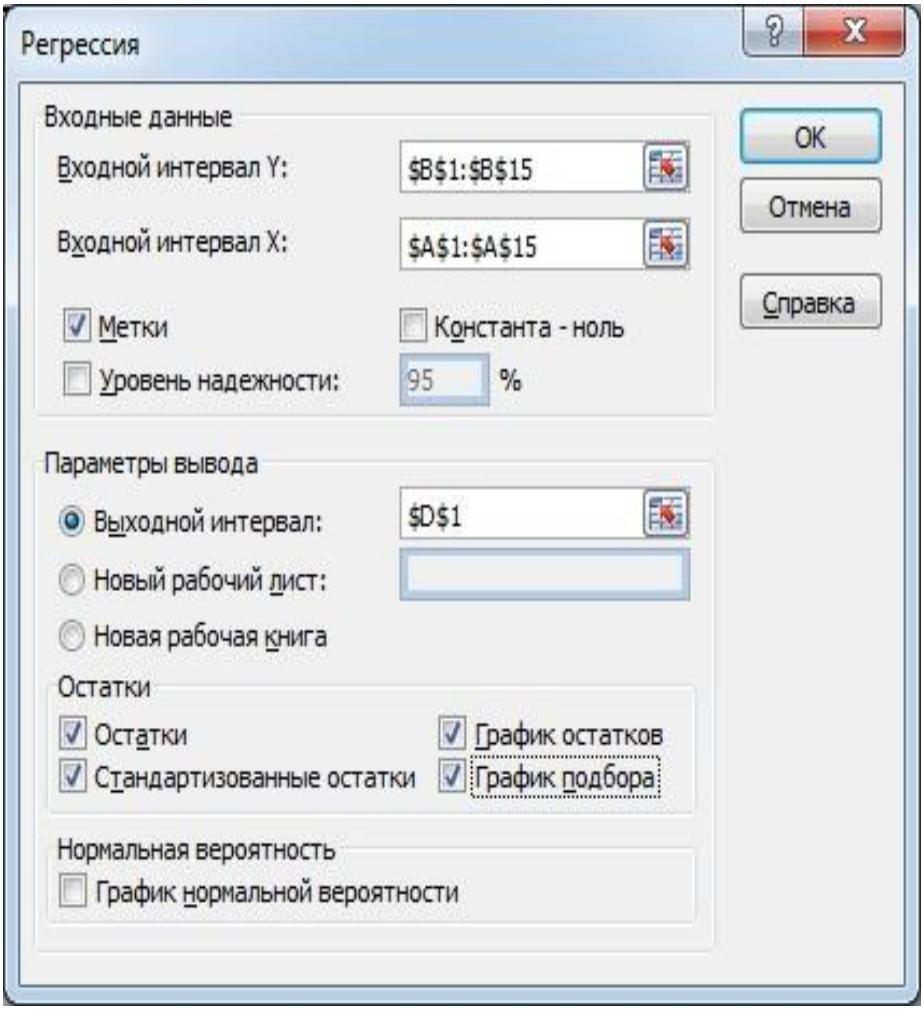

Рис. 33. Диалоговое окно «Регрессия»

Часть результатов проведения регрессии показана на рис. 34.

Для проведения анализа целесообразно поставить рядом исходные данные и полученные остатки без номеров наблюдений. Для этого надо скопировать эти данные и поставить вместе (рис. 35).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,954524599
R-квадрат	0,91111721
Нормированный R-квадрат	0,90371031
Стандартная ошибка	159,5655448
Наблюдения	14

Дисперсионный анализ

	Df	*SS*	*MS*
Регрессия	1	3131958,9	3131958,9
Остаток	12	305533,957	25461,163
Итого	13	3437492,857	

	Коэффициенты	*Стандартная ошибка*	*t-статистика*
Y-пересечение	688,014	128,354	5,360
Жилая площадь	56,449	5,089	11,091

Рис. 34. Итоговые результаты проведения регрессии

Жилая площадь	Цена в тыс. руб.	*Предсказанное Цена в тыс. руб.*	*Остатки*
13	1500	1421,862282	78,13771802
13,5	1650	1450,087209	199,9127907
14	1650	1478,312137	171,6878634
15	1490	1534,761991	-44,76199128
16,5	1590	1619,436773	-29,43677326
19	1550	1760,56141	-210,5614099
21	1500	1873,461119	-373,4611192
27	2100	2212,160247	-112,1602471
28	2300	2268,610102	31,38989826
30	2500	2381,509811	118,490189
31	2500	2437,959666	62,0403343
33	2650	2550,859375	99,140625
35	2700	2663,759084	36,2409157
37	2750	2776,658794	-26,6587936

Рис. 5. Цены продаж, предсказанные цены продаж и остатки

Интерпретация регрессии

Смещение и наклон аппроксимирующей прямой представлены в таблице «Вывод итогов» в графе «Коэффициенты» на рис. 34. Уравнение регрессии выглядит так:

Предсказанная цена = 688,014 + 56,45 * Жилая площадь

Предсказанные цены, приведенные в остатках на рис. 35 (иногда называемые подобранными значениями), являются результатами оценивания стоимости каждого объекта недвижимости с помощью уравнения регрессии.

Остатки равны разнице между фактическими и подобранными значениями цены продажи. Например, первый объект имеет жилую площадь 13 кв. м. Ожидается, что он стоит 1421862,3 руб., а реальная стоимость равна 1500000 руб. Оста-

ток для данного объекта равен 1500000-1421862,3, то есть + 78137,7. Реальная цена продажи на 78137,7 больше ожидаемой цены продажи. Остатки также называют отклонениями или ошибками.

Оценка приближения может быть осуществлена с помощью следующих характеристик: стандартная ошибка, R^2.

Стандартная ошибка 159,56 тыс. руб. и выражается в тех же единицах, что и зависимая переменная – цена продажи. Стандартная ошибка характеризует разброс цен относительно линии регрессии. Итак, стандартная ошибка (разброс цен, или стандартная ошибка оценки) равна 159565 руб.

Значение R^2 (0,91111721) характеризует долю изменений зависимой переменной, описываемых кривой регрессии. Данное число должно быть в пределах от нуля до единицы и может выражаться в процентах. В данном примере приблизительно 91% колебаний цены описывается моделью с жилой площадью в качестве независимой переменной линейного уравнения.

4.3. Лабораторная работа № 4

Задание № 1

Имеются данные о наличии нежилых помещений и месячной стоимости их аренды. Исходные данные представлены в табл. 5.

Таблица 5

Площадь нежилых помещений и месячная стоимость их аренды

Площадь нежилого помещения, в кв. м	Месячная стоимость аренды в руб.
100	50000
20	10000
70	45000
80	47000
40	30000

Продолжение табл. 5

110	45000
80	40000
60	30000
30	32500
50	27500

Выполнить следующие исследования.

1. Задать на отдельном листе исходные данные.

2. Создать точечную диаграмму и добавить линию тренда.

3. Провести анализ добавленной линии тренда:

– указать уравнение, задающее линейную зависимость между площадью нежилого помещения (X) и месячной стоимостью аренды (Y);

– оценить степень приближения полученного уравнения к реальной зависимости между X и Y;

– используя полученное уравнение, вычислить расчетную (предсказанную) месячную стоимость аренды.

4. Задать исходные данные на другом листе и реализовать инструмент анализа Регрессия.

5. Провести анализ результатов линейной регрессии:

– создать таблицу из четырех столбцов: площадь нежилого помещения, месячная стоимость аренды, предсказанная месячная стоимость аренды, остатки;

– записать уравнение регрессии;

– проанализировать оценку приближения, используя стандартную ошибку и значение R^2.

Задание № 2

Имеются данные о тестировании студентов: время на подготовку к тестированию и полученные баллы. Данные приведены в табл. 6.

Таблица 6

Затраты времени на подготовку и полученные баллы

Студент	Время на подготовку к тестированию в час.	Балл
1	5	54
2	10	56
3	4	63
4	8	64
5	12	62
6	9	61
7	10	63
8	12	73
9	15	78
10	12	72
11	12	74
12	20	78
13	16	83
14	14	86
15	22	83
16	18	81
17	30	88
18	21	87
19	28	89
20	24	93

Выполнить следующие исследования.

1. Задать на отдельном листе свои исходные данные с помощью генератора случайных чисел. Среднее и стандартное отклонение – по заданию от преподавателя.

2. Создать точечную диаграмму и добавить линию тренда.

3. Провести анализ добавленной линии тренда:

– указать уравнение, задающее линейную зависимость между временем на подготовку к тестированию (X) и полученным баллом (Y);

– оценить степень приближения полученного уравнения к реальной зависимости между X и Y;

– используя полученное уравнение, вычислить расчетный (предсказанный) балл.

4. Задать исходные данные на другом листе и реализовать инструмент анализа Регрессия.

5. Провести анализ результатов линейной регрессии:

– создать таблицу из четырех столбцов: время на подготовку к тестированию, полученный балл, предсказанный балл, остатки;

– записать уравнение регрессии;

– проанализировать оценку приближения, используя стандартную ошибку и значение R^2.

Контрольные вопросы

1. Как добавить линию тренда на график?
2. Как проанализировать полученное уравнение и коэффициент?
3. Как осуществить регрессию?
4. Как интерпретировать результаты регрессии?

5. КОНТРОЛЬНАЯ РАБОТА ДЛЯ СТУДЕНТОВ ЗАОЧНОЙ ФОРМЫ ОБУЧЕНИЯ

В рамках контрольной работы студенты заочной формы обучения выполняют лабораторные работы №1 и №2.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Кулаичев, А. П. Методы и средства комплексного анализа данных: учеб. пособие / А. П. Кулаичев. – М.: ФОРУМ, 2010.

2. Каплан, А. В. Статистическая обработка и анализ экономических данных / А. В. Каплан. - Ростов н/Д: Феникс, 2007.

3. Мидлтон, М. Р. Анализ статистических данных с использованием MS Excel для Office XP / М. Р. Мидлтон. – М.: Бином, 2013.

СОДЕРЖАНИЕ

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
к выполнению лабораторных работ
и контрольной работы
для студентов направления 09.03.01
«Информатика и вычислительная техника» (профиль
«Вычислительные машины, комплексы, системы и сети»)
очной и заочной форм обучения

Составители:
Сергеева Татьяна Ивановна,
Петрухнова Галина Викторовна

Компьютерный набор Сергеева Т.И., Петрухновой Г. В.

Подписано к изданию .09.2019

Уч.-изд. л. 3.3

ФГБОУ ВО «Воронежский государственный технический университет»
394026 Воронеж, Московский просп., 14