МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Логистика и организация производства»

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Методические рекомендации к лабораторным работам для студентов специальности 1-25 80 01 «Экономика» очной и заочной форм обучения

Могилев 2020

УДК 519.25
ББК 65.051
Т38

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Логистика и организация производства» «14» января 2020 г., протокол № 10

Составитель ст. преподаватель Т. М. Лобанова

Рецензент канд. экон. наук, доц. А. В. Александров

Методические рекомендации к лабораторным работам предназначены для студентов специальности 1-25- 80 01 «Экономика» очной и заочной форм обучения.

Учебно-методическое издание

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Ответственный за выпуск	М. Н. Гриневич
Редактор	А. А. Подошевко
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс. Печать трафаретная. Усл. печ. л. . Уч.-изд. л. . Тираж 16 экз. Заказ №

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, Могилев.

Содержание

Введение

Целью лабораторных работ по дисциплине «Технологии интеллектуального анализа данных» является получение магистрантами практических навыков применения интеллектуальных алгоритмов анализа и обработки данных.

Интеллектуальный анализ данных рассматривает весь спектр проблем, связанный с процессом извлечения знаний из баз данных. В его основе лежат математические методы, такие как оптимизация, распознавание образов, статистика, а также использующие визуальное представление информации.

В результате изучения дисциплины, магистр осваивает основные алгоритмы анализа данных (классификация, кластеризация, регрессия); учится применять стандартные методы и разработанные технологии к решению вероятностных и статистических задач, обрабатывать статистическую информацию и получать статистически обоснованные выводы; приобретает навыки работы с основными программными технологиями и методами интеллектуальной обработки данных, применения современных пакетов программ для интеллектуального анализа данных на ЭВМ.

Отчёт по лабораторным работам представляет собой файл с выполненным заданием. В процессе защиты работы студент поясняет отдельные этапы выполнения задания, при необходимости выполняет в присутствии преподавателя аналогичные задания.

1 Лабораторная работа № 1. Изучение возможностей применения встроенных функций Excel для статистической обработки информации

Цель работы: научиться использовать встроенные функции для статистической обработки и вычислений больших массивов информации.

Задачи: использовать встроенные функции MS Excel для автоматизации статистической обработки данных.

1.1 Математические функции

В MS Excel имеются следующие функции округления.

ОКРУГЛ(Число;Число_разрядов) – округляет число до указанного количества десятичных разрядов (по правилам математики).

ОКРУГЛВНИЗ(Число;Число_разрядов) – округляет число до ближайшего меньшего по модулю до указанного количества десятичных разрядов.

ОКРУГЛВВЕРХ(Число;Число_разрядов) – округляет число до ближайшего большего по модулю до указанного количества десятичных разрядов.

ОКРВНИЗ(Число;Точность) – округляет число до ближайшего меньшего по модулю целого, кратному указанному значению.

ОКРВВЕРХ(Число;Точность) – округляет число до ближайшего большего по модулю целого, кратному указанному значению.

ЦЕЛОЕ(Число) – округляет число до ближайшего меньшего целого.

Задание 1

Имеется набор исходных значений:

7474,9727 5097,257 1501,66667 5750 4957,5 454,7177 7714,75 44,6667 7497,777

Получить результаты округления исходных значений различными функциями округления:

- по правилам математики до одного знака в дробной части;
- в меньшую сторону до одного знака в дробной части;
- в большую сторону до одного знака в дробной части;
- значение, которое делится на 10 без остатка в меньшую сторону;
- значение, которое делится на 10 без остатка в большую сторону;
- определить только целую часть числа.

Функции суммирования. Функции суммирования позволяют выполнять сложение всех числовых аргументов или только значений, которые отвечают заданным критериям.

СУММ(Число1;Число2) – суммирует только числовые аргументы.

СУММЕСЛИ(Диапазон; Критерий; Диапазон_суммирования) – суммирует ячейки, заданные указанным условием.

СУММЕСЛИМН (Диапазон_суммирования; Диапазон_условия1; Условие1) – суммирует ячейки, удовлетворяющие заданному набору условий.

В качестве условий можно использовать следующие символы:

больше >

меньше <

не более <=

не менее >=

не равно <>

Для текстовых значений: ? – замена одного символа, * – замена символов.

Примеры использования символа ? и *:

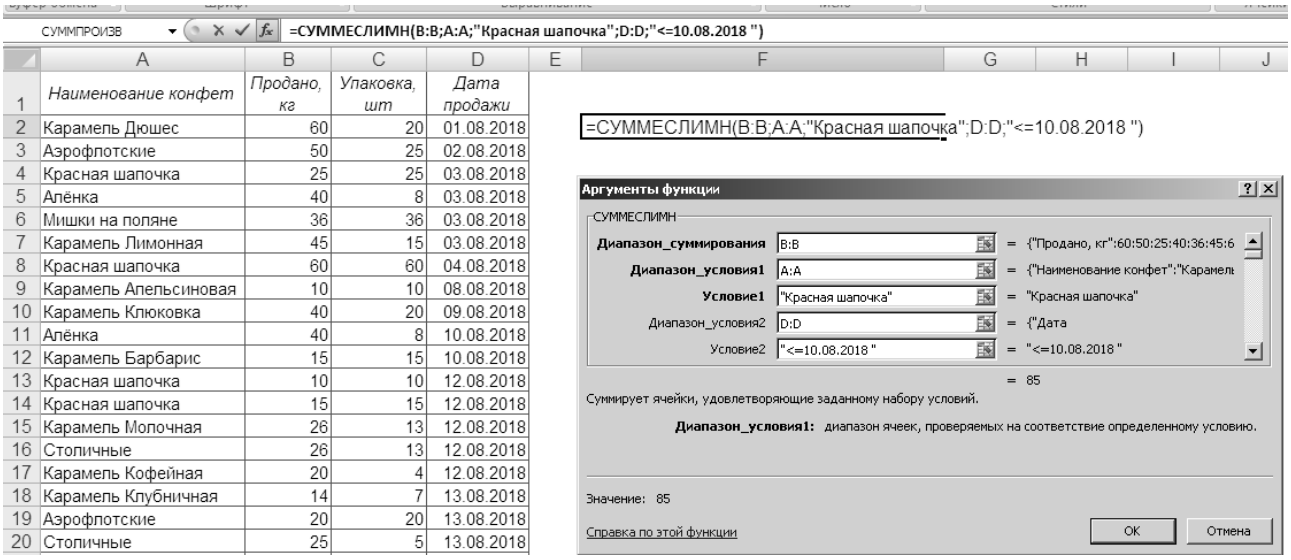
к?т – слово из трёх букв: первая – к, третья – т и обязательно один символ между ними. Может быть кит, кот, кэт, к-т, к8т и т. д.;

*дом – заканчивается на дом;

дом* – начинается с дом;

дом – содержит дом.

На рисунке 1.1 приведён пример использования функции СУММЕСЛИМН для вычисления объёма продаж конфет «Красная шапочка» в период до 10.08.2008 г.



	A	B	C	D
1	*Наименование конфет*	*Продано, кг*	*Упаковка, шт*	*Дата продажи*
2	Карамель Дюшес	60	20	01.08.2018
3	Аэрофлотские	50	25	02.08.2018
4	Красная шапочка	25	25	03.08.2018
5	Алёнка	40	8	03.08.2018
6	Мишки на поляне	36	36	03.08.2018
7	Карамель Лимонная	45	15	03.08.2018
8	Красная шапочка	60	60	04.08.2018
9	Карамель Апельсиновая	10	10	08.08.2018
10	Карамель Клюковка	40	20	09.08.2018
11	Алёнка	40	8	10.08.2018
12	Карамель Барбарис	15	15	10.08.2018
13	Красная шапочка	10	10	12.08.2018
14	Красная шапочка	15	15	12.08.2018
15	Карамель Молочная	26	13	12.08.2018
16	Столичные	26	13	12.08.2018
17	Карамель Кофейная	20	4	12.08.2018
18	Карамель Клубничная	14	7	13.08.2018
19	Аэрофлотские	20	20	13.08.2018
20	Столичные	25	5	13.08.2018

Рисунок 1.1 – Функция СУММЕСЛИМН

Задание 2

По предложенным данным рассчитайте:

– общий объём продаж всех конфет;

– объём продаж каждого вида конфет;

– объём продаж карамели (независимо от разновидностей);

– объём продаж конкретного вида конфет в указанный промежуток времени.

1.2 Статистические функции

МИН(Число1;Число2;) – вычисление наименьшего значения из списка аргументов, логические и текстовые значения игнорируются.

МАКС(Число1;Число2;) – вычисление наибольшего значения из списка аргументов, логические и текстовые значения игнорируются.

СРЗНАЧ(Число1;Число2;) – определение среднего арифметического своих аргументов, которые могут быть числами, именами или ссылками на ячейки с числами.

СЧЁТ(Значение1;Значение2;) – подсчитывает количество ячеек в диапазоне, которые содержат числа.

СЧЁТЗ(Значение1;Значение2;) – подсчитывает количество непустых ячеек в указанном диапазоне.

СЧИСТАТЬПУСТОТЫ(Диапазон) – подсчитывает количество пустых ячеек в указанном диапазоне.

СЧЁТЕСЛИ(Диапазон;Критерий) – подсчитывает количество ячеек в диапазоне, удовлетворяющее заданному условию.

Функция СЧЁТЕСЛИ подсчитывает количество ячеек только при выполнении одного критерия, если критериев несколько, то нужно использовать функцию СЧЁТЕСЛИМН.

СЧЁТЕСЛИМН(Диапазон_условия1;Условие1;) – подсчитывает количество ячеек в диапазоне, удовлетворяющее заданному набору условий.

СРЗНАЧЕСЛИ(Диапазон;Условие;Диапазон_усреднения) – подсчитывает среднее значение из диапазона, удовлетворяющего условию.

СРЗНАЧЕСЛИМН(Диапазон_усреднения;Диапазон_условия1;Условие1) подсчитывает среднее арифметическое для ячеек, удовлетворяющих заданному набору условий.

Задание 3

Имеется список сотрудников организации из 200 записей. О сотрудниках содержится следующая информация: фамилия, имя, отчество, пол, дата рождения, семейное положение, название отдела, должность, оклад.

С использованием вышеперечисленных функций определить:

– минимальный, максимальный, средний оклад;
– те же показатели отдельно для мужчин и женщин;
– те же показатели отдельно для каждого отдела;
– количество сотрудников и суммарную зарплату по отделам;
– количество молодых специалистов (возраст до 30 лет);
– определить количество сотрудников предпенсионного возраста (возраст более 60 лет).

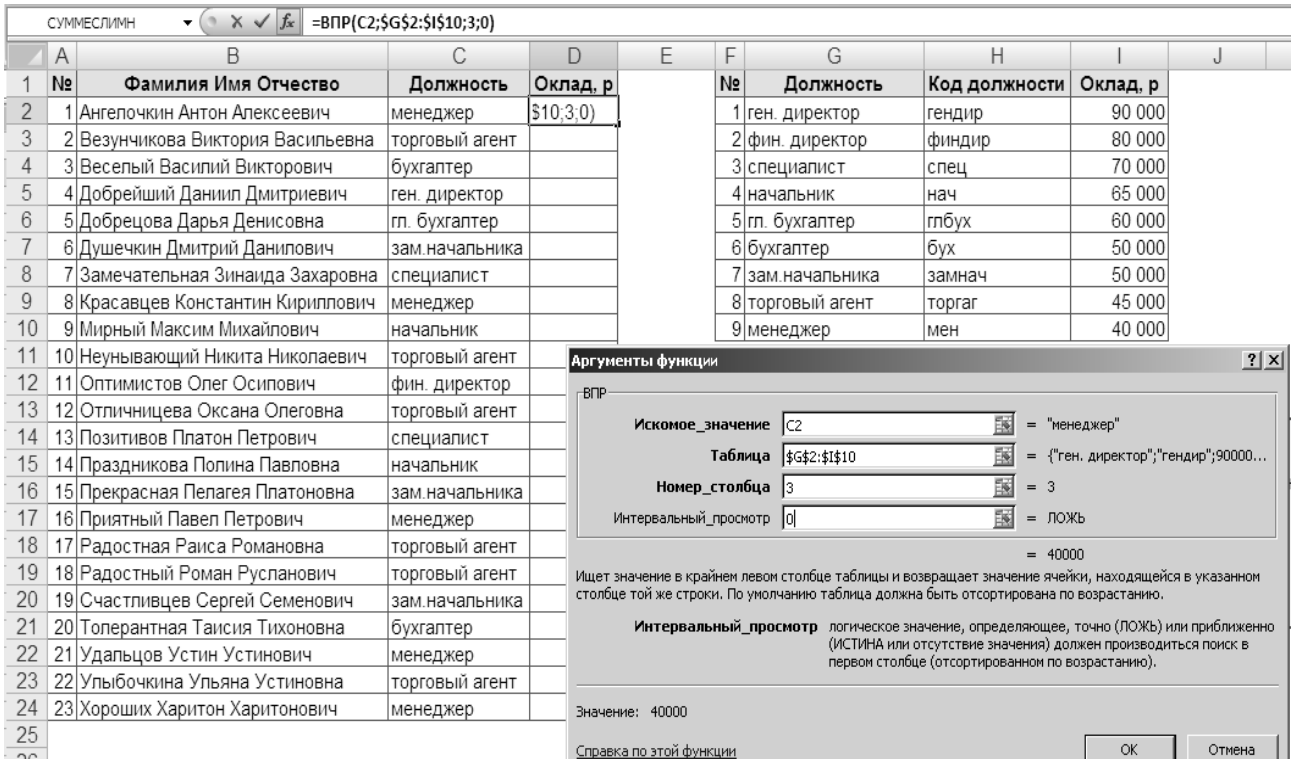
1.3 Функции ссылок и подстановки

При работе с большими таблицами для быстрого получения отдельных записей из этих списков можно использовать функции подстановок. Эти функции нужны для поиска связанных записей в таблицах. При использовании таких функций задача формулируется следующим образом – есть значение, для которого нужно найти совпадение в другой таблице и получить в ответ значение, которое хранится в ячейке соответствующей строки или столбца другой таблицы. Основное применение функций – подставлять данные, осуществлять сравнение двух таблиц.

Использование функций ВПР и ГПР зависит от расположения исходных данных в таблицах, из которых осуществляется подстановка.

В случае, если данные хранятся в столбцах, используется функция ВПР (применяется для вертикальных таблиц) (рисунок 1.2).

ВПР(Искомое_значение;Таблица;Номер_столбца;Интервальный_просмотр) – ищет значение в крайнем левом столбце таблицы и возвращает значение в той же строке из указанного столбца таблицы.



	A	B	C	D	E	F	G	H	I	J
1	№	Фамилия Имя Отчество	Должность	Оклад, р		№	Должность	Код должности	Оклад, р	
2	1	Ангелочкин Антон Алексеевич	менеджер	$10;3;0)		1	ген. директор	гендир	90 000	
3	2	Везунчикова Виктория Васильевна	торговый агент			2	фин. директор	финдир	80 000	
4	3	Веселый Василий Викторович	бухгалтер			3	специалист	спец	70 000	
5	4	Добрейший Даниил Дмитриевич	ген. директор			4	начальник	нач	65 000	
6	5	Добрецова Дарья Денисовна	гл. бухгалтер			5	гл. бухгалтер	глбух	60 000	
7	6	Душечкин Дмитрий Данилович	зам.начальника			6	бухгалтер	бух	50 000	
8	7	Замечательная Зинаида Захаровна	специалист			7	зам.начальника	замнач	50 000	
9	8	Красавцев Константин Кириллович	менеджер			8	торговый агент	торгаг	45 000	
10	9	Мирный Максим Михайлович	начальник			9	менеджер	мен	40 000	
11	10	Неунывающий Никита Николаевич	торговый агент							
12	11	Оптимистов Олег Осипович	фин. директор							
13	12	Отличницева Оксана Олеговна	торговый агент							
14	13	Позитивов Платон Петрович	специалист							
15	14	Праздникова Полина Павловна	начальник							
16	15	Прекрасная Пелагея Платоновна	зам.начальника							
17	16	Приятный Павел Петрович	менеджер							
18	17	Радостная Раиса Романовна	торговый агент							
19	18	Радостный Роман Русланович	торговый агент							
20	19	Счастливцев Сергей Семенович	зам.начальника							
21	20	Толерантная Таисия Тихоновна	бухгалтер							
22	21	Удальцов Устин Устинович	менеджер							
23	22	Улыбочкина Ульяна Устиновна	торговый агент							
24	23	Хороших Харитон Харитонович	менеджер							
25										

Рисунок 1.2 – Функция ВПР

В результате работы функции из столбца «Оклад» во второй таблице будет выбрано значение, для которого должности обоих таблиц будут совпадать.

В случае если данные хранятся в строках, то используется функция ГПР (применяется для горизонтальных таблиц).

ГПР(Искомое_значение;Таблица;Номер_строки;Интервальный_просмотр) – ищет значение в крайней верхней строке таблицы и возвращает значение в том же столбце из указанной строки таблицы.

Если не подходит функция ВПР или ГПР, то задачи можно решать с использованием функций ПОИСКПОЗ и ИНДЕКС.

Задание 4

1 Определить значение оклада для каждого сотрудника в зависимости от его должности по данным таблицы.

2 Определить значение стоимости доставки для каждого заказа в зависимости от его массы по данным таблицы.

3 Определить изменения в массе по каждому наименованию товара, которые произошли в отчётном году по сравнению с прошлым.

4 Вычислить значение бонуса с продажи как процент бонуса каждого сотрудника от стоимости заказа.

5 Определить по коду заказа значение кода клиента. Проверить, что код заказа 10500 отсутствует в исходной таблице.

6 Определить сумму доставки по значениям кода заказа.

1.4 Логические функции

Логические функции используются в случаях, когда результат обработки зависит от выполнения некоторого условия, заданного в виде логического выражения.

ЕСЛИ(Лог_выражение;Значение_если_истина;Значение_если_ложь) – возвращает одно значение, если заданное условие при вычислении дает значение ИСТИНА, и другое значение, если ЛОЖЬ.

Например, требуется рассчитать премию сотрудникам исходя из условия: если стаж сотрудника более 7 лет, то премия составляет 30 % от оклада, в противном случае – 50 р.

Для этого в ячейку E2 записываем следующую формулу: =ЕСЛИ(C2>7;0,3*D2;50)

В результате в ячейке появится значение 50. Для следующего сотрудника премия будет уже определяться по другому правилу (рисунок 1.3)

E3 | =ЕСЛИ(C3>7;0,3*D3;50)

	A	B	C	D	E	F
1	№	Ф.И.О.	Стаж работы	Оклад, р	Премия, р	Вывод
2	1	Антонов Антон Алексеевич	7	550	50	
3	2	Викторов Виктор Васильевич	14	420	126	
4	3	Васильев Василий Викторович	5	450	50	
5	4	Данилов Даниил Дмитриевич	10	650		
6	5	денисов Денис Давидович	12	700		
7	6	Дмитриев Дмитрий Данилович	8	750		

Рисунок 1.3 – Функция ЕСЛИ

Если имеется несколько условий и все они должны быть выполнены одновременно, используют функцию **И()**.

И(Логическое_значение1;Логическое_значение2;) – проверяет, все ли аргументы имеют значение ИСТИНА, и возвращает значение ИСТИНА, если истинны все аргументы.

ИЛИ(Логическое_значение1;Логическое_значение2;) – проверяет, имеет ли хотя бы один из аргументов значение ИСТИНА. Значение ЛОЖЬ возвращается только в том случае, если все аргументы имеют значение ЛОЖЬ.

Использование только функций И и ИЛИ позволяет получить ответ в ячейке как ИСТИНА или ЛОЖЬ, поэтому их часто используют в логической функции ЕСЛИ, чтобы задать более сложные условия.

Задание 5

По предложенным данным произвести следующие вычисления, используя логические функции.

1 Начислить премию сотрудникам исходя из условия: если стаж работы превышает 10 лет, то премия составляет 40 % от оклада, в противном случае – 60 р.

2 Определить доплаты сотрудникам в размере 20 р., которые работают в 1-ю или 3-ю смену.

3 Начислить премию сотрудникам в размере 100 р., которые работают более 5 лет и при этом их коэффициент надежности не менее 0,8.

4 Начислить годовую премию сотрудникам как коэффициент премии от оклада. Коэффициент зависит от стажа работы следующим образом:

- при стаже менее 5 лет – 2;
- при стаже от 5 до 10 лет включительно – 3;
- при стаже свыше 10 лет – 10.

5 Определить класс доступа сотрудника в зависимости от отдела, в котором он работает.

6 Скорректировать результат расчета премии сотрудников как произведение коэффициента и оклада. Коэффициенты каждого отдела для расчета премий указаны в отдельной таблице.

Контрольное задание

С использованием встроенных функций MS Excel произвести вычисления по предложенным преподавателем данных.

2 Лабораторная работа № 2. Анализ данных на наличие аномальных наблюдений и обработка пропусков

Цель работы: изучить методы и алгоритмы обработки аномальных наблюдений и пропусков.

Задачи: освоить различные способы поиска аномальных наблюдений и пропусков, применить различные варианты их заполнения.

2.1 Обработка пропущенных значений

Пропуски значений в наборах данных могут быть по нескольким причинам:

– отсутствие данных о некоторых характеристиках объекта;
– искажение или потеря информации;
– сокрытие данных.

Типы пропусков

1 MCAR (data are missing completely at random – пропуски в данных полностью случайные). Вероятность пропуска не зависит ни от значений наблюдаемых, ни от значений пропущенных данных.

2 MAR (missing at random – пропуски в данных случайные). Вероятность пропуска зависит от значений других факторов.

3 NMAR (missing not at random – вероятность пропуска зависит от значений и наблюдаемых, и от значений пропущенных данных).

Методы обработки пропущенных значений

1 Удаление объектов или признаков с пропущенными значениями.
2 Замена случайным значением.
3 Замена средним значением признака, медианой, модой.
4 Замена средним по группе (кластеру) значением признака, медианой, модой.
5 Вычисление пропущенных значений с помощью уравнения регрессии.

Задание 1

1 В заданном массиве данных подсчитать количество пропущенных значений по каждому фактору и по каждому объекту. Можно использовать встроенную функцию **СЧИСТАТЬПУСТОТЫ** (Диапазон) – подсчитывает количество пустых ячеек в указанном диапазоне.

2 Получите основные статистические характеристики выборки при помощи встроенного пакета анализа «Описательная статистика». (Закладка Данные/Анализ данных). Результат представлен на рисунке 2.1.

3 Замените пропуски в значениях первого признака средним его значением, второго признака – модой, третьего признака – медианой. Значения среднего, моды и медианы были получены на предыдущем шаге. Рекомендуется использовать встроенную логическую функцию

ЕСЛИ (Лог_выражение; Значение_если_истина; Значение_если_ложь). Для удобства вычислений рекомендуется преобразовать диапазон данных в таблицу. На вкладке **Вставка** в группе **Таблицы** выбрать **Таблица** (рисунок 2.2) и указать диапазон расположения данных таблицы.

Топливо Расход		*Пробег*		*Расход на 100 км*		*Дней в пути*	
Среднее	1032,742073	Среднее	3076,676829	Среднее	33,26927254	Среднее	9,213414634
Стандартная оши	51,49329363	Стандартная оши	149,2666283	Стандартная оши	0,218801	Стандартная ошибка	0,42495876
Медиана	751,5	Медиана	2192,5	Медиана	33,1825307	Медиана	8,5
Мода	433	Мода	4836	Мода	32,0754717	Мода	6
Стандартное откл	659,435913	Стандартное откл	1911,545531	Стандартное откл	2,767638053	Стандартное отклонен	5,442127475
Дисперсия выбор	434855,7233	Дисперсия выбор	3654006,318	Дисперсия выбор	7,659820393	Дисперсия выборки	29,61675146
Эксцесс	-0,686223263	Эксцесс	-0,923060113	Эксцесс	13,79048548	Эксцесс	-0,764372061
Асимметричност	0,398162005	Асимметричност	0,294729749	Асимметричност	2,179387183	Асимметричность	0,450476594
Интервал	3271	Интервал	9314	Интервал	24,73949339	Интервал	22
Минимум	45	Минимум	139	Минимум	27,82488661	Минимум	1
Максимум	3316	Максимум	9453	Максимум	52,56438	Максимум	23
Сумма	169369,7	Сумма	504575	Сумма	5323,083607	Сумма	1511
Счет	164	Счет	164	Счет	160	Счет	164

Рисунок 2.1 – Пример вывода результатов описательной статистики

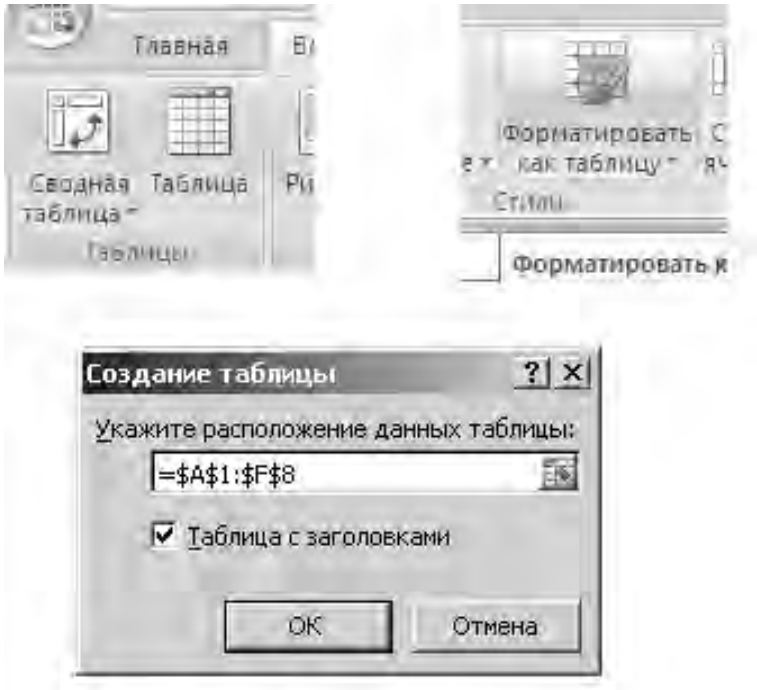


Рисунок 2.2 – Элементы преобразования диапазона данных в таблицу

Для вычисления новых данных достаточно написать формулу в одной ячейке. Формула автоматически будет скопирована вниз до окончания списка.

После заполнения пропущенных значений получите по обновлённому массиву данных новую описательную статистику (шаг 2). Результат разместите

рядом с предыдущим. Сравните, как изменились статистические характеристики признаков и объясните почему.

2.2 Обработка аномальных значений

Трудность обнаружения грубых ошибок обусловлена следующим обстоятельством. Если число измерений мало, то доверительный интервал широк, и даже значительные отклонения от среднего в него укладываются. Если же велико, то возрастает вероятность того, что хотя бы одно измерение сильно отклонится от среднего на «законных основаниях», т. е. случайно.

Рассмотрим несколько способов поиска аномальных значений в пакете Excel.

Выделите столбец, в котором требуется осуществить проверку данных. На закладке **Данные** выберите пункт «Проверка данных…» (рисунок 2.3).

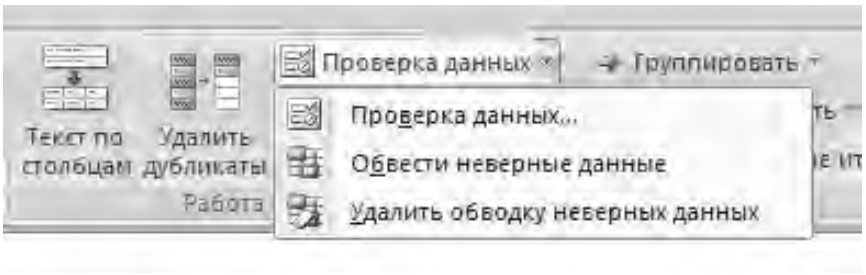


Рисунок 2.3 – Меню проверки данных

Предположим, требуется проверить корректность введённых значений в столбце «Возраст». В окне «Проверка вводимых значений» укажите, что это должно быть целое число, которое находится в диапазоне от 0 до 120. При желании можно ввести текст сообщения об ошибке (рисунок 2.4). Нажмите кнопку «ОК».

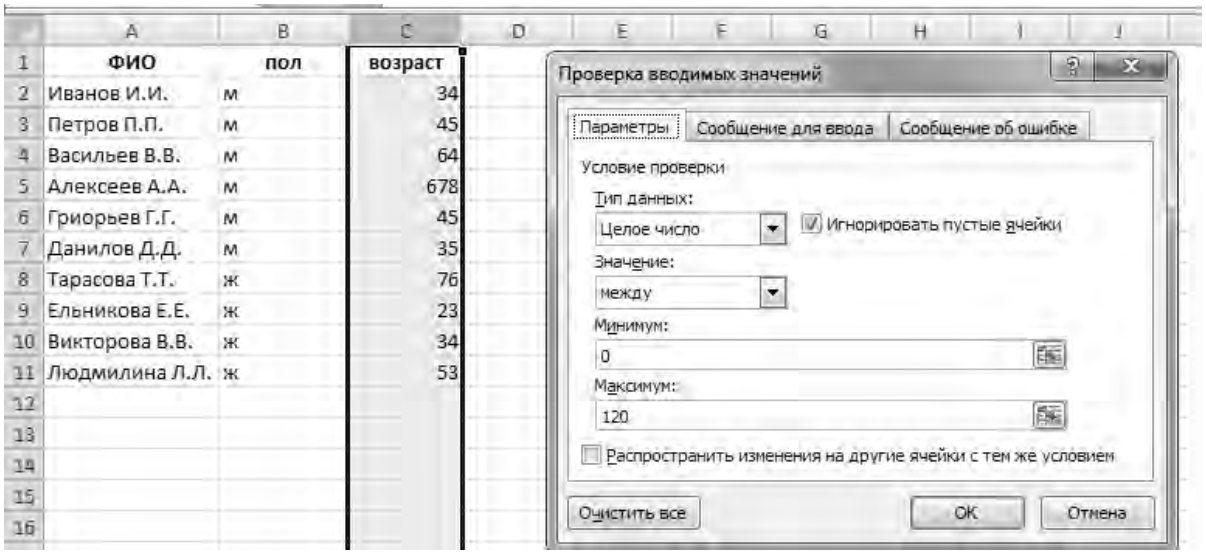


Рисунок 2.4 – Определение параметров проверяемых значений

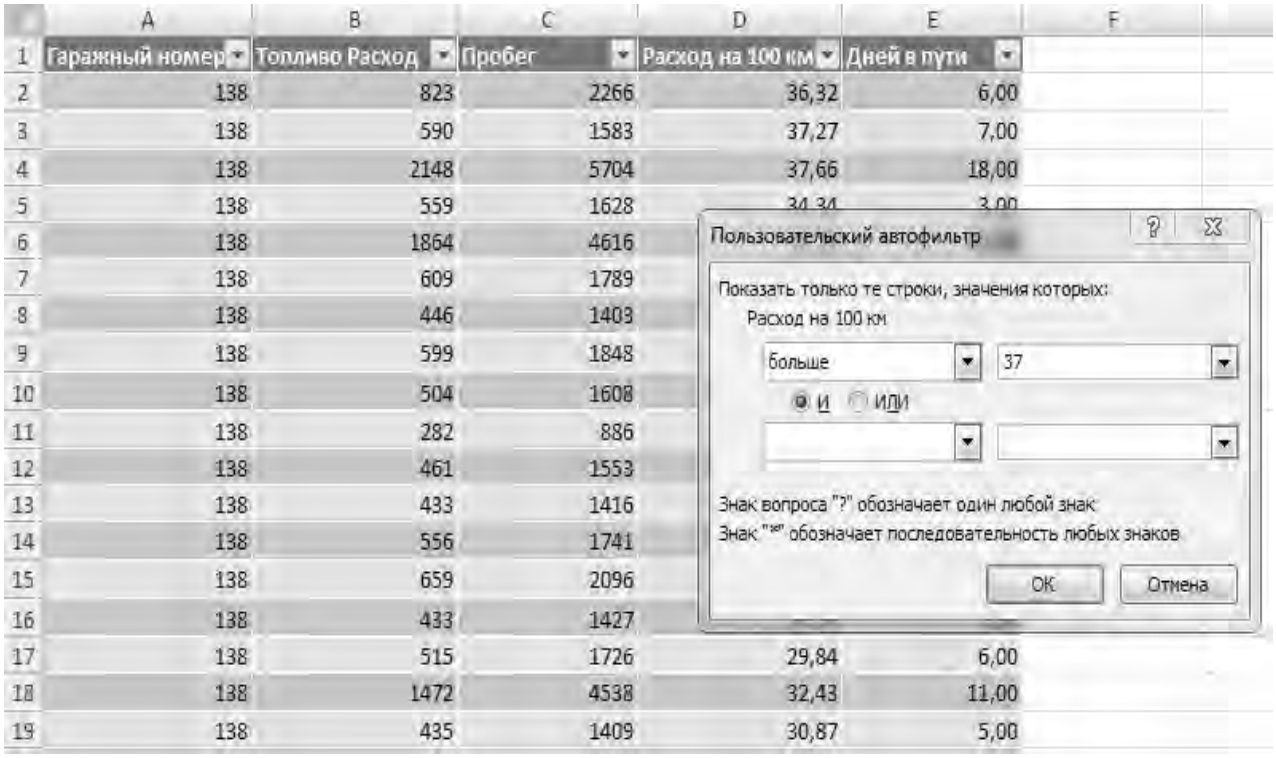
Далее выберите «Проверка данных → Обвести неверные значения». Результат будет следующий (рисунок 2.5)

	A	B	C
1	**ФИО**	**пол**	**возраст**
2	Иванов И.И.	м	34
3	Петров П.П.	м	45
4	Васильев В.В.	м	64
5	Алексеев А.А.	м	678
6	Гриорьев Г.Г.	м	45
7	Данилов Д.Д.	м	35
8	Тарасова Т.Т.	ж	76
9	Ельникова Е.Е.	ж	23
10	Викторова В.В.	ж	34
11	Людмилина Л.Л.	ж	53
12			

Рисунок 2.5 – Выделение подозрительных значений

Для этих же целей можно использовать графики и условное форматирование.

При работе с большими массивами данных для выявления аномальных значений удобно использовать фильтры, предварительно преобразовав диапазон в таблицу. В нашем примере, применив к полю «Расход топлива на 100 км» условие «больше 37», из 164 данных останется только 11 (рисунок 2.6).



	A	B	C	D	E
1	Гаражный номер	Топливо Расход	Пробег	Расход на 100 км	Дней в пути
2	138	823	2266	36,32	6,00
3	138	590	1583	37,27	7,00
4	138	2148	5704	37,66	18,00
5	138	559	1628	34,34	3,00
6	138	1864	4616		
7	138	609	1789		
8	138	446	1403		
9	138	599	1848		
10	138	504	1608		
11	138	282	886		
12	138	461	1553		
13	138	433	1416		
14	138	556	1741		
15	138	659	2096		
16	138	433	1427		
17	138	515	1726	29,84	6,00
18	138	1472	4538	32,43	11,00
19	138	435	1409	30,87	5,00

Рисунок 2.6 – Выделение подозрительных значений

В прикладном статистическом анализе к аномальным значениям принято относить те, которые отклоняются от среднего на величину трёх среднеквадратических отклонений (правило трёх сигм). Найдём в таблице значения, выходящие за указанные пределы с помощью условного форматирования. Выделяем столбец с данными о суммах выданных кредитов. На закладке **Главная** выбираем меню «Условное форматирование → Правила выделения ячеек → Другие правила». В диалоговом окне указываем критерии, которым должны соответствовать выделяемые ячейки (рисунок 2.7): для суммы кредита – правило трёх сигм, для возраста – меньше 18 и больше 120.

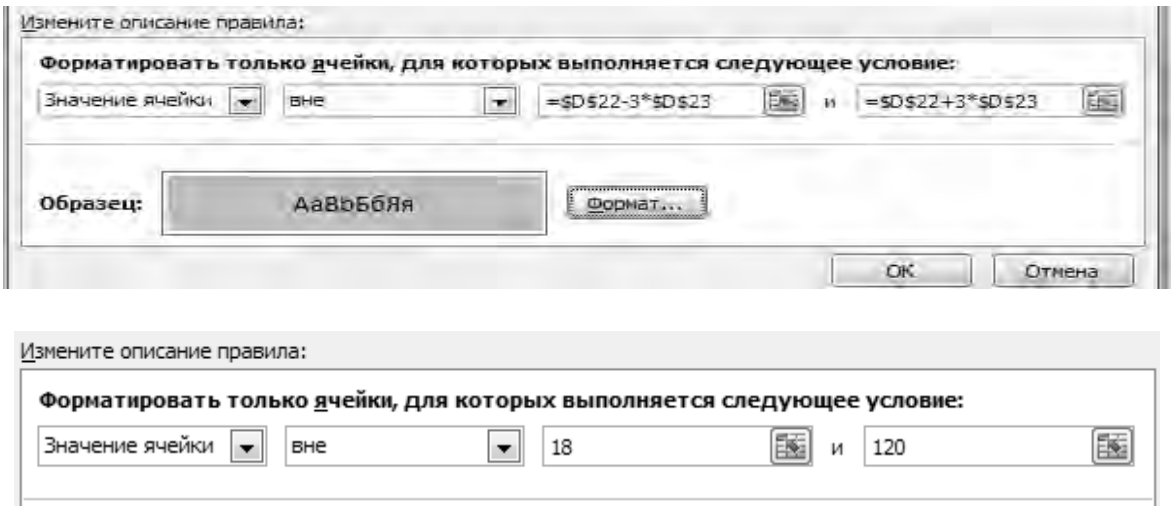


Рисунок 2.7 – Создание правила для выделения значений

Таблица с исходными данными и результатом условного форматирования приведена на рисунке 2.8.

После нахождения аномальных значений к ним применяют одно из следующих действий:

– удаление записи с аномальным значением;
– ручная замена аномальных значений;
– сглаживание и фильтрация данных;
– интерполяция данных;
– замена на наиболее вероятное значение.

Контрольные вопросы и задания

1 Какие методы замены пропущенных значений при каких типах пропусков допускается применять?

2 Продемонстрируйте на примере алгоритм выявления признаков с пропущенными значениями.

3 Назовите причины появления аномальных значений.

4 Какие существуют методы их выявления?

	A	B	C	D
1	**ФИО**	**пол**	**возраст**	**сумма кредита**
2	Алексеев А.А.	м	65	12 000
3	Борисова Б.Б.	ж	27	28 000
4	Васильев В.В.	м	64	64 000
5	Викторова В.В.	ж	34	987 000
6	Воробьёв В.В.	м	358	73 000
7	Воронова В.В.	ж	58	53 000
8	Грибов Г.Г.	м	37	54 000
9	Гриорьев Г.Г.	м	45	46 000
10	Данилов Д.Д.	м	35	34 000
11	Ельникова Е.Е.	ж	23	65 000
12	Иванов И.И.	м	34	10 000
13	Лесная Л.Л.	ж	15	79 000
14	Людмилина Л.Л.	ж	53	23 000
15	Михайлов М.М.	м	23	25 000
16	Петров П.П.	м	45	35 000
17	Тарасова Т.Т.	ж	76	78 000
18	Титов Т.Т.	м	43	64 000
19	Юрьева Ю.Ю.	ж	42	43 000
20	Яблочкин Я.Я.	м	46	25 000
21				
22	**Среднее**		59,11	94631,58
23	**Среднеквадратическое отклонение**		74,07	217175,10
24				

Рисунок 2.8 – Выделение аномальных значений в массиве данных

3 Лабораторная работа № 3. Анализ одномерной количественной выборки

Цель работы: научиться определять основные характеристики одномерной количественной выборки.

Задачи: рассчитать и пояснить значение основных статистических характеристик одномерной выборки.

Целью одномерного анализа является описание одной характеристики выборки в определенный момент времени. Описательная статистика является базовым и наиболее общим методом анализа данных. Она предполагает определение следующих характеристик:

- среднее;
- дисперсия выборки;
- стандартное отклонение;
- стандартная ошибка;
- минимум;
- максимум;

– сумма;
– асимметрия;
– эксцесс;
– медиана;
– мода;
– количество наблюдений.

Для проверки соответствия нормальному закону распределения строится гистограмма частот и значение критерия, определяются эмпирические и теоретические частоты.

В MS Excel основные статистические характеристики выборки можно получить при помощи встроенного пакета анализа «Описательная статистика». (Закладка Данные/Анализ данных) (рисунок 3.1).

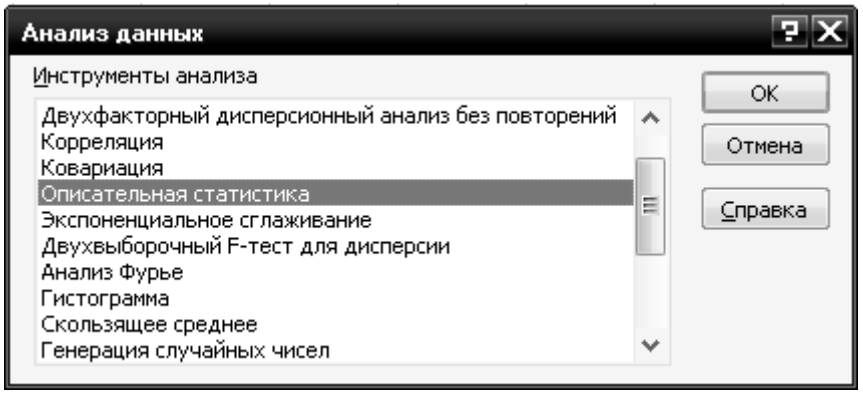


Рисунок 3.1 – Инструменты анализа данных

Задание

По предложенной одномерной выборке рассчитать все вышеперечисленные характеристики ряда. Проверить гипотезу о соответствии случайной величины нормальному закону распределения.

Контрольные вопросы

1 Пояснить смысл каждого из рассчитанных параметров одномерной выборки.

2 Для каких целей проводится тот или иной вид анализа?

4 Лабораторная работа № 4. Решение задач на проверку параметрических гипотез

Цель работы: научиться выдвигать и проверять параметрические гипотезы.

Задачи: сформулировать нулевую и альтернативную параметрические гипотезы и проверить их по предложенному ниже алгоритму.

Статистические гипотезы называются параметрическими? если выдвинутое предположение касается неизвестного значения параметра определённого вида распределения. Предполагается, что случайная величина имеет нормальный закон распределения.

Алгоритм проверки статистических гипотез

1 По выборочным данным формулируют основную H_0 и альтернативную H_1 гипотезы.

2 Задают уровень значимости α (0,05 или 0,01).

3 В зависимости от H_0 определяют статистический критерий K, имеющий известное распределение.

4 По выборке и формуле критерия K рассчитывают наблюдаемое значение критерия $K_{набл}$.

5 В зависимости от вида H_1 определяют вид критической области W и критические точки по соответствующим таблицам для распределения критерия K.

6 По результатам проверки принадлежности $K_{набл}$ критической области делают вывод о принятии или отклонении гипотезы H_0. Формулируют общий вывод, исходя из поставленной задачи.

Проверка гипотез о равенстве числовому параметру:

– дисперсии $\sigma^2 = \sigma_0^2$;

– математического ожидания $\mu = \mu_0$;

– вероятности $p = p_0$.

Проверка гипотез о равенстве числовых характеристик:

– дисперсии $D(X) = D(Y)$;

– математических ожиданий $M(X) = M(Y)$;

– вероятностей $p_X = p_Y$.

Для проверки параметрических гипотез в зависимости от количества рассматриваемых данных, используются следующие методы:

– t-тест для одной выборки;

– t-тест для двух зависимых выборок;

– t-тест для двух независимых выборок;

– для трех и более выборок применяется однофакторный или многофакторный дисперсионный анализ.

t-тест состоит в расчете *t*-статистики (коэффициента Стьюдента) по имеющимся данным, сравнение его с табличным значением и формулированием вывода о равенстве средних.

Все вышеуказанные тесты реализованы в надстройке Excel «Анализ данных» (рисунок 4.1).

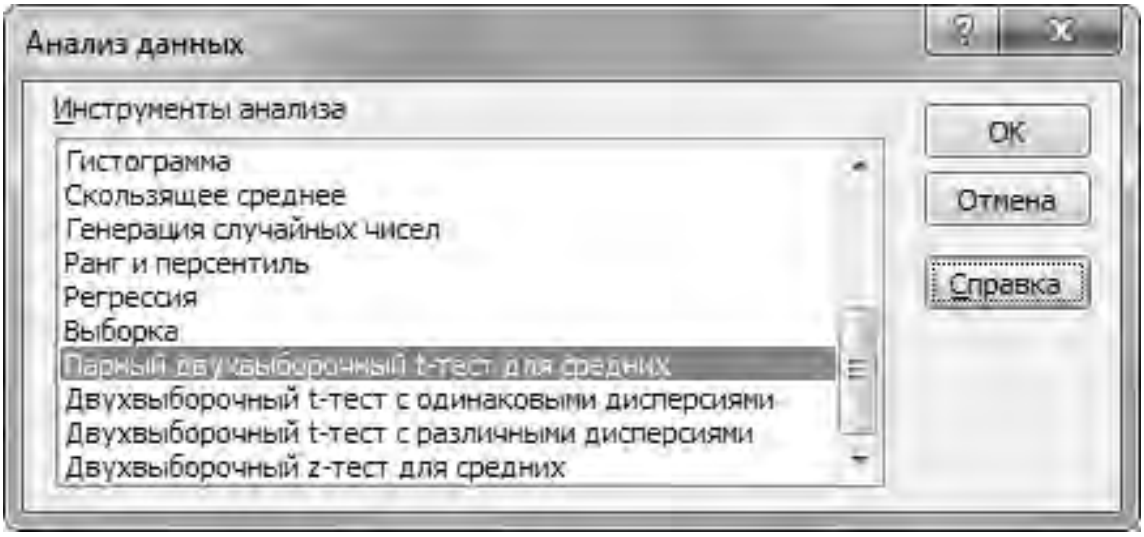


Рисунок 4.1 – Инструменты для проверки гипотез в Excel

Двухвыборочный *t*-тест проверяет равенство средних значений генеральной совокупности по каждой выборке. Разновидности этого теста допускают следующие условия: равные дисперсии генерального распределения, дисперсии генеральной совокупности не равны, а также представление двух выборок до и после наблюдения по одному и тому же субъекту.

Двухвыборочный F-тест применяется для сравнения дисперсий двух генеральных совокупностей. Например, можно использовать *F*-тест по выборкам результатов заплыва для каждой из двух команд. Это средство предоставляет результаты сравнения нулевой гипотезы о том, что эти две выборки взяты из распределения с равными дисперсиями, с гипотезой, предполагающей, что дисперсии различны в базовом распределении.

Парный двухвыборочный t-тест для средних. Парный тест используется, когда имеется естественная парность наблюдений в выборках, например, когда генеральная совокупность тестируется дважды – до и после эксперимента. Этот инструмент анализа применяется для проверки гипотезы о различии средних для двух выборок данных. В нем не предполагается равенство дисперсий генеральных совокупностей, из которых выбраны данные.

Двухвыборочный t-тест с различными дисперсиями. Этот инструмент анализа выполняет двухвыборочный *t*-тест Стьюдента, который используется для проверки гипотезы о равенстве средних для двух выборок данных из разных генеральных совокупностей. Эта форма *t*-теста предполагает несовпадение дисперсий генеральных совокупностей и обычно называется

гетероскедастическим *t*-тестом. Если тестируется одна и та же генеральная совокупность, необходимо использовать парный тест.

Инструмент анализа «***Двухвыборочный z-тест для средних***» выполняет двухвыборочный *z*-тест для средних с известными дисперсиями, который используется для проверки основной гипотезы о неразличии между средними двух генеральных совокупностей относительно односторонней и двусторонней альтернативных гипотезах. При неизвестных значениях дисперсий следует воспользоваться функцией *z*-тест.

Задача 1. По результатам $n = 7$ независимых измерений найдено, что $\bar{x} = 82{,}48$ мм, а $S = 0{,}08$ мм. Допустив, что ошибки измерения имеют нормальное распределение проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0: $\sigma^2 = 0{,}01$ мм^2. против конкурирующей гипотезы H_0: $\sigma^2 = 0{,}005$ мм^2.

Задача 2. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2 мкм^2. Выборочная дисперсия, найденная по измерениям 121 детали, оказалась равной 0,3 мкм^2. Предполагая нормальное распределение размеров, определить, можно ли принять партию при уровне значимости 0,01.

Задача 3. Часовая выработка рабочих механического цеха имеет характеристику рассеяния 1,4 дет^2/ч^2. После чего с целью проверки соответствия нормативной выработке (равной 21 дет/ч) была сформирована выборка из 36 рабочих. На базе выборки найдена средняя выработка рабочего, составившая 21,6 дет/ч. Предполагая нормальное распределение выработки рабочих цеха определить: соответствует ли средняя выработка станочников выработке по норме (принять $\alpha = 0{,}05$)?

Задача 4. С целью проверки соответствия нормативной выработке рабочих механического цеха (равной 15 дет/ч) была проведена оценка часовой выработки 20 станочников. Результаты выборочной проверки показали, что средняя выработка рабочего составляет 16 дет/ч, а выборочная дисперсия – 4,28 дет^2/ч^2. Предполагая нормальное распределение выработки рабочих цеха определить: соответствует ли средняя выработка станочников выработке по норме (принять $\alpha = 0{,}05$).

Задача 5. Компания не осуществляет инвестиционных вложений в ценные бумаги с дисперсией годовой доходности более чем 0,04. Выборка из 52 наблюдений по активу А показала, что выборочная дисперсия ее доходности равна 0,045. Выяснить, допустимы ли для данной компании инвестиционные вложения в актив А на уровне значимости: 0,05; 0,01.

Контрольные вопросы и задания

1 Как формулируются статистические параметрические гипотезы?

2 По каким критериям оцениваются вероятности ошибок принятия параметрической гипотезы?

3 Каков алгоритм проверки статистических параметрических гипотез?
4 Какие выводы делаются по результатам статистической проверки?
5 Какие возможности предоставляет Excel для проверки гипотез?

5 Лабораторная работа № 5. Решение задач на проверку непараметрических гипотез

Цель работы: научиться выдвигать и проверять непараметрические гипотезы.

Задачи: сформулировать нулевую и альтернативную непараметрические гипотезы и проверить их на основе рассмотренных ниже критериев.

Статистические гипотезы называются непараметрическими, при проверке которых вид распределения неизвестен.

Одним из факторов, ограничивающих применения критериев, основанных на предположении нормальности, является объем выборки. До тех пор пока выборка достаточно большая (100 и более наблюдений), можно считать, что выборочное распределение нормально. Если выборка мала, эти критерии следует использовать только при наличии уверенности, что переменная действительно имеет нормальное распределение. На малой выборке нет возможности проверить это предположение.

Для анализа малых и применяют непараметрические методы.

Различия между независимыми группами. Если имеются две выборки (например, мужчины и женщины), которые нужно сравнить относительно некоторого среднего значения, например, среднего давления или количества лейкоцитов в крови, то можно использовать *t*-тест для независимых выборок.

Непараметрическими альтернативами этому тесту являются критерий серий Вальда – Вольфовица, Манна – Уитни (*z*-тест и двухвыборочный критерий Колмогорова – Смирнова.

Различия между зависимыми группами. Если вы хотите сравнить две переменные, относящиеся к одной и той же выборке, например, влияние нового метода организации процесса на производительность оборудования, обычно используется *t*-критерий для зависимых выборок.

Альтернативными непараметрическими тестами являются критерий знаков и критерий Вилкоксона.

Проверка гипотез о законе распределения:

– критерий согласия Пирсона (χ^2);
– критерий согласия Колмогорова $F_{эмп}(x_i) \leftrightarrow F_{теорет}(x_i)$.

Проверка гипотез об однородности выборок:

– критерий Колмогорова-Смирнова $F_{эмп\,1}(x_i) \leftrightarrow F_{эмп\,2}(x_i)$;

– ранговый критерий Вилкоксона.

Задание

Решите задачи из лабораторной работы № 4 без учёта информации о нормальности распределения случайных величин с применением непараметрических критериев.

Контрольные вопросы и задания

1 Как формулируются статистические непараметрические гипотезы?
2 По каким критериям оцениваются вероятности ошибок принятия непараметрической гипотезы?
3 Каков алгоритм проверки статистических непараметрических гипотез?
4 Какие выводы делаются по результатам статистической проверки?
5 Какие возможности предоставляет Excel для проверки гипотез?

6 Лабораторная работа № 6. Построение и эконометрический анализ двумерной регрессионной модели

Цель работы: научиться подбирать уравнение для описания стохастической зависимости между случайными величинами.

Задачи: установить наличие зависимости между случайными величинами, подобрать уравнение регрессии и оценить его качество.

Уравнения двумерной линейной регрессии для факторов x и y имеет следующий вид:

$$\hat{y} = b_0 + b_1 \cdot x .$$

Исследование зависимости между случайными величинами начинается с определения направления и тесноты связи при помощи коэффициента корреляции.

При построении уравнения регрессии для нахождения коэффициентов используется метод наименьших квадратов.

В MS Excel для определения взаимосвязи между двумя признаками используется функция КОРРЕЛ(массив1; массив2).

В случае многомерной выборки (более двух факторов) следует воспользоваться встроенным пакетом анализа данных «Корреляционный анализ».

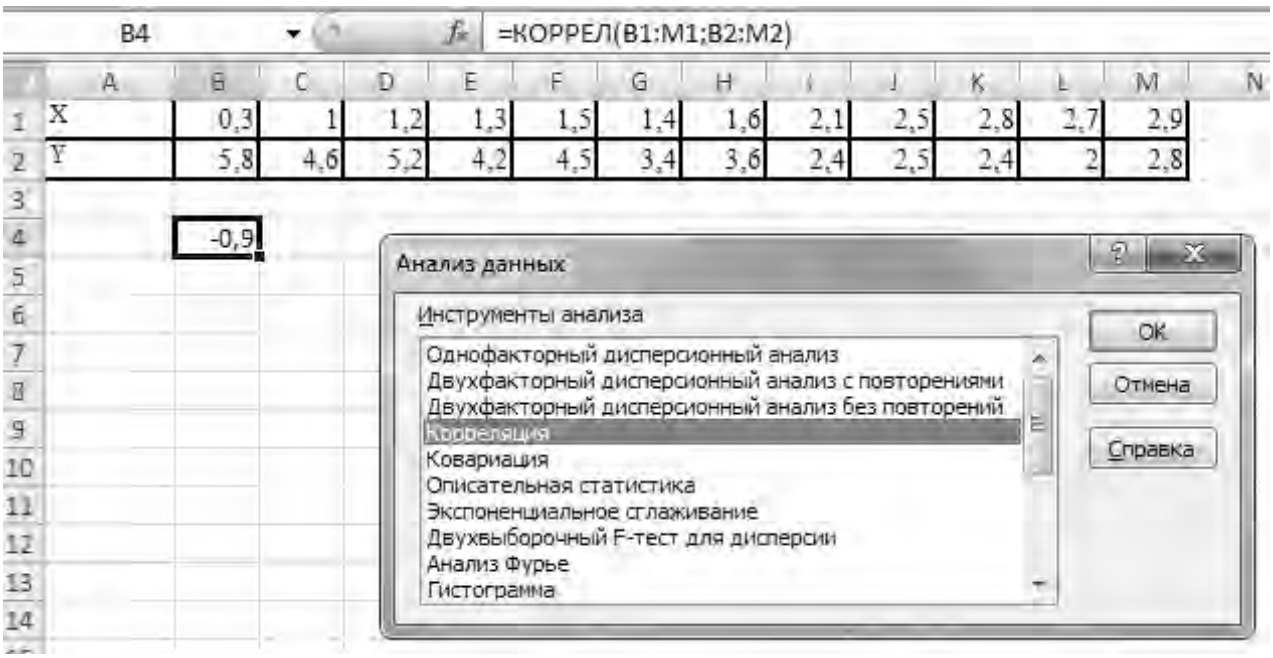


Рисунок 6.1 – Корреляционный анализ

Для получения параметров уравнения линейной регрессии предназначен пакет анализа данных «Регрессия». Результат его работы представлен на рисунке 6.2.

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный	0,91282							
R-квадрат	0,83324							
Нормированный	0,816564							
Стандартная оши	0,530819							
Наблюдения	12							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	1	14,07897392	14,07897392	49,96632	3,421E-05			
Остаток	10	2,817692742	0,281769274					
Итого	11	16,89666667						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	6,06789	0,37912	16,0052	1,87E-08	5,223156	6,912616	5,223156	6,912616
X	-1,38097	0,19536	-7,0687	3,42E-05	-1,816267	-0,945670	-1,816267	-0,945670

Рисунок 6.2 – Нахождение параметров множественной (парной) регрессии при помощи пакета анализа данных MS Excel

Отдельные характеристики линейной регрессии можно получить при помощи некоторых встроенных функций Excel.

ОТРЕЗОК (известные_значения_x;известные_значения_y) – вычисляет коэффициент b_0.

НАКЛОН (известные_значения_y;известные_значения_x) – возвращает наклон линии линейной регрессии (коэффициент b_1).

СТОШYX (известные_значения_y;известные_значения_x) – возвращает стандартную ошибку предсказанных значений *y* для каждого значения *x* в регрессии.

КВПИРСОН (известные_значения_y;известные_значения_x) – возвращает квадрат коэффициента корреляции Пирсона.

ПРЕДСКАЗ (x;известные_значения_y;известные_значения_x) – вычисляет или предсказывает будущее значение по существующим значениям.

После определения параметров уравнения регрессии проводится оценка их значимости. Проверка значимости каждого параметра уравнения регрессии осуществляется на основе соответствующих *t*-статистик. Для них определяются критические значения или расчетные уровни значимости, на основе которых и принимаются решения о значимости или незначимости соответствующих оценок.

Полученные точечные оценки дополняются интервальными оценками. Доверительные интервалы дают дополнительную информацию о надежности точечных оценок и позволяют повысить надежность суждений о точечных оценках.

Для определения доверительных интервалов используются *t*-статистики Стьюдента.

Задача 1. По 12 торговым точкам проводился анализ взаимосвязи объёмов продаж *x* и цены товара *y*. Признаки *x* и *y* имеют нормальный закон распределения (таблица 6.1).

Таблица 6.1 – Исходные данные

X	0,4	1,2	1,4	1,6	1,8	1,7	1,9	2,5	3,0	3,4	3,2	3,5
Y	7,0	5,5	6,2	5,0	5,4	4,1	4,3	2,9	3,0	2,9	2,4	3,4

Задание

1 Постройте корреляционное поле и сформулируйте гипотезу о форме связи между ценой товара А и объемом продаж данного товара.

2 Рассчитайте оценки параметров уравнения парной линейной регрессии.

3 Оцените тесноту связи между ценой товара А и объемом продаж данного товара с помощью парного коэффициента корреляции. Проверьте значимость коэффициента корреляции ($\alpha = 0{,}05$).

4 Рассчитайте выборочный коэффициент детерминации. Сделайте экономический вывод.

5 Проверьте значимость оценки коэффициента регрессии с помощью критерия Стьюдента при уровне значимости $\alpha = 0{,}05$.

6 Постройте доверительный интервал для коэффициента регрессии. Дайте экономическую интерпретацию.

7 Составьте таблицу дисперсионного анализа.

8 Оцените с помощью *F*-критерия Фишера – Снедекора значимость уравнения линейной регрессии ($\alpha = 0{,}05$).

9 Рассчитайте объем продаж данного товара, если его цена составит 11 долл. Постройте доверительный интервал для прогнозного значения объясняемой переменной. Сделайте экономический вывод.

10 Рассчитайте средний коэффициент эластичности $\overline{Э}$. Сделайте экономический вывод.

11 Определите среднюю ошибку аппроксимации.

12 На поле корреляции постройте линию регрессии.

Контрольные вопросы

1 Дайте понятие двумерной регрессионной модели, цель её построения.
2 По каким критериям оценивается качество регрессионной модели?
3 Что такое доверительный интервал и уровни значимости?
4 Как оценивается значимость коэффициентов уравнения?
5 Поясните отдельные этапы построения модели.

7 Лабораторная работа № 7. Использование фиктивных переменных в регрессионном анализе

Цель работы: научиться использовать качественные характеристики при построении уравнения регрессии.

Задачи: учесть в регрессионном уравнении качественные характеристики объекта, проверить значимость его влияния на результирующий признак.

В регрессионных моделях наряду с количественными переменными часто используются качественные переменные, такие как профессия, пол, образование, климатические условия и т. п. Такого рода переменные в эконометрике называются *фиктивными* (*структурными*, или *искусственными*) переменными, а также *индикатором*. Чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные *цифровые метки*, т. е. качественные переменные должны быть преобразованы в количественные.

Фиктивные переменные отражают два противоположных состояния качественного фактора: фактор действует – фактор не действует. (Например, сезон летний – сезон зимний, пол мужской – пол женский, есть высшее образование – нет высшего образования).

В этом случае фиктивные переменные выражаются в двоичной форме

$$z = \begin{cases} 0, \ \textit{признак не действует}; \\ 1, \ \textit{признак действует}. \end{cases}$$

(Например, $z = 0$, если потребитель не имеет высшего образования, $z = 1$, если потребитель имеет высшее образование).

Таким образом, кроме моделей, содержащих только количественные переменные x_i, в регрессионном анализе рассматриваются также модели, содержащие лишь качественные переменные z_i, либо те и другие одновременно.

Простейшая модель с одной количественной и одной качественной переменными имеет вид:

$$y = a + b \cdot x + g \cdot z + e,$$

где y – заработная плата сотрудника фирмы;

x – стаж работы;

z – пол сотрудника.

Если

$$z = \begin{cases} 0, \ \textit{если сотрудник женщина}; \\ 1, \ \textit{если сотрудник мужчина}, \end{cases}$$

тогда для женщин ожидаемое значение заработной платы при x годах трудового стажа будет $\hat{y} = a + b \cdot x$, а для мужчин – $\hat{y} = a + b \cdot x + g = (a + g) + b \cdot x$.

Эти зависимости являются линейными относительно стажа работы x и различаются только величиной свободного члена. Если коэффициент g является статистически значимым, то можно сделать вывод, что в фирме имеет место дискриминация в заработной плате по половому признаку. При $g > 0$ она будет в пользу мужчин, при $g < 0$ – в пользу женщин. На графике такие зависимости изображаются параллельными прямыми.

Задача 1. Дан набор данных о сдельной зарплате производственных рабочих, который включает табельный номер, ФИО работника, пол, производственный участок, зарплата.

Требуется:

– сформулировать и проверить гипотезу о существовании существенных различий в зарплате, во-первых, между мужчинами и женщинами, во-вторых, между производственными участками;

– построить уравнение регрессии для определения величины заработной платы в зависимости от пола работника и/или производственного участка.

Какой вывод можно сделать, если коэффициент при фиктивной переменной положительный, а какой – если отрицательный?

Задача 2. Дан набор данных о продаже прохладительных напитков в нескольких торговых точках, который включает номер торговой точки, торговую площадь, дату продажи, наименование напитка, количество, цену, стоимость.

Требуется:

– сформулировать и проверить гипотезу о существовании существенных различий в объёмах продаж напитков в зависимости, во-первых, между торговыми точками, во-вторых, в зависимости от сезона (предварительно добавить фактор «сезон», который заполнить на основе даты продаж);

– построить уравнение регрессии для определения объёмов продаж напитков в зависимости от торговой точки, торговой площади и сезона.

Какой вывод можно сделать, если коэффициент при фиктивной переменной положительный, а какой – если отрицательный?

Контрольные вопросы

1 Поясните смысл фиктивных переменных.

2 В каких случаях фиктивные переменные добавляют в уравнение регрессии?

3 Как интерпретируются коэффициенты при фиктивных переменных?

4 По какому критерию определяется значимость коэффициента при фиктивной переменной.

8 Лабораторная работа № 8. Разбиение совокупности объектов на группы различными методами

Цель работы: научиться выполнять группировку объектов, характеризующихся несколькими признаками, различными методами с использованием программных инструментов.

Задачи: изучить существующие методы кластеризации и особенности их применения, рассмотреть различные способы измерения расстояния между классами, применить графические методы для визуализации результатов.

Методы кластерного анализа позволяют выделить из исследуемой совокупности объектов *кластеры* – скопления объектов с близкими значениями параметров.

Методы кластерного анализа можно разделить на две большие категории по алгоритму действия. Первая группа методов называется *иерархическими*, т. к. в процессе работы метода строится иерархия вложенности кластеров, обычно представляемая на графике – *дендрограмме*. На каждом шаге агломеративной иерархической процедуры объединяется пара ближайших кластеров. Методы второй категории называются *итерационными*, т. к. они

основаны на поиске оптимального положения центров кластеров на каждой итерации – последовательного рассмотрения всех объектов исходной выборки.

Среди итерационных методов наиболее распространённым *является метод k-средних*. На первом его шаге необходимо задать требуемое количество кластеров k и начальные центры их тяжести. В качестве этих начальных центров обычно используются первые k наблюдений выборки, однако в некоторых случаях это может привести к недостаточному качеству полученного решения. Поэтому возможно использовать иерархическую процедуру на случайной выборке и затем использовать полученные центры в итерационной процедуре.

Чаще всего близость объектов x и y измеряется с помощью следующих метрик расстояния, если их характеристики измерены в интервальной шкале:

- расстояние Евклида: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$;
- расстояние Манхэттена: $\sum_{i=1}^{n}|x_i - y_i|$;
- расстояние Чебышева: $\max_{i=1,N}|x_i - y_i|$.

Расстояние между кластерами определяется с помощью следующих основных методов:

- связь между группами – расстояние между двумя кластерами определяется как среднее значение расстояний между всеми возможными парами наблюдений, причём одно наблюдение берётся из одного кластера, а второе – из другого;
- связь внутри групп – расстояние между двумя кластерами определяется как среднее значение расстояний между всеми возможными парами наблюдений из обоих кластеров, включая пары наблюдений внутри кластеров;
- ближний сосед – расстояние между двумя кластерами определяется как минимальное из всех расстояний между всеми возможными парами наблюдений из разных кластеров;
- дальний сосед – расстояние между двумя кластерами определяется как максимальное из всех расстояний между всеми возможными парами наблюдений из разных кластеров;
- центроидная кластеризация – расстояние между двумя кластерами определяется как расстояние между центрами тяжести обоих кластеров;
- медианная кластеризация – расстояние между двумя кластерами определяется как взвешенное центроидное расстояние между кластерами, где веса соответствуют размеру каждого кластера;
- метод Варда – в этом методе объединяются только те два кластера, для которых прирост внутрикластерной дисперсии минимален.

Наиболее универсальными методами являются метод Варда и метод межгрупповой связи.

Задача 1. По *иерархическому агломеративному алгоритму* провести классификацию $n = 10$ предприятий, данные о деятельности которых характеризуются показателями рентабельность $x1$ и производительность труда $x2$. Данные представлены в таблице 8.1.

Таблица 8.1 – Исходные данные для кластерного анализа

Показатель	Предприятие									
	1	2	3	4	5	6	7	8	9	10
$x1$, %	23,4	17,5	9,7	18,2	6,6	8,0	9,8	19,1	15,2	6,8
$x2$, р./чел.	9,1	5,2	5,5	9,4	7,5	5,7	5,7	7,2	4,9	7,9

В качестве расстояния между объектами принять:

– обычное евклидово расстояние;

– взвешенное евклидово расстояние с весами 0,1 и 0,9;

– сравнить разбиения на два кластера по критерию минимума суммы внутриклассовых дисперсий.

Расстояние между кластерами определить по принципу «ближайшего соседа», не нормализуя данные.

Задача 2. Решить задачу 1, предварительно нормализовав исходные данные.

Задача 3. Решить задачу 1, измеряя расстояние между кластерами по принципу «дальнего соседа», не нормализуя предварительно исходные данные.

Задача 4. Решить задачу 1, измеряя расстояние между кластерами по «центрам тяжести» групп, не нормализуя предварительно исходные данные.

Задача 5. По агломеративному алгоритму провести классификацию $n = 10$ хозяйств, данные о деятельности которых характеризуются показателями объема реализованной продукции растениеводства $x1$ и животноводства $x2$ с 1 га пашни (р./га). Данные представлены в таблице 8.2.

Таблица 8.2 – Исходные данные для кластерного анализа

Показатель	Хозяйство									
	1	2	3	4	5	6	7	8	9	10
$x1$	2,49	1,51	1,17	1,67	2,73	2,78	1,19	2,15	1,93	1,21
$x2$	0,38	0,51	0,28	0,29	0,34	0,39	0,25	0,29	0,30	0,27

В качестве расстояния между объектами принять обычное евклидово расстояние, а расстояние между кластерами измерять по принципу:

– «ближайшего соседа»;

– «дальнего соседа»;

– сравнить разбиения на два кластера по критерию минимума суммы внутриклассовых дисперсий. Исходные данные не нормализовывать.

Задача 6. Решить задачу 5, нормализовав исходные данные.

Задача 7. Анализируются сведения об уровне развития промышленности в девяти странах (С1, С2, ..., С9). Показатели, характеризующие уровень развития этих стран, представлены в таблице 8.3.

Таблица 8.3 – Исходные данные для кластерного анализа

Страна	С1	С2	С3	С4	С5	С6	С7	С8	С9
Доля промышленной продукции в валовом национальном продукте (ВНП), %	68	74	23	35	67	71	43	54	32
Доля высокотехнологичной продукции в ВНП, %	17	42	12	27	49	20	24	18	25

Требуется выделить группы стран, имеющих сходные значения показателей.

При решении задачи с использованием *метода k-средних* выделить следующие группы:

1) страны с высокой долей промышленной и высокотехнологичной продукции в ВНП;

2) страны с низкой долей промышленной и высокотехнологичной продукции в ВНП;

3) страны со средними значениями обоих показателей.

Контрольное задание

Выполнить группировку объектов методами кластерного анализа на основе самостоятельно подобранных данных из [7, 8].

9 Лабораторная работа № 9. Классификация объектов с учителем с помощью дискриминантного анализа

Цель работы: научиться относить объекты, характеризующиеся набором признаков, к тому или иному классу при помощи дискриминантного анализа.

Задачи: научиться формировать классы объектов на основе обучающей выборки, определять принадлежность тому или иному классу нового объекта.

Дискриминантный анализ включает в себя статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками («классификация с учителем»).

Задача 1. В таблице 9.1 представлены группы передовых и отстающих предприятий. Характеризуется деятельность предприятий такими показателями как рентабельность и производительность труда. С помощью дискриминантного анализа требуется классифицировать три последних предприятия.

Таблица 9.1 – Исходные данные для дискриминантного анализа

Номер предприятия	Группа предприятий	Показатель	
		Рентабельность	Производительность труда
1	Передовые	23,4	9,1
2		19,1	6,6
3		17,5	5,2
4		17,2	10,1
5	Отстающие	5,4	4,3
6		6,6	5,5
7		8,0	5,7
8		9,7	5,5
9		9,1	6,6
10	Подлежат классификации	9,9	7,4
11		14,2	9,4
12		12,9	6,7

Контрольные вопросы

1 Для чего используется дискриминантный анализ?
2 Что такое дискриминирующая переменная?
3 Что такое обучающая выборка?

10 Лабораторная работа № 10. Классификация объектов с помощью логистической регрессии

Цель работы: изучить алгоритм классификации объектов с помощью логистической регрессии.

Задачи: научиться строить уравнение логистической регрессии, рассчитывать параметры качества уравнения, интерпретировать получаемые результаты.

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная y, принимающая лишь одно из двух значений – как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) – вещественных x_1, x_2, ..., x_n, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

Логистическая функция имеет следующий вид:

$$y = f(z) = \frac{1}{1+e^{-z}},$$

где z – уравнение регрессии, описывающее зависимость результирующего показателя от независимых переменных,

$$z = F(x_1, x_2, \ldots, x_n) \ .$$

Подбор коэффициентов регрессионного уравнения осуществляется на обучающей выборке.

Графической интерпретацией результатов классификации с помощью логистической регрессии служит ROC-кривая. А площадь под ней (AUC) является показателем качества классификационной модели (таблица 10.1).

Таблица 10.1 – Соответствие значения AUC и качества логистической модели

Интервал значений AUC	Качество модели
0,9…1,0	Отличное
0,8…0,9	Очень хорошее
0,7…0,8	Хорошее
0,6…0,7	Среднее
0,5…0,6	Неудовлетворительное

Для оценки качества логистической модели строится матрица сопряжённости, рассчитываются коэффициенты чувствительности и специфичности, которые варьируются в зависимости от порога отсечения.

Задание

Кредитным отделом банка накоплена информация о кредитополучателях и их платёжной дисциплине. Информация о клиентах содержит сведения о их возрасте, образовании, месте работы, зарплате, наличии недвижимого имущества и другие сведения, а также их кредитную историю. Одним из характеристик является наличие или отсутствие пени за просрочку платежей. Требуется построить логистическую классификационную модель, которая будет определять, будет ли новый клиент нарушать график платежей или нет. Сравнить несколько вариантов логит-модели с различными регрессионными уравнениями.

Контрольные вопросы

1 Для чего используется логистическая модель?
2 Какие показатели рассчитываются на основе матрицы сопряжённости?
3 Поясните алгоритм построения ROC-кривой.
4 Что показывает показатель AUC? В каких диапазонах он может изменяться?

Список литературы

1 **Мхитарян, В. С.** Анализ данных в MS Excel: учебное пособие / В. С. Мхитарян, В. Ф. Шишов, А. Ю. Козлов. – Москва: КУРС, 2019. – 368 с.

2 Эконометрика: учебник для магистров / И. И. Елисеева [и др.]; под ред. И. И. Елисеевой. – Москва: Юрайт, 2014. – 453 с.

3 **Кулешова, О. В.** Microsoft Excel 2016/2013. Расширенные возможности. Решение практических задач / О. В. Кулешова. – Москва: Специалист, 2016. – 100 с.

4 **Ниворожкина, Л. И.** Многомерные статистические методы в экономике: учебник / Л. И. Ниворожкина, С. В. Арженовский. – Москва: РИОР; ИНФРА-М, 2017. – 203 с.

5 **Дубров, А. М.** Многомерные статистические методы: учебник / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – Москва: Финансы и статистика, 2011. – 354 с.

6 Центр справки и обучения Office [Электронный ресурс]. – Режим доступа: https://support.office.com. – Дата доступа: 27.12.2019.

7 Открытые данные [Электронный ресурс]. – Режим доступа: http://opendata.by/. – Дата доступа: 27.12.2019.

8 Открытые данные [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/. – Дата доступа: 27.12.2019.