МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра биомедицинской инженерии

УТВЕРЖДАЮ
Проректор по учебной работе
_____ О.Г. Локтионова
«_11_» _03_ 2018 г.

МЕТОДЫ ОБРАБОТКИ МНОГОМЕРНЫХ СИГНАЛОВ И ДАННЫХ

Методические рекомендации по организации и выполнению самостоятельной работы для аспирантов направлений подготовки 09.06.01

Курск 2018

УДК 004.93:61

Составитель: С.А. Филист.

Рецензент

Доктор технических наук, профессор А.Ф. Рыбочкин

Методы обработки многомерных сигналов и данных: Методические рекомендации по организации и выполнению практической работы / Юго-Зап. гос. ун-т; сост.: С.А. Филист. - Курск, 2018. - 55 с.

Методические указания по структуре, содержанию и стилю изложения материала соответствуют требованиям, предъявляемым к учебным и методическим пособиям. Предназначены для аспирантов направлений подготовки 09.06.01 «Информатика и вычислительная техника (Системный анализ, управление и обработка информации (технические и медицинские системы))»

Текст печатается в авторской редакции

Подписано в печать 01.02.18. Формат 60x84 1/16.

Усл.печ.л. 3,3. Уч.-изд.л. 3,1 Тираж 100 экз. Заказ 1426 Бесплатно.
Юго-Западный государственный университет.

305040, г. Курск, ул. 50 лет Октября, 94.

Содержание

Самостоятельная работа 1
Характеристика и модели данных

Цель работы: представление статистических данных, построение вариационных рядов, вычисление средних величин и показателей вариации.

Исходные данные. Исследован размер заработной платы работников предприятия. Данные представлены в таблице (Приложение 1)

Порядок выполнения работы:

1) В соответствии с вариантом выбрать данные из таблицы исходных данных.

2) Упорядочить исходные данные (провести сортировку по возрастанию)

3) На основе *исходных* данных определить:

а) среднее значение показателя, моду и медиану

б) размах вариации, среднее линейное отклонение, дисперсию, стандартное отклонение, коэффициент вариации

4) На основе исходных данных построить *дискретный* вариационный ряд и определить:

а) среднее значение показателя, моду и медиану

б) размах вариации, среднее линейное отклонение, дисперсию, стандартное отклонение, коэффициент вариации

в) первый и третий квартили

г) построить диаграммы распределения работников по заработной плате.

5). На основе исходных данных построить *интервальный* вариационный ряд с равными интервалами. Число интервалов задано в каждом варианте. Определить:

а) среднее значение показателя, моду и медиану

б) размах вариации, среднее линейное отклонение, дисперсию, стандартное отклонение, коэффициент вариации

в) первый и третий квартили

г) построить диаграммы распределения работников по заработной плате.

6) Провести сравнительный анализ полученных результатов.

7) Оформить отчет.

Самостоятельная работа 2

Первичный анализ данных на компьютере в среде Microsoft Excel. Вычисление статистических характеристик показателей с использованием встроенных функций.

Исходные данные. Основные социально-экономические показатель субъектов ЦФО РФ представлены в таблице (Приложение 2)

Цель работы: построение и оценка качества группировки. Изучение взаимосвязи признаков методом аналитической группировки.

Порядок выполнения лабораторной работы:

1) По номеру варианта выбрать из таблицы Приложения 2 столбец, содержащий значения показателя.

2) Представить графически (столбиковая диаграмма) значения показателя у субъектов.

3) Пользуясь статистическими процедурами Excel, определить: Максимальное и минимальное значение признака (МАКС, МИН)

Среднее значение (СРЗНАЧ), медиану (МЕДИАНА), моду (МОДА)

Дисперсию (ДИСПР) и среднее квадратическое отклонение (СТАНДОТКЛОН)

4) Вычислить коэффициент вариации. Сделать выводы относительно однородности совокупности.

5) Определить удельный вес каждого субъекта в общем объеме признака в СФО

6) Представить графически (круговая диаграмма) структуру совокупности.

6) Провести сравнительный анализ полученных результатов.
7) Оформить отчет

Определить вариант лабораторной работы и выбрать данные из таблицы ПРИЛОЖЕНИЯ 3

Вариант	Фактор (номер показа-теля)	Результат-ВРП	Вариант	Фактор (номер показа-теля)	Результат-ВРП
1.	2	1	6.	7	1
2.	3	1	7.	8	1
3.	4	1	8.	9	1
4.	5	1	9.	10	1
5.	6	1	10.	11	1

Исходные данные

Сформировать исходную таблицу, содержащую названия регионов и указанные в варианте показатели социально-экономического развития регионов ЦФО (приложение 2).

Регион	Показатель (фактор)	ВРП (результат)
	x	*y*
…	…	…

Порядок выполнения лабораторной работы:

1) Провести сортировку по значению фактора.

2) Провести по всей совокупности для каждого признака расчет среднего значения, дисперсии, стандартного отклонения, коэффициента вариации.

3) Построить *точечную* диаграмму зависимости результата от фактора.

4) Провести группировку регионов по *значению фактора*, выделив 3 группы: «Малые», «Средние», «Крупные». Границы группировочного показателя задать самостоятельно и уметь обосновать их.

Для каждой группы определить и занести в таблицу1:
-частоту группы,
-групповые средние значения показателей *x* и *y*
-групповые дисперсии показателя *x* и *y*,
-групповые коэффициенты вариации показателей *x* и *y* .

Таблица 1 - Статистические характеристики группировки

Группа	Интервалы признака–фактора	Частота группы	*Признак – фактор x*			*Признак – результат y*		
		fj	*Среднее*	*Дисперсия*	*Коэффициент вариации*	*Среднее*	*Дисперсия*	*Коэффициент вариации*
Малые								
Средние								
Крупные								

5) Дать оценку качества построенной группировки по признаку-фактору. При расчете коэффициента детерминации R^2 рассчитать межгрупповую дисперсию.

6) Провести анализ наличия связи, направления связи между *x* и *y*

7) По величинам b_{yx}..определить линейность (нелинейность) связи между *x* и *y*

8) Рассчитать по формуле межгрупповую дисперсию по показателю – *фактору*

9) Дать оценку силы связи на основе расчета коэффициента детерминации R^2. При расчете коэффициента детерминации R^2 рассчитать:

- среднюю групповых дисперсий группировки по признаку-*результату*;

- используя правило сложения дисперсий, вычислить межгрупповую дисперсию для группировки по признаку-результату

10) вычислить эмпирическое корреляционное отношение по формуле (7).

11) Провести анализ полученных результатов.

12) Оформить отчет.

«Обработка и оценка результатов исследования»

Цель работы: научиться использовать возможности MS Excel для проведения корреляционного и регрессионного анализа исследовательских данных, планирования и обработки результатов факторного эксперимента.

Учебные вопросы:

1. Возможности прикладного программного обеспечения на этапах обработки и оценки результатов исследования.

Изучив данную тему, студент должен: знать:

-назначение существующих современных средств компьютеризации научных исследований, их функциональные возможности и особенности применения;

уметь:

- производить обработку и оценку результатов исследования.

1.1. Краткое изложение основных теоретических и методических аспектов работы

Параметрический корреляционный анализ

Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между выборками (наборами числовых данных каких-либо величин). Обычно связь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей различают *корреляцию* и *регрессию*.

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами *X* и *Y*. В качестве меры такой связи используется коэффициент корреляции. Коэффициент корреляции оценивается по выборке объема *n* связанных пар наблюдений *(xi, yi)* из совместной генеральной совокупности *X* и *Y*. Существует несколько типов коэффициентов корреляции, применение которых зависит от измерения (способа шкалирования) величин *X* и *Y*.

Для оценки степени взаимосвязи величин *X* и *Y*, измеренных в количественных шкалах, используется коэффициент линейной корреляции (коэффициент Пирсона), предполагающий, что выборки *X* и *Y* распределены по нормальному закону.

Линейный коэффициент корреляции – параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле:

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - y)^2}} \qquad (1)$$

где xi – значения, принимаемые в выборке *X*,
yi – значения, принимаемые в выборке *Y*;
$\bar{x}\bar{y}$ – средняя по *X*, – средняя по *Y*.

Коэффициент корреляции изменяется от –1 до 1. Когда при расчете получается величина большая +1 или меньшая –1 – следовательно, произошла ошибка в вычислениях. При значении 0 линейной зависимости между двумя выборками нет.

Знак коэффициента корреляции очень важен для интерпретации полученной связи (таблица 1). Если знак коэффициента линейной корреляции «+», то связь между коррелирующими признаками такова, что большей величине одного признака (переменной) соответствует большая величина другого признака (другой переменной). Иными словами, если один показатель (переменная) увеличивается, то соответственно увеличивается и другой показатель (переменная). Такая зависимость носит название прямо пропорциональной зависимости.

Таблица 1. Теснота связи и величина коэффициента корреляции.

Коэффициент корреляции *rxy*	**Теснота связи**
±(0,91 … 1,00)	Очень сильная
±(0,81… 0,90)	Весьма сильная
± (0,25… 0,44)	Сильная
± (0,45… 0,64)	Слабая
До ± 0,25	Очень слабая

знак «+» – прямая зависимость, «–» – обратная зависимость

Большей величине одного признака соответствует меньшая величина другого. Иначе говоря, при наличии знака минус, увеличению одной переменной (признака, значения) соответствует уменьшение другой переменной. Такая зависимость носит название обратно пропорциональной зависимости.

t-статистика Стьюдента

Для того чтобы оценить наличие связи между двумя переменными, также можно использовать *t-статистику Стьюдента*, которая оценивает отношение величины линейного коэффициента корреляции к среднему квадратическому отклонению и рассчитывается по формуле

$$t_{расч} = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}, \qquad (2)$$

Полученную величину *tрасч* сравнивают с табличным значением *tтабл* критерия Стьюдента с $n - 2$ степенями свободы. Если *tрасч* > *tтабл*, то практически невероятно, что найденное значение обусловлено только случайными совпадениями величин *X* и *Y* в выборке из генеральной совокупности, т.е.

существует зависимость между *X* и *Y*. И наоборот, если *tрасч* < *tтабл*, то величины *X* и *Y* независимы.

Исследование связей между двумя переменными в Excel

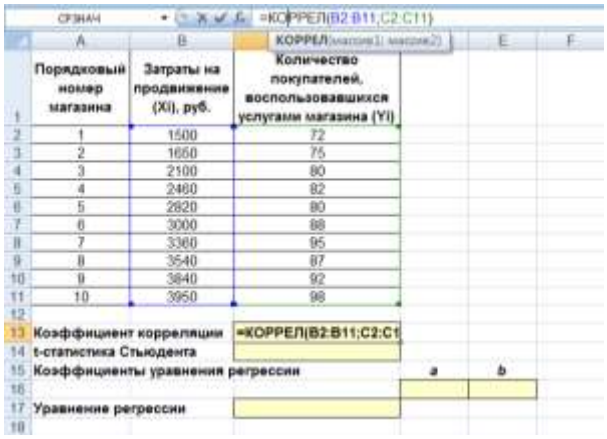
Условие задачи: По 10 интернет-магазинам были определены затраты на рекламную раскрутку сайтов и количество покупателей, воспользовавшихся после ее проведения услугами каждого магазина. Определить коэффициент корреляции между исследуемыми признаками.

Ход выполнения:

1. Открываем новую книгу MS Excel и создаем таблицу согласно рисунку 2.
2. Рассчитываем в ячейке C12 коэффициент корреляции, используя функцию КОРРЕЛ из категории Статистические.

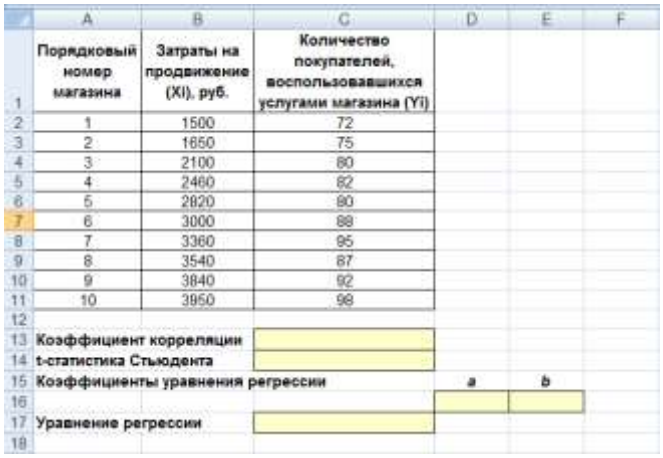
Синтаксис функции: КОРРЕЛ (<массив 1>;<массив 2>), где <массив 1> – ссылка на диапазон ячеек первой выборки (X); <массив 2> – ссылка на диапазон ячеек второй выборки (Y).

В нашей задаче формула будет иметь вид: =КОРРЕЛ(B2:B11;C2:C11) – см. рисунок 3.



	A	B	C	D	E	F
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)			
2	1	1500	72			
3	2	1650	75			
4	3	2100	80			
5	4	2460	82			
6	5	2820	90			
7	6	3000	88			
8	7	3360	95			
9	8	3540	87			
10	9	3840	92			
11	10	3950	98			
12						
13	Коэффициент корреляции		=КОРРЕЛ(B2:B11;C2:C1			
14	t-статистика Стьюдента					
15	Коэффициенты уравнения регрессии			a	b	
16						
17	Уравнение регрессии					
18						

Рисунок 2 - Исходные данные для исследования связей между двумя переменными



	A	B	C	D	E	F
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)			
2	1	1500	72			
3	2	1650	75			
4	3	2100	80			
5	4	2460	82			
6	5	2820	80			
7	6	3000	88			
8	7	3360	95			
9	8	3540	87			
10	9	3840	92			
11	10	3950	98			
12						
13	Коэффициент корреляции					
14	t-статистика Стьюдента					
15	Коэффициенты уравнения регрессии			a	b	
16						
17	Уравнение регрессии					
18						

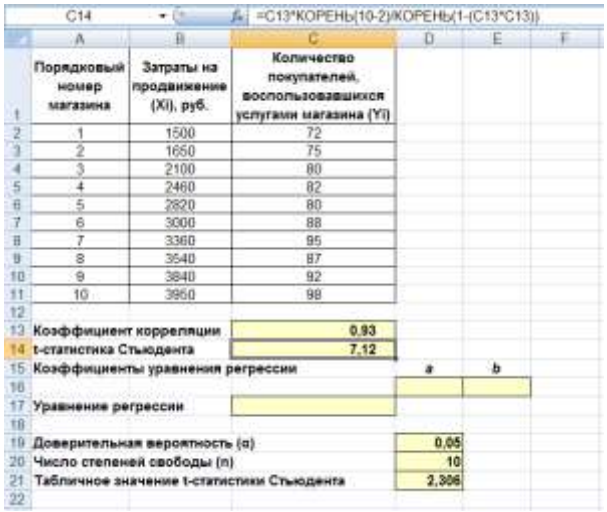
Рисунок 3 - Вычисление коэффициента корреляции

3. Сделаем вывод о тесноте связи между затратами на рекламную раскрутку сайтов и количество покупателей.

После ввода формулы получаем в ячейке C13 значение коэффициента корреляции равное 0,93. По таблице 2 делаем вывод, что связь между переменными очень сильная, т.е. имеет место линейная зависимость (прямая пропорциональность).

4. Оценим значимость коэффициента корреляции. С этой целью рассмотрим две гипотезы. Основную H0: rxy=0 и альтернативную H1: rxy≠0. Для проверки гипотезы H0 рассчитаем в ячейке C14 t-статистику Стьюдента по формуле, указанной в 3.1.2. В нашем случае число степеней свободы $\nu = n - 2 = 10 - 2 = 8$ и формула будет следующей: =C13*КОРЕНЬ(10-2)/КОРЕНЬ(1-(C13*C13)).

После ввода формулы получаем в ячейке C13 t-статистику Стьюдента (*tрасч*) равную 7,12 (рисунок 4).



C14 =C13*КОРЕНЬ(10-2)/КОРЕНЬ(1-(C13*C13))

	A	B	C	D	E	F
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)			
2	1	1500	72			
3	2	1650	75			
4	3	2100	80			
5	4	2460	82			
6	5	2820	80			
7	6	3000	88			
8	7	3360	95			
9	8	3540	87			
10	9	3840	92			
11	10	3950	98			
12						
13	Коэффициент корреляции		0,93			
14	t-статистика Стьюдента		7,12			
15	Коэффициенты уравнения регрессии			a	b	
16						
17	Уравнение регрессии					
18						
19	Доверительная вероятность (α)			0,05		
20	Число степеней свободы (n)			10		
21	Табличное значение t-статистики Стьюдента			2,306		
22						

Рисунок 4 - Вычисление t-статистики Стьюдента (*tрасч*)

5. Сравним полученное значение с критическим значением $t_{\nu,\alpha}$, табличное распределения Стьюдента (при $\nu = 8$ и доверительной вероятности $\alpha = 0{,}05$, $t_{\nu,\alpha,табл} = 2{,}306$). $t_{\nu,\alpha,табл}$ можно найти либо в специальной таблице (приложение 1), либо воспользовавшись встроенной статистической функцией СТЬЮДРАСПОБР(вероятность; степени_ свободы). В нашем случае это будет формула: =СТЬЮДРАСПОБР(D19;D20-2).

6. Сделаем вывод о наличии связи между исследуемыми величинами – так как $t_{расч} > t_{\nu,\alpha,табл}$ (7,12 > 2,306), то между переменными существует зависимость и найденный коэффициент корреляции значим.

Регрессионный анализ

Цель регрессионного анализа – определить количественные связи между зависимыми случайными величинами. Одна из этих величин полагается зависимой и называется откликом, другие – независимые, называются факторами. Для установления степени зависимости между откликом и факторами используются вычисляемые величины ковариации и коэффициент корреляции. Если коэффициент корреляции по абсолютной величине близок к единице, то для построения зависимости используется линейная модель. Для других случаев используются более сложные нелинейные модели (например, полиномиальные и экспоненциальные). В данной работе изучим линейную модель.

Уравнение линейной регрессии имеет вид:

$Y = a_1X_1 + a_2X_2 + \ldots + a_kX_k$,

где *a1, a2, …, ak* – параметры, подлежащие определению методом наименьших квадратов (МНК). Обычно находят первые два параметра, которые принято обозначать *a* и *b*. В этом случае

уравнение линейной регрессии имеет вид $Y = a \cdot X + b$.

Коэффициенты a и b вычисляются следующим образом (формулы 3 – 4):

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2}, \qquad (3)$$

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x^2{}_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2}, \qquad (4)$$

где i – номер измерения, xi и yi – значения переменных при i-том измерении, n – число измерений при моделировании системы.

В среде MS Excel для нахождения модели регрессии (т.е., фактически коэффициентов a и b) можно использовать несколько способов:

- использовать встроенную функцию ЛИНЕЙН;
- графический способ – построение линии тренда на диаграмме с показом уравнения регрессии;
- инструмент Регрессия из Пакета анализа;
- использовать встроенную функцию СУММКВРАЗН и инструмент Поиск решения;
- использовать встроенные функции НАКЛОН (вычисляет коэффициент a) и ОТРЕЗОК (вычисляет коэффициент b).

Построение регрессионной модели средствами Excel

Рассмотрим на примере первые три из перечисленных способов нахождения модели регрессии.

1-й способ. Функция ЛИНЕЙН.

В первом способе для получения коэффициентов a и b линейного уравнения регрессии $Y = a \cdot X + b$, описывающего зависимость количества привлеченных покупателей от затрат на рекламную раскрутку сайтов, воспользуемся статистической функцией ЛИНЕЙН. Для этого выделите две ячейки D16:E16 и выполните вставку функции ЛИНЕЙН с аргументами согласно рисунку 5.

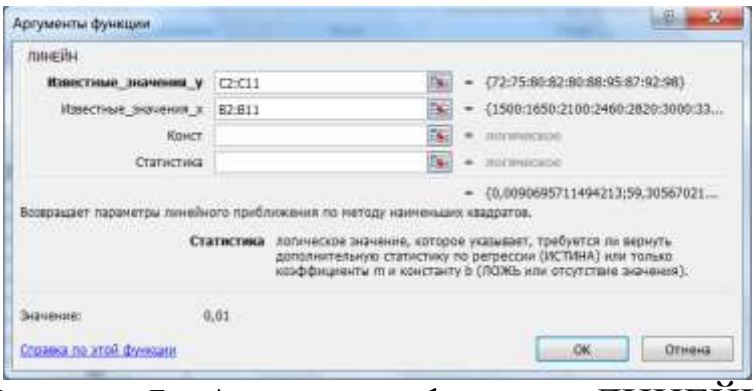


Рисунок 5 - Аргументы функции ЛИНЕЙН

Здесь *«Известные_значения_y»* – диапазон значений *«Количество покупателей»*, *«Известные_значения_x»* – диапазон значений *«Затраты на продвижение»*. Нажмите комбинацию клавиш SHIFT+CTRL+ENTER.

Получаем следующие значения коэффициентов регрессии – *a = 0,01* (ячейка D16), *b = 59,32* (ячейка E16). В ячейку D17 введем уравнение *Y = 0,01□X + 59,31*, чтобы продемонстрировать уравнение регрессии:

15	**Коэффициенты уравнения регрессии**		***a***	***b***
16			**0,01**	**59,31**
17	**Уравнение регрессии**	***Y=0,01·X+59,31***		

2-й способ (графический). Построение линии тренда

1. Для получения уравнения регрессии построим корреляционное поле переменных *X* (затраты на продвижение) и *Y* (количество покупателей).

2. Выделим диапазон ячеек B2:C11, запустим мастер диаграмм и выберем тип диаграммы – Точечная (в Excel 2007 выберем на панели инструментов «Вставка» кнопку «Точечная» и выберем подтип «Точечная с маркерами», после этого диаграмма будет создана и помещена на текущий лист, после чего ее можно будет дооформить). Задаем для диаграммы имя – *«Корреляционное поле»*, название оси *X* – *«Затраты на продвижение, руб.»*, оси *Y* – *«Количество покупателей»* (в Excel 2007 данные действия выполняются на вкладке «Макет» после выделения диаграммы – команды «Название диаграммы» и «Названия осей»). На последнем шаге мастера указываем место расположения – текущий лист.

3. Добавим линию тренда на точечный график (рис. 6). Для этого необходимо выделить диаграмму и выполнить команду меню «Диаграмма/Добавить линию тренда» (в Excel 2007 на вкладке «Макет» выберите команду «Анализ» и далее «Линия тренда» и «Линейное приближение»), либо выполнить данную команду из контекстного меню «Добавить линию тренда…», щелкнув по любой точке графика правой кнопкой мыши. Линия тренда – графическое представление направления изменения ряда данных.

4. Выбираем тип тренда «Линейный», который используется для аппроксимации данных по методу наименьших квадратов в соответствии с уравнением: $Y = a \square X + b$, где a – угол наклона (в радианах) и b – координата пересечения оси абсцисс (оси Y).

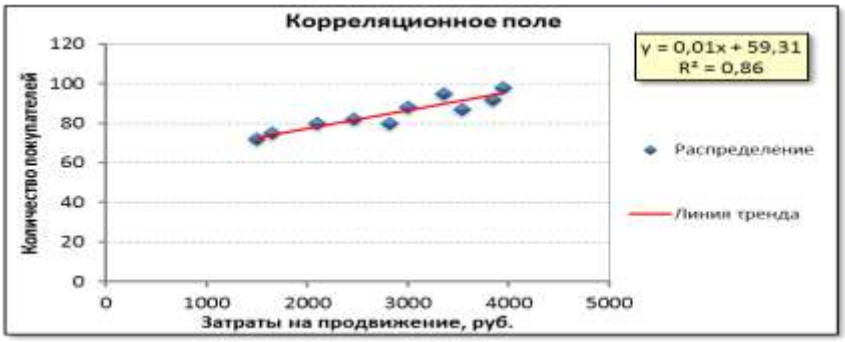


Рисунок 6 - Диаграмма с линией и уравнением тренда

5. На вкладке Параметры устанавливаем флажки «Показать уравнение на диаграмме» и «Поместить на диаграмму величину достоверности аппроксимации R^2». Щелкаем по кнопке ОК. Далее можно отформатировать эти уравнения, выделив их и в контекстном меню выбрав «Формат подписи линии тренда». R^2 – это число от 0 до 1, которое отражает близость линии тренда к фактическим данным. Линия тренда наиболее соответствует действительности, когда значение близко к 1.

6. Сравниваем уравнение регрессии, полученное графическим методом, с уравнением, рассчитанным с помощью функции ЛИНЕЙН. Как видим, эти уравнения одинаковые.

3-й способ. Инструмент анализа Регрессия.

1. Прежде чем мы начнем использовать этот инструмент, нужно убедится, что был активизирован Пакет анализа (меню «Сервис» есть команда «Анализ данных»). Если нет, то выполните команду «Сервис/Надстройки». В диалоговом окне «Надстройки» установите флажок «Пакет анализа» и щелкните по кнопке ОК (в Excel 2007 этот инструмент находится на вкладке «Данные» – «Анализ данных»).

2. Далее выполните команду «Сервис/Анализ данных». Выберите инструмент анализа «Регрессия» из списка «Инструменты анализа». Щелкните по кнопке ОК.

3. На экране появится диалоговое окно «Регрессия» (рисунок 7):

- в текстовом поле «Входной интервал Y» введите диапазон со значениями зависимой переменной $C$2:$C$211.
- в текстовом поле «Входной интервал X» введите диапазон со значениями независимых переменных $B$2:$B$11.
- Убедитесь, что в поле Уровень надежности введено 95% и переключатель «Параметры вывода» установлен в положении «Новый рабочий лист».
- Щелкните по кнопке ОК.

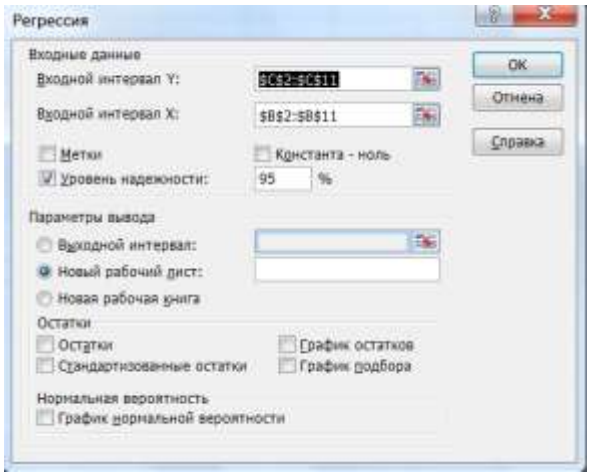
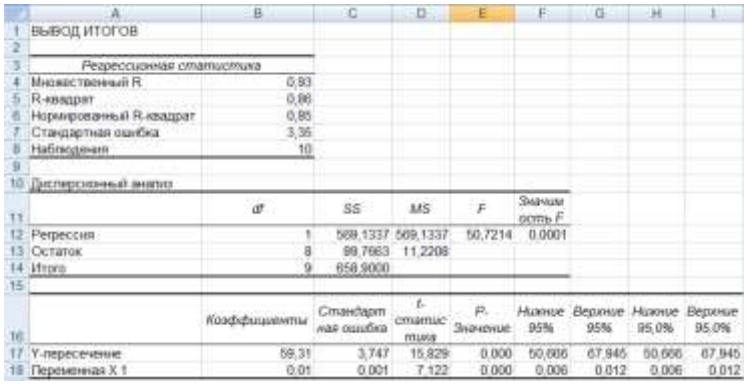


Рисунок 7 - Диалоговое окно инструмента анализа «Регрессия»

4. В результате на новом листе будет отображены результаты использования инструмента «Регрессия» (рисунок 8).



	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,93							
5	R-квадрат	0,86							
6	Нормированный R-квадрат	0,85							
7	Стандартная ошибка	3,35							
8	Наблюдения	10							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	569,1337	569,1337	50,7214	0,0001			
13	Остаток	8	89,7663	11,2208					
14	Итого	9	658,9000						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	59,31	3,747	15,829	0,000	50,666	67,945	50,666	67,945
18	Переменная X 1	0,01	0,001	7,122	0,000	0,006	0,012	0,006	0,012

Рисунок 8 - Вывод итогов инструмента «Регрессия»

5. Среди полученных результатов после применения инструмента Регрессия есть столбец «Коэффициенты», содержащий значение *b* в строке «Y-пересечение», *a* – в строке «Переменная X1».

6. Сравним полученные результаты с ранее рассчитанными коэффициентами *a* и *b* – результаты полностью совпадают.

7. Следует обратить также внимание на следующие показатели:

а) Столбец «*df*» – число степеней свободы (используется при проверке адекватности модели по статистическим таблицам):

- в строке «Регрессия» находится *k1* – количество коэффициентов уравнения, не считая свободного члена *b*;

- в строке «Остаток» находится $k2 = n - k1 - 1$, где *n* – количество исходных данных. б) Столбец «*SS*» (сумма квадратов):

- в строке Регрессия: $SS_{resid} = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$, где $-Y_i$ - модельные значения Y, полученные путем подстановки значений X в построенную модель;– среднее значение Y;

$SS_{resid} = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$ - в строке Остаток: .

в) Столбец «*MS*» – вспомогательные величины: в строке Регрессия: в строке Остаток: $S_r^2 = SS_{reg} / k_1$ $S_b^2 = SS_{reg} / k_2$.

$F = S_r^2 / S_b^2$

г) Столбец «*F*» – критерий Фишера. Используется для проверки

адекватности модели.

д) Столбец «*Значимость F*» – оценка адекватности построенной модели. Находится по значениям F, и с помощью функции FРАСП. Если значимость F меньше 0,05, то модель может считаться адекватной с вероятностью 0,95.

е) «*Стандартная ошибка*», «*t-статистика*» – это вспомогательные величины, используемые для проверки значимости коэффициентов модели.

ж) «*P-Значение*» – оценка значимости коэффициентов модели. Если «*P-Значение*» меньше 0,05, то с вероятностью 0,95 можно считать, что соответствующий коэффициент модели значим (т.е. его нельзя считать равным нулю и Y значимо зависит от соответствующего X).

и) Нижние и верхние 95% – доверительные интервалы для коэффициентов модели.

Прогнозирование данных

Кроме нахождения уравнения регрессии, часто необходимо на основании этого уравнения предсказать теоретические значения *Y* при известных значениях *X*.

Это можно сделать тремя способами (рисунок 9).

	A	B	C	D	E
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)		
2			Способ 1	Способ 2	Способ 3
3	1	1500	72	72	72
4	2	1650	75	75	75
5	3	2100	80	80	80
6	4	2460	82	82	82
7	5	2820	80	80	80
8	6	3000	88	88	88
9	7	3360	95	95	95
10	8	3540	87	87	87
11	9	3840	92	92	92
12	10	3950	98	98	98
13	**11**	**5000**			
14	**12**	**10000**			
15	**13**	**50000**			
16					
17	**Коэффициенты уравнения регрессии:**				
18	*a*	*b*			
19	0,01	59,31			

Рисунок 9 - Исходные данные для прогнозирования

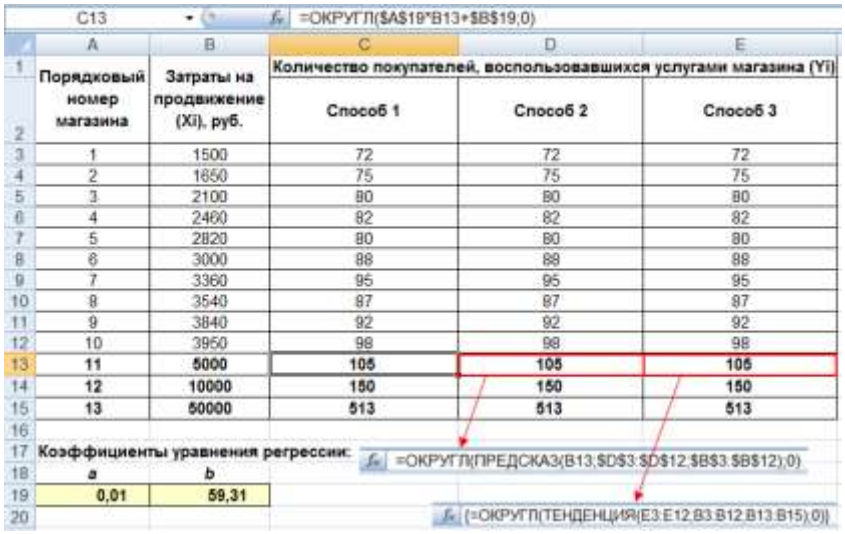
1. *Способ 1*. Создать в Excel обычную формулу, основанную на уравнении регрессии $Y = a \cdot X + b$, типа

C13=$A$19*B13+$B$19, где C13 – адрес ячейки с прогнозным значением функции *Y*, B13 – адрес ячейки со значением переменной *X*, для которого мы хотим спрогнозировать значение *Y*, $A$19 – абсолютный адрес ячейки со значением коэффициента *a*, $B$19 – абсолютный адрес ячейки со значением коэффициента *b*. В нашем случае нужно округлить до целого с помощью функции ОКРУГЛ($A$19*B13+$B$19;0). После чего скопируем формулу в ячейки C14 и C15.

2. *Способ 2.* Также можно вычислить теоретическое значение *Y* при *X* из ячейки B13 с помощью функции ПРЕДСКАЗ. Ее синтаксис – ПРЕДСКАЗ(Xi;<массив Y>;<массив X>). Аргумент Xi – это точка данных из массива X, для которой предсказывается теоретическое значение Yi. Теоретическое значение в ячейке D13 вычислим по формуле = ПРЕДСКАЗ(B13;$D$3:$D$12;$B$3:$B$12). После чего скопируем формулу в ячейки D14 и D15.

3. *Способ 3.* Еще один способ прогнозирования – вычислить значения уравнения линейной регрессии Y для целого диапазона значений независимой переменной X с помощью функции ТЕНДЕНЦИЯ. Ее синтаксис – ТЕНДЕНЦИЯ(<массив Y>;<массив X>;<новые значения X>;[<константа>]). Аргумент <новые значения X > – это массив значений X, для которых функция ТЕНДЕНЦИЯ возвращает соответствующие значения Y. Новые значения зависимой переменной вычислим в ячейках E13:B15 по формуле =ТЕНДЕНЦИЯ(E3:E12;B3:B12;B13:B15). **Важно** оформить эту функцию в ячейках E13:E15 как массив, для чего после ввода формулы в ячейку B12 нажать клавишу ENTER, выделить ячейки E13:E15, нажать клавишу F2, после этого нажать комбинацию клавиш SHIFT+CTRL+ENTER.

4. *Способ 4.* Сравним полученные результаты для всех трех способов (рисунок 10). Видим, что все три способа дают одинаковые результаты, что не удивительно, так как во всех случаях используются линейная регрессия.



	A	B	C	D	E
1	Порядковый номер магазина	Затраты на продвижение (Xi), руб.	Количество покупателей, воспользовавшихся услугами магазина (Yi)		
2			Способ 1	Способ 2	Способ 3
3	1	1500	72	72	72
4	2	1650	75	75	75
5	3	2100	80	80	80
6	4	2460	82	82	82
7	5	2820	80	80	80
8	6	3000	88	88	88
9	7	3360	95	95	95
10	8	3540	87	87	87
11	9	3840	92	92	92
12	10	3950	98	98	98
13	11	5000	105	105	105
14	12	10000	150	150	150
15	13	50000	513	513	513
16					
17	Коэффициенты уравнения регрессии:				
18	a	b			
19	0,01	59,31			
20					

Рисунок 10 - Результаты прогнозирования тремя способами

1.2.1. Введение в теорию факторного планирования эксперимента

Если необходимо изучить влияние, например, количества углерода X на прочность стали Y проводят однофакторный эксперимент. И чем больше различных значений примет X, тем более полно мы узнаем изучаемую зависимость Y(X).

Допустим, что исследуем влияние на прочность стали Y количества углерода X1 и количества хрома X2. Последовательно проведя 2 серии однофакторных экспериментов получим всего лишь 2 линии на двумерном экспериментальном поле – основная область возможных сочетаний факторов останется неисследованной (рисунок 19).

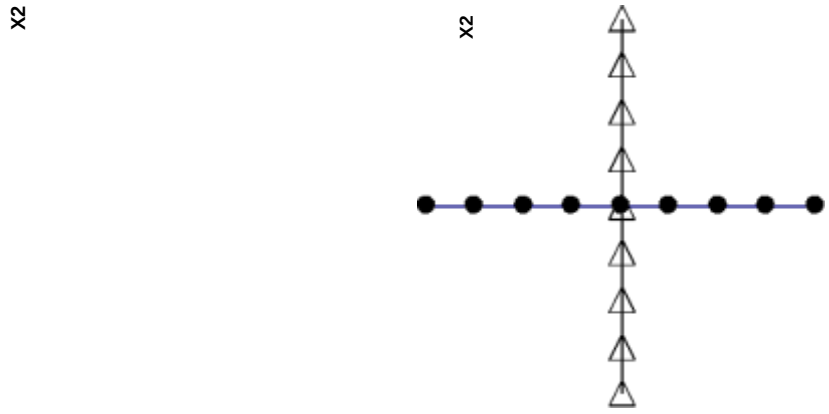


Рисунок 19 - Схема эксперимента «крест»

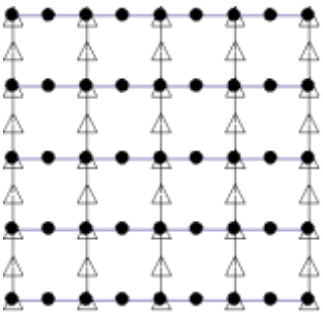


Рисунок 20 - Схема эксперимента «решетка»

Попытка «заштриховать» всё поле эксперимента экспериментальными линиями (рисунок 20) приведет к недопустимо высоким затратам по времени и по средствам. На рисунках 19 и 20 треугольниками обозначены серии с варьированием X_2 при постоянном X_1, точками – серии с варьированием X_1 при постоянном X_2. Решение – провести отдельные эксперименты в точках, расположенных на границах, в углах и в центре исследуемой области. Это пример факторного планирования эксперимента (рисунок 21).

Факторное планирование эксперимента имеет цель: за минимальное количество экспериментов описать исследуемую область с достаточной для экспериментатора точностью.

Факторный эксперимент – мощное средство эмпирического изучения процессов, обеспечивающее точное математическое описание отклика системы при минимальном количестве экспериментов.

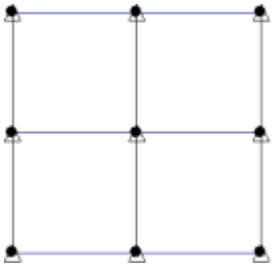


Рисунок 21 - Схема полного факторного эксперимента (ПФЭ) 3^2

Не приводя строгих определений терминов, связанных с факторным планированием, опишем их упрощенно.

Фактор, X – величина, которую экспериментатор меняет (варьирует). Отклик Y – величина, которую экспериментатор измеряет.

Факторное пространство – служит для мысленного расположения в нем экспериментальных точек. Количество измерений равно количеству факторов.

План полного факторного эксперимента ПФЭ обозначается m^k где m – число уровней

варьирования факторов, k – число факторов. Например, если 3 фактора варьируются на 2-х уровнях, то план ПФЭ обозначится 2^3 и будет состоять из 8 опытов на различных сочетаниях факторов. Очевидно, что план 3^2 состоит из 9 опытов. Планы 2^k называют планами первого порядка, планы 3^k – планы второго порядка. Планы больших порядков используют редко – для повышения точности выгоднее повторить эксперимент, сузив диапазоны варьирования.

Пример

Гипотеза: Чтобы корова меньше ела и давала больше молока – ее надо меньше кормить и чаще доить.

1. Постановка задачи.

Корова (рисунок 22) представляет собой систему. Система имеет на входе контролируемые воздействия (варьируемые факторы) X1 и X2 и неконтролируемые воздействия (случайные факторы), например X3, X4, … . Случайные факторы не учитываем – полагаем систему детерминированной. Из выходных характеристик системы, Y, Y1, Y2, … в соответствии с целями исследования для контроля выбираем Y.

Рисунок 22 - Исследуемая система

В рамках принятой модели (рисунок 23) исследуем зависимость количества молока в сутки Y от количества корма X1 и числа доений X2.

Рисунок 23 - Принятая модель исследуемой системы

2. Планирование и обработка результатов эксперимента

План эксперимента 2^2.

Технически возможные пределы изменения факторов. Количество корма от 0 до 100 кг. Количество доений от 1 до 10.

Пределы варьирования факторов не должны превышать технически возможных и выбираются на усмотрение экспериментатора. С учетом гуманного отношения к животным принимаем пределы варьирования: X1 = 10…70 кг, X2 = 2…5 шт. (таблица 10).

В планах первого порядка два уровня варьирования факторов, верхний, обозначаемый «+» или «1» и нижний, обозначаемый «–» или «–1». При этом от натуральных значений факторов (X) переходят к кодированным (x) и оформляют в виде таблицы.

Следующая таблица (таблица 11) – матрица эксперимента – состоит из уровней варьирования факторов, взаимодействий и отклика.

Столбцы взаимодействий получаются перемножением соответствующих кодированных значений факторов.

Таблица 10. Уровни и интервалы варьирования факторов

Факторы	**д. изм.**	**Е Кодовое обозначение**	**Интервал варьирования**	**Натуральные уровни соответствующие кодированным**	
				1	–1
X1 – количество корма	г	к x1	60	70	10
X2 – число доений	т.	ш x2	3	5	2

Таблица 11. Матрица эксперимента

№ опыта	**x1**	**x2**	**x1x2**	**Y**
1	1	1	1	12
2	–1	1	–1	2
3	1	–1	–1	17
4	–1	–1	1	6

Для описания результатов планов первого порядка используют полиномы первого порядка, в данном случае:

$$Y=a0+a1x1+a2x2+a12x1x2 . \quad (5)$$

Коэффициенты при кодированных факторах дают информацию о влиянии факторов или из сочетаний на отклик.

В данном случае применив надстройку Excel «Поиск решения» получаем:

a0	a1	a2	a12
9,24999 7	5,24999 9	–2,25	–0,25

Это говорит о том, что повышение количества корма сильно влияет на количество молока, а вот повышения частоты доений – немного снижает количество молока, совместное влияние факторов не выражено.

Таким образом – чтобы было больше молока, корову надо больше кормить и реже доить.

3. Анализ полученных результатов

Вывод несколько противоречит сложившимся представлениям – если корову совсем не доить то, скорее всего, молока не будет, да и бесконечно увеличивать кормление тоже нецелесообразно.

Предполагаем, что существует оптимальное количество корма и числа доений, соответствующее максимуму молока. Проведем уточняющий эксперимент 3^2 сузив диапазоны варьирования и перейдя в предполагаемую оптимальную область. Поместим центр плана в точку с x1=1(X1 =70кг) и x2 = –1 (X2 = 2 шт.).

Таблица 12. Уровни и интервалы варьирования факторов

Факторы	**д. изм.**	**Кодовое обозначение**	**Интервал варьирования**	**Натуральные уровни соответствующие кодированным**		
				1		–1
X1 – количество корма	г	x1	20	90	0	50
X2 – число доений	т.	x2	1	3		1

Кодированные значения факторов *x* связаны с натуральными *X* через диапазон варьирования *e* и натуральное значение фактора в центре плана *X0*:

$$x = \frac{XX_0}{e} \qquad (6)$$

В факторном планировании часто при переходе от одного плана к другому стараются использовать данные предыдущего плана – в данном случае опыт 5 плана второго порядка можно не проводить, т.к. он соответствует опыту 3 плана второго порядка.

Для описания результатов плана второго порядка применяют полиномы второй степени, в

данном случае:

$$Y=a0+a1x1+a2x2+a12x1x2+a11x1^2+a22x2^2 \qquad (7)$$

Таблица 13. Матрица эксперимента

№ опыта	**x1**	**x2**	**x1x2**	**$x1^2$**	**$x2^2$**	**Y**
1	1	1	1	1	1	11
2	0	1	0	0	1	13
3	–1	1	–1	1	1	14
4	1	0	0	1	0	15
5	0	0	0	0	0	17
6	–1	0	0	1	0	18
7	1	–1	–1	1	1	9
8	0	–1	0	0	1	11
9	–1	–1	1	1	1	12

Построим в Excel таблицу на основе таблицы 13. Справа от столбца Y с экспериментальными данными располагаем столбец Yp где с помощью формул строим выражение (7) положив в качестве начальных значений всех коэффициентов ноль (для этого нужно для каждого коэффициента выбрать ячейку в Excel и записать в нее «0» – рисунок 24).

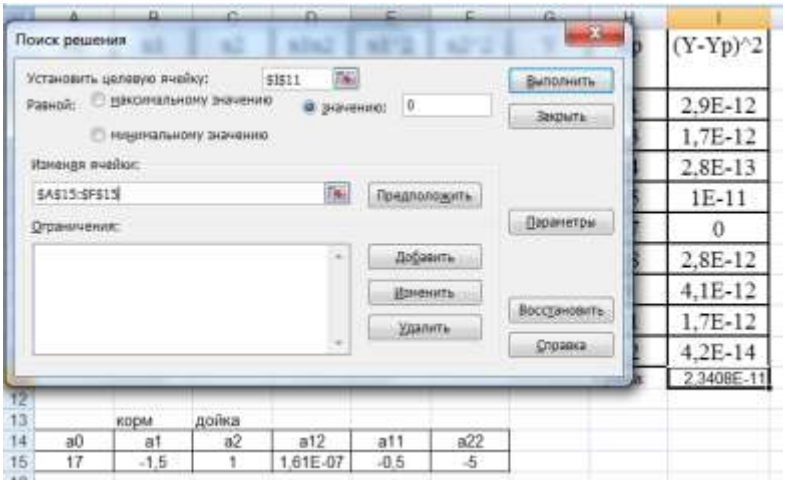


Рисунок 24 - Подготовка таблицы эксперимента для применения надстройки «Поиск решения»

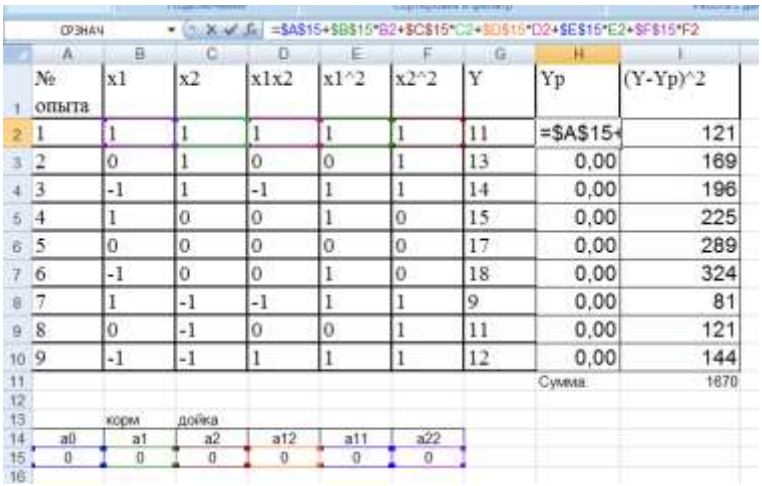


Рисунок 25 - Применение надстройки «Поиск решения»

Применив надстройку «Поиск решения», положив значение целевой ячейки «0» (см. рисунок 25), получаем:

a0	a1	a2	a12	a11	a22
17	–1,5	1	1,61E-07	–0,5	–5

Разные знаки при квадратичных и линейных коэффициентах указывают, что возможно, оптимум лежит внутри исследованной области.

Найти искомый оптимум можно с использованием надстройки «Поиск решения» положив в качестве целевой функции (7) с найденными коэффициентами и изменяя ячейки x1 и x2.

Результат:

x1	x2
–1,49999	0,1

Перейдя от кодированных значений к натуральным по (6) получаем, что максимальный суточный надой 18,2 л возможен при 2-х разовом доении и кормлении в объеме 40 кг в день. Оптимум по сену лежит за пределами диапазона варьирования и, строго говоря, нуждается в дополнительной экспериментальной проверке.

Вопросы статистической обработки при планировании и обработке результатов факторного эксперимента в данном примере не рассмотрены – для лучшего понимания основных принципов факторного планирования.

Требуемый результат исследования – оптимальное сочетание факторов – достигнут за 12 опытов.

Контрольные вопросы

1. В чем цель корреляционного анализа?
2. Что такое коэффициент корреляции?
3. Для чего используется t-статистика Стьюдента?
4. Какими способами можно определить коэффициент корреляции в MS Excel?
5. В чем цель регрессионного анализа?
6. Опишите уравнение линейной регрессии.
7. Какими способами можно найти модель регрессии в MS Excel? Коротко опишите эти способы.
8. В чем задача прогнозирования данных?
9. Какими способами осуществить прогнозирование в MS Excel?
10. Что обозначает план эксперимента 34?
11. Как подключить надстройку «Поиск решения»?
12. Для чего выполняют кодирование переменных при планировании и обработке результатов эксперимента?

13. Что такое «целевая ячейка»?

14. Для чего используются относительные, абсолютные и смешанные ссылки в формулах?

15. Полный факторный план какого порядка целесообразно применить при 8 факторном эксперименте?

16. Чем отличаются уравнения регрессии в описании планов первого и второго порядка?

17. Для чего выполняется переход от натуральных размерных значений факторов к кодированным безразмерным?

1.2. Порядок выполнения задания

1. Перед выполнением задания_1 изучить теоретическую часть практикума (1.1) и ответить на контрольные вопросы.

2. Открыть новую книгу Excel и сохранить под именем «Статфункции.xls». 3. В книге выполнить задание со следующими условиями:

Имеются данные по двум экономическим показателям X и Y:

Цена (X)	995	983	1001	1012	1011	1017	978	997	1010	989	900	1100	5000
Спрос (Y)	122	144	114	100	100	90	150	130	95	155	?	?	?

Необходимо:

- вычислить коэффициент корреляции;
- построить корреляционное поле (диаграмму) на отдельном листе;
- построить регрессионную модель (с использованием функции ЛИНЕЙН);
- спрогнозировать значение Y для 3-х новых значений X с помощью функции ПРЕДСКАЗ.

Все действия (в том числе форматирование таблицы) необходимо выполнять, опираясь на образец.

4. На диаграмме разместить линию тренда с уравнением регрессии и оформить их как показано в образце. Дополнить диаграмму спрогнозированными данными (кроме последнего значения цены 5000).

5. Используя инструмент «Регрессия» на отдельном листе построить регрессионную модель с учетом новых спрогнозированных значений. Записать на листе уравнение регрессии на основании данных из «Вывода итогов».

6. Представить файл с выполненной работой преподавателю для проверки.

7. Перед выполнением задания_2 изучите теоретическую часть работы (1.2) и ответьте на контрольные вопросы.

8. Создать книгу MS Excel с названием «Корова». Лист 1 озаглавить «2-2» и воспроизвести на нем пример плана первого порядка. Лист 2 озаглавить «3-2» и воспроизвести на нем пример плана второго порядка и поиск оптимальных значений. Скопировать лист «2-2» на новый лист «доить-не-кормить» и путем подбора результатов эксперимента подтвердить проверяемую гипотезу: «Чтобы корова меньше ела и давала больше молока ее надо меньше кормить и чаще доить».

9. В п.3 примера принято решение, в результате которого для достижения цели понадобилось 12 опытов. Предложить вариант решения с достижением цели за опытов. Создать книгу «Корова2» и провести в ней расчеты по новому варианту с описанием проводимых действий в отчете по работе по аналогии с примером.

10. Оформить отчет по работе, содержащий: цель работы, описание действий, выводы по работе.

1.3. Требования к оформлению, процедура защиты

Отчет по данной работе должен содержать распечатку каждого листа книги «Статфункции.xls». При защите необходимо дать требуемые пояснения к содержанию каждого листа книги, продемонстрировать выполнение работы в файле книги «Статфункции.xls» и ответить на контрольный вопрос.

Самостоятельная работа 3 .
Методы снижения размерности многомерных данных

Цель работы: Построение информативной системы признаков. Снижение размерности признакового пространства. Применение алгоритмов факторного анализа для построения интегрированных показателей.

Исходные данные

Определить вариант работы и выбрать данные из таблицы

Вари-ант	Номера исследуемых показателей			Вари-ант	Номера исследуемых показателей		
1.	2	1	11	6.	7	1	2
2.	3	1	10	7.	8	1	3
3.	4	1	9	8.	9	1	4
4.	5	1	8	9.	10	1	5
5.	6	1	7	10.	11	1	6

Сформировать в EXCEL исходную таблицу, содержащую названия регионов и указанные в варианте показатели социально экономического развития регионов СФО.

Регион	Показатель	Показатель	Показатель
	x	*y*	*z*
…	…	…	

Порядок выполнения работы

1) Запустить модуль STA_FAC.EXE. Скопировать файл исходных данных и EXCEL в систему STATISTICA, предварительно увеличив число строк (случаев – *case*) до нужного количества.

2) Определить средствами STATISTICA описательные статистики показателей (среднее, дисперсия и др.), коэффициент корреляции

3) Построить диаграммы распределения регионов по значению признаков.

4) Провести факторный анализ:

- на стартовой панели модуля Factor Analysis (Факторный анализ) выберите все 3 переменные
- задайте метод выделения факторов (по умолчанию принимается метод Главных компонент),
- число факторов (максимальное число факторов в случае трех переменных равно 3)

5) В окне Factor Analysis Results проанализировать результаты факторного анализа:

- объясненная дисперсия собственные значения
- таблица факторных нагрузок общности решения
- значения факторов для каждого региона.

6) Сделать содержательную интерпретацию полученных результатов.

7) Оформить отчет

Самостоятельная работа 4
Методы многомерного анализа данных.
Классификация.
Кластерный и дискриминантный анализы.

Краткие сведения из теории

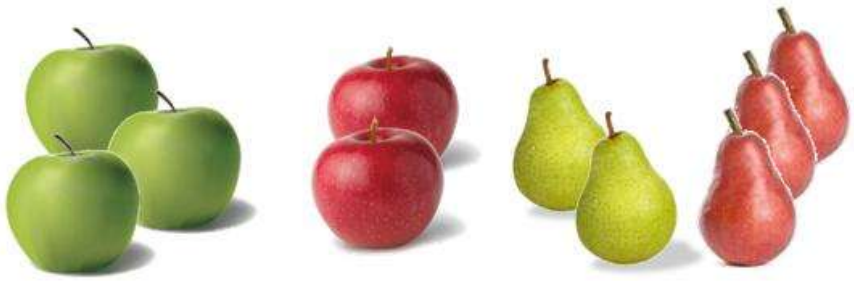
Классификацией называют разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы.

Различают классификацию при наличии обучающих выборок (дискриминантный анализ) и классификацию без обучения. К классификации без обучения относят методы автоматической ***классификации (****кластерный анализ****)***.

Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры.

Но, с другой стороны, можно сгруппировать фрукты по цветам:

Но можно сформировать больше групп, основываясь на цвете и на типе фрукта:



А если появится новый, неопознанный фрукт?

В какую группу его отнести? Или выделить под него новую группу?

В научных исследованиях задачи возникают куда более сложные и трудоемкие нежели, чем в выше приведенном примере. При наличии огромных массивов разнородных данных осуществить подобное разделение на группы (классифицировать объекты) — непростая задача.

Кластерным анализом называются разнообразные формализованные процедуры построения классификаций объектов. Лидирующей в развитии кластерного анализа наукой является биология.

Другими словами, задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами.

Предмет кластерного анализа (от англ. «cluster» — гроздь, пучок, группа) был сформулирован в 1939 г. психологом Робертом Трионом. «Классиками» кластерного анализа являются американские систематики Роберт Сокэл и Питер Снит. Одно из важнейших их достижений в этой области — книга «Начала численной таксономии», выпущенная в 1963 году. В соответствии с основной идеей авторов, классификация должна строится не на

смешении плохо формализованных суждений о сходстве и родстве объектов, а на результатах формализованной обработки результатов математического вычисления сходства/отличий классифицируемых объектов. Для выполнения этой задачи нужны были соответствующие процедуры, разработкой которых и занялись авторы.

Еще пример, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т. е. с обезьянами), чем с «отдаленными

Основные этапы кластерного анализа таковы:

1. Выбор сравниваемых друг с другом объектов.

2. Выбор множества признаков (характеристик), по которому будет проводиться сравнение и описание объектов по этим признакам.

3. Вычисление меры сходства между объектами (или меры различия объектов) в соответствии с избранной метрикой.

4. Группировка объектов в кластеры с помощью той или иной процедуры объединения.

5. Проверка применимости полученного кластерного решения (проверка построенной модели).

Кластер — это тип объектов, схожих по определенному признаку.

Если вы взглянете на географическую карту и увидите на ней горы или посмотрите на звездное небо и увидите там созвездия, то поймете, что такое кластеры.

Важно еще раз отметить, что задача кластеризации не является тривиальной. Сложность задач кластерного анализа состоит в том, что *реальные объекты являются многомерными*, т. е. описываются не одним, а несколькими параметрами, и объединение объектов в группы проводится в пространстве многих измерений, что весьма непросто. Кроме того, данные могут носить нечисловой характер.

Методы кластеризации

В целом методы кластеризации делятся на **агломеративные** (от слова агломерат — скопление) и итеративные **дивизивные** (от слова division — деление, разделение).

В ***агломеративных***, или объединительных, методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде **дендрограммы**, или дерева объединения. Это удобное представление позволяет наглядно представить кластеризацию агломеративными алгоритмами.

На каждом шаге ее составления алгоритм находит два самых близких объекта по расстоянию, подсчитанному специальным методом. Это расстояние откладывается по оси y. В итоге, исходя из расстояний на дендрограмме, можно определить необходимое количество групп.

Дендрограмма

Исходными данными для анализа могут быть собственно объекты и их параметры. Данные для анализа могут быть также представлены матрицей расстояний между объектами.

Расстояние между объектами — одна из мер сходства, чем меньше расстояние между объектами, тем они более схожи.

В биологических науках кластеризация имеет множество приложений в самых разных областях. Например, в биоинформатике с помощью неѐ анализируются сложные сети взаимодействующих генов, состоящие порой из сотен или даже тысяч элементов. Кластерный анализ позволяет выделить подсети, узкие места, концентраторы и другие скрытые свойства изучаемой системы, что позволяет в конечном счете узнать вклад каждого гена в формирование изучаемого феномена.

Дискриминантный анализ

Дискриминантный анализ является одним из методов многомерного статистического анализа.

Цель дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик (признаков, параметров) объекта классифицировать его, т. е. отнести к одной из нескольких групп (классов) некоторым оптимальным способом.

Под оптимальным способом понимается либо минимум средних потерь, либо минимум вероятности ложной классификации.

Этот вид анализа является *многомерным*, так как измеряется несколько параметров объекта, по крайней мере, больше одного, например, температура, влажность в технологическом процессе, давление, состав крови, температура больного и т. д.

Типичные **области применения** дискриминантного анализа — биология, медицина, управление производством, экономика, геология, контроль качества. В медицине объектом исследования является пациент, когда по результатам измерений различных параметров, проведения диагностических тестов врач определяет, например, необходимо ли хирургическое вмешательство при лечении. Медик может регистрировать различные переменные, относящиеся к состоянию больного, чтобы выяснить, какие переменные лучше предсказывают, что пациент, вероятно, **выздоровел** полностью (группа 1), частично (группа 2) или совсем не выздоровел (группа 3). Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы.

Задача дискриминантного анализа

Предположим, имеется n объектов с m характеристиками. В результате измерений каждый объект характеризуется вектором из m характеристик: $x1 \ldots x_m$, $m > 1$. Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких ранее определенных групп (классов) $G1,\ldots G_k$, $k >= 2$.

Иными словами, нужно построить решающее правило, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. Число групп заранее известно,

также известно, что объект заведомо принадлежит к определенной группе.

Рассмотрим простой пример:

Предположим, что вы измеряете рост в случайной выборке из 50 мужчин и 50 женщин. Женщины в среднем не так высоки, как мужчины, и эта разница должна найти отражение для каждой группы средних (для переменной Рост). Поэтому переменная «Рост» позволяет вам провести дискриминацию между мужчинами и женщинами лучше, чем, например, вероятность, выраженная следующими словами: «Если человек большой, то это, скорее всего, мужчина, а если маленький, то это вероятно женщина».

Основная идея дискриминантного анализа заключается в том, чтобы определить, отличаются ли совокупности по среднему значению какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе. Другими словами, вы хотите построить «модель», позволяющую лучше всего предсказать, к какой совокупности будет принадлежать тот или иной образец.

Алгоритм дискриминантного анализа

Решение задач дискриминации (дискриминантный анализ) состоит в разбиении всего выборочного пространства (множества реализации всех рассматриваемых многомерных случайных величин) на некоторое число областей.

Пусть имеются две генеральные совокупности X и Y, имеющие многомерный (трехмерный) нормальный закон распределения с неизвестными, но равными ковариационными матрицами.

Из этих совокупностей взяты обучающие выборки объемами n_1 и n_2 соответственно:

$$x_{11}\ x_{12}\ x_{13}$$
$$y_{11}\ \ y_{12}\ y_{13}$$
$$Xx_{21}\ x_{22}\ x_{23}\ ;$$
$$Yy_{21}\ \ y_{22}\ y_{23}$$
$$x_{n\,1}\ \ x_{n\,2}\ x_{n\,3}$$

$y_{n2}1 \quad y_{n2}2 \; y_{n2}3$

Целью дискриминантного анализа в этом случае является отнесение нового наблюдения (строки) из матрицы:

$Z\, z11 \;\; z12 \;\; z13\, z21 \;\; z22 \;\; z23$
либо к X, либо к Y.

$z_l1 \quad z_l2 \; z_l3$

Классификация населенных пунктов, расположенных в зоне радиоактивного загрязнения

Для классификации населенных пунктов по степени первоочередности проведения мероприятий радиационного или социального характера были учтены следующие факторы: демографический (численность населения, возрастная структура населения i-го населенного пункта);
хозяйственный (отношение числа жителей к числу коров); радиационный (средние значения суммарной годовой эффективной индивидуальной дозы и удельной активности молока по i-
му населенному пункту). На основе этой информации для каждого населенного пункта были рассчитаны социально-экономические и радиологические показатели, которые затем были *отнормированы* на максимальное значение. Соответствующие значения в баллах, присваиваемые i-му населенному пункту приведены в таблице. Максимальное значение — 1 балл.

№ НП	Соц.-экон.	Радиолог.
1	0,24719329	0,29540061
2	0,49097333	0,49549701
3	0,76815313	0,24014938
4	0,83789641	0,35430514
5	0,92087343	0,31977715
6	0,83693199	0,40867036
7	0,64208613	0,30919644
8	0,75447239	0,28769678
9	0,84431659	0,33989095
10	0,43312923	0,24221189
11	0,92254975	0,28911137
12	0,823976	0,24254329
13	0,96296219	0,33465852
14	0,80014316	0,22299802
15	0,82842084	0,69562283
16	1	0,36552184
17	0,71545294	0,35440414

№ НП	Соц.-экон.	Радиолог.
18	0,8853492	0,57296466
19	0,65832115	0,27772873
20	0,73896951	0,63227918
21	0,68365851	0,35711119
22	0,68365851	0,59168568
23	0,51156682	0,42806439
24	0,76328687	0,77716795
25	0,45186182	0,33931741
26	0,46782343	0,42848911
27	0,84907526	0,29800686
28	0,76096616	0,25151952
29	0,29841345	0,20134517
30	0,82047266	0,43816498
31	0,88639612	0,2515159
32	0,76052719	0,53358935
33	0,8274442	0,36113046
34	0,89421902	1
35	0,38097797	0,45873837
36	0,7502056	0,30138691
37	0,61300179	0,51226607
38	0,44290376	0,36935574
39	0,85262386	0,23098893
40	0,78977027	0,39782878
41	0,74822401	0,52675873
42	0,89249236	0,36927656
43	0,73758751	0,65779693
44	0,61224609	0,24290559
45	0,77375606	0,23284717
46	0,93288537	0,2708384

Необходимо разделить населенные пункты на соответствующие классы при помощи процедуры кластерного анализа в программе «Statistica».

1. Запустите программу и укажите исходные настройки как показано на рисунке:

2. Скопируйте показатели в окно программы:

3. Запустите процедуру кластерного анализа: *Анализ — Многомерный разведочный анализ — Кластерный анализ*.

4. Выберите метод кластеризации К-средних.

5. В появившемся окне укажите *значения переменных* (выберите все), *объекты установите — наблюдения (строки)*. Попытаемся разбить объекты на *3 кластера*.

6. После нажатия кнопки «ОК» появляются результаты обработки:

7. Перейдите во вкладку «*Дополнительно»*.

8. Нажмите кнопку «Элементы кластеров и расстояния». В результате вы получите 3 таблицы, показывающие, какие объекты относятся к одному из трех кластеров.

В строках таблиц указано расстояние от каждого населенного пункта до центра кластера.

9. Вернитесь в окно анализа и нажмите кнопку «*Средние кластеры и евклидовы расстояния»*.

Результат обработки появится на экране.

Над диагональю в таблице даны квадраты расстояний между кластерами. 10. Вернитесь в окно анализа и нажмите кнопку «*График средних».* В результате строятся следующие графики средних значений характеристик населенных пунктов для каждого кластера.

11. По результатам проведения анализа (пункт 8) создайте таблицу.

Сколько кластеров вы получили? Можно ли было сделать больше кластеров или меньшее ихколичество?

12. Построение дендрограммы. Вернитесь в окно анализа и закройте его. В появившемся окне выберите «*Иерархическая классификация»*.

13. Задайте установки как показано на следующих рисунках:

14. После нажатия кнопки «ОК» появится окно результатов вычисления. В котором необходимо нажать «*Вертикальная дендрограмма»*.

Дискриминантный анализ. Задача

В предыдущем анализе мы получили таблицу, в которой показано то как объекты распределены по трем классам в зависимости от двух параметров. Преобразовав эту таблицу, выполним дискриминантный анализ данных и проверим, к какому классу можно отнести новый населенный пункт (социально-экономические показатели — 0,78; радиологические — 0,61).

1. Задайте исходные настройки нового файла программы.

2. Скопируйте таблицу в появившееся поле и оформите согласно рисунку.

3. Запустите процедуру дискриминантного анализа: Анализ — Многомерный разведочный анализ — Дискриминантный анализ.

4. Укажите Группирующую переменную как 3-Klass (Var3), а независимые переменные — 1 Soc-ec и 2 Radio (Var1 и 2).

5. Нажмите на кнопку коды переменной, затем «ОК».

6. Нажмите «ОК» и в появившемся окне задайте установки как представлено на рисунке. Затем «ОК».

Информационная часть окна сообщает, что использовано: Число переменных в модели: 2;

Лямбда Уилкса: 0,0689068;

приближ. F (4,84) = 58,999 (Приближенное значение F-статистики), связанной с лямбдой Уилкса;

p — уровень значимости F-критерия для значения 58,999; значения статистики лямбда Уилкса лежат в интервале 0–1.

Значения статистики Уилкса, лежащие около нуля, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около единицы, свидетельствуют о плохой дискриминации.

Иными словами, это можно выразить следующим образом: если значения лямбды Уилкса близки к нулю, то мощность дискриминации (мощность = 1 — вероятность ошибки) близка к 1, если лямбда Уилкса близка к единице, то мощность близка к нулю.

7. Перейдите во вкладку дополнительно и нажмите кнопку Переменные в модели.

Результаты свидетельствуют о хорошей дискриминации.

8. Просмотрите разделение групп на графике. Для этого инициируйте кнопку «Канонический анализ». В появившемся

диалоговом окне перейдите во вкладку *Канонические значения* и выберите «Диаграмма рассеяния для канонических значений».

Населенные пункты, находящиеся во втором кластере приоритетны в распределении инвестиций, направленных на их социально-экономическое развитие. При этом наиболее перспективными являются населенные пункты, расположенные в правой части этого кластера. Эти пункты имеют стабильный социально-экономический фактор.

9. Апостериорные вероятности. Нажав кнопку «Апостериорные вероятности», вы увидите таблицу с апостериорными вероятностями принадлежности объекта к определенному классу.

Что характеризует с какой вероятностью объект принадлежит тому или иному классу.

10. Добавьте новую строку в исходную таблицу в окне программы (двойной щелчок по пустому затемненному полю слева от таблицы под цифрой 46). И введите туда значения нового исследуемого населенного пункта, у которого социально-экономические показатели — 0,78; радиологические — 0,61. Определите, к какому классу относится объект.

11. Для этого не закрывая предыдущий анализ запустите новый, процедура та же, что и описанная выше.

12. В появившемся окне выберите «Апостериорные вероятности».

13. По результатам вычисления видно, что новый населенный пункт № 47 с вероятностью 99 % относится к 3 классу.

Контрольные вопросы

1. Что означает классифицировать объект?
2. Какие статистические методы классификации вам известны?
3. Для чего используется кластерный анализ?
4. Что такое кластер?
5. Результат кластерного анализа.
6. Для чего используется дискриминантный анализ?
7. Что такое дискриминирующая переменная?
8. Что такое обучающая выборка?
9. Цель дискриминантного анализа.

ЛИТЕРАТУРА

1. *Гланц, С.* Медико-биологическая статистика: пер. с англ. / Гланц С. — М., Практика, 1998. — 459 с.

2. Платонов, А. Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А. Е. Платонов. — М.: Издательство РАМН, 2000. — 52 с.

3. *Жученко, Ю. М.* Информационные технологии в биологии и химии: лабораторный практикум для студентов вузов по специальности 1-31 01 01 «Биология» / Ю. М. Жученко. М-во образования РБ, Гомельский гос. ун-т им. Ф. Скорины. — Гомель: ГГУ им. Ф. Скорины, 2010. — 148 с.

4. *Реброва, О. Ю.* Статистический анализ медицинских данных. Применение пакета рикладных программ STATISTICA / О. Ю. Реброва. — 3-е изд. — М., МедиаСфера, 2006. — 312 с.

5. *Халафян, А. А.* STATISTICA 6. Статистический анализ данных / А. А. Халафян. — 3-е изд. — М.: ООО «Бином-Пресс», 2007. — 512 с.

6. Обработка экспериментальных данных в MS Excel: методические указания к выполнению лабораторных работ для студентов дневной формы обучения / сост. Е. Г. Агапова, Е. А. Битехтина. — Хабаровск: Изд-во Тихоокеан. гос. ун-та, 2012. — 32 с.

7. STATISTICA 6.0 — фирменное руководство. Компания StatSoft. Электронная публикация, 1995.

8. *Максимов, С. И.* Статистический анализ и обработка данных с применением MSExceli SPSS: учеб.-метод. пособие / С. И. Максимов. — Минск: РИВШ, 2012. — 114 с.

9. *Джелен, Б.* Сводные таблицы в MSExcel 2010.: пер. с англ. / Б. Джелен, М. Александер. — М.: ООО «И. Д. Вильямс», 2011. — 464 с.

10. *Лялин, В. С.* Статистика: теория и практика в Excel: учеб. пособие / В. С. Лялин, И. Г. Зверева, Н. Г. Никифорова. — М.: Финансы и статистика; ИНФРА-М, 2010. — 448 с.

11. *Mario, F. Triola*. Elementary statistics / F. Triola Mario. — 10 th ed. — 770 с. 12. *Боровиков, В.* Statistica. Искусство анализа данных на компьютере:

Для профессионалов / В. Боровиков. — 2-е изд. — Спб.: Питер, 2003. — 688 с.

Самостоятельная работа 5. Цифровая обработка изображений

Функции геометрического преобразования изображений

Цель работы: Изучение и использование стандартных функций среды MATLAB выполняющих геометрические преобразования изображений.

Справочные сведения

1. Загрузите изображение в среду MATLAB. Определите размеры изображения (число строк и столбцов).

Функция size(Im) возвращает количество строк и столбцов в изображении.

2. Выполните кадрирования изображения. Необходимо вырезать $\frac{1}{4}$ изображения.

Функция imcrop возвращает изображение, ограниченное заданным прямоугольником. Используя функции rIm=imcrop(Im, rect), можно явно определить ограничивающий прямоугольник, где rect - вектор из четырех элементов: [xmin ymin w h], которые задают положение левого верхнего угла (xmin ymin), ширину (w) и высоту (h) .

3. Над полученным изображением (из задания 2) выполните масштабирование с коэффициентами 4 и 0.25 и заданным методом интерполяции. Результат и исходное изображение выведите в одно графическое окно. Проанализируйте наличие артефактов.

Функция zIm=imresize(кIm, m, method) создает изображение zIm, размеры которого в m раз отличаются от размеров исходного изображения. Если m принадлежит диапазону от 0 до 1, то происходит уменьшение изображения, если m больше 1, то увеличение. Для изменения размеров используется один из предопределенных методов интерполяции, который задается во входном параметре method в виде одной из следующих строк:

‘nearest’ - использовать значение ближайшего пиксела (установлен по умолчанию, и данный параметр может быть опущен при вызове функции);

‘bilinear’ - использовать интерполяцию по билинейной поверхности; ‘bicubic’ - использовать интерполяцию по бикубической поверхности.

4. Выполните поворот изображения на 360^0 с шагом 15^0 и заданным методом интерполяции. При выполнении задания использовать возможности анимации для демонстрации результата.

Функция rIm=imrotate(Im, angle, method) создает изображение rIm, соответствующее повернутому исходному изображению, используя один из методов интерполяции. Метод интерполяции определяется входном параметре method в виде одной из следующих строк:

"nearest' - использовать значение ближайшего пиксела (установлено по умолчанию, и данный параметр может быть опущен при вызове функции);

"bilinear' - использовать интерполяцию по билинейной поверхности; "bicubic' - использовать интерполяцию по бикубической поверхности. Угол поворота angle задается в градусах. Положительные значения данного параметра соответствуют повороту против часовой стрелки, а отрицательные - по часовой стрелке.

Функция getframe создает фрейм для анимации. Функция movie воспроизводит анимацию.

Пример. Необходимо анимировать масштабирование с коэффициентами: 0.5, 1, 1.5. Предусмотреть 10 повторов анимации.

```
z=0.5; for t=1:3
zIm=imresize(Im, z); imshow(zIm); z=z+0.5; M(t)=getframe;

end
movie(M, 10)% 10  количество повторов
```

5. Разработайте М – функцию автоматизирующую выполнение заданий лабораторной работы.

Геометрические преобразования векторных изображений

Цель работы: Изучение и программная реализация в среде MATLAB алгоритмов геометрических преобразований на плоскости (2-D преобразований).

1. Разработать и протестировать M - функцию построения случайного многоугольника для различных максимальных углов приращения W: 90^0, 120^0, 150^0, 180^0. Алгоритм построение приведен в Приложении 1. Координаты вершин должны храниться в матрице.
2. Построить диаграмму процентного соотношения количества вершин многоугольника от угла W.

Справочные сведения

Функция rand('seed', x0) начальному значению генератора случайных чисел значение x0.

Функция X = rand(n) формирует массив размера n на n, элементами которого являются случайные величины, распределенные по равномерному закону в интервале (0, 1).

Функция rand без аргументов формирует случайное число, подчиняющееся равномерному закону распределения в интервале (0, 1).

1. Разработать и протестировать M - функцию, выполняющую 2-D преобразование над случайным многоугольником. Вариант преобразования задается преподавателем (см. Таблица 1).

Таблица 1. Варианты преобразований

Вариант	Преобразование	Матрица преобразования
1	Поворот относительно произвольной точки О	$\begin{bmatrix} \cos(\varphi) & \sin(\varphi) & 0 \\ -\sin(\varphi) & \cos(\varphi) & 0 \\ X & Y & 1 \end{bmatrix}$ $X = o_x(1-\cos(f)) + o_y \sin(f)$ $Y = o_y(1-\cos(f)) + o_x \sin(f)$
2	Перемещение	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ dx & dy & 1 \end{bmatrix}$
3	Масштабирование	$\begin{bmatrix} zx & 0 & 0 \\ 0 & zy & 0 \\ 0 & 0 & 1 \end{bmatrix}$
4	Отражение от прямой $x=a$	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 2a & 0 & 1 \end{bmatrix}$
5	Сдвиг	$\begin{bmatrix} 1 & S2 & 0 \\ S1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

2. Произвести анимацию преобразования путем дискретизации. Результат анимации сохранить в файл формата GIF. Разработать соответствующую М – функцию.

3. Выведите информацию о полученном изображении. Для этого воспользуйтесь функцией **imfinfo**.

Функция comet(y) рисует движение точки по траектории, заданной одномерным массивом y, в виде головы и хвоста кометы.

Функция comet(x, y) рисует движение точки по траектории, заданной массивами x и y.

Функция plot(x, y) соответствует построению обычной функции, когда одномерный массив x соответствует значениям аргумента, а одномерный массив y - значениям функции. Когда один из массивов X или Y либо оба двумерные, реализуются следующие построения:

- если массив Y двумерный, а массив x одномерный, то строятся графики для столбцов массива Y в зависимости от элементов вектора x;
- если двумерным является массив X, а массив y одномерный, то строятся графики столбцов массива X в зависимости от элементов вектора y;
- если оба массива X и Y двумерные, то строятся зависимости столбцов массива Y от столбцов массива X.

Функция rgb2ind создает палитровое изображение из полноцветного, используя один из четырех способов: запись в виде палитрового изображения без уменьшения количества цветов, установка равномерной палитры, оптимальный подбор палитры, использование некоторой предопределенной палитры.

Команда [im, map] =rgb2ind(RGB) создает палитровое изображение im из полноцветного f, составляя палитру map из всех уникальных цветов, представленных в исходном изображении. Результирующая палитра map может быть очень большого размера.

Команда [im,map] = rgb2ind(f,256) создает палитровое изображение im из полноцветного f, 256 — ограничение на количество цветов в палитре.

Функция imwrite(im,map,'test.gif','DelayTime',0,'LoopCount',inf) записывает анимацию в GIF – файл.

DelayTime определяет время задержки между кадрами анимации, LoopCount задает число повторений. LoopCount=inf зацикливает анимацию.

Функция imfinfo(filemane) возвращает информацию об изображении и способе его хранения из файла с именем filemane. Информация заносится в структуру info. Структуры info различаются для разных форматов файлов, однако первые 9 полей всегда содержат следующую общую информацию, которую можно представить в виде Таблицы 2.

Таблица 2 - Структура info

Имя поля	**Тип**	**Описание**
FileName	Строк а	Имя файла, если файл находится в текущей директории, или полный путь к файлу
FileModeDa te	Строк а	Дата и время последней модификации файла
FileSize	Число	Размер файла в байтах
Format	Строк а	Формат файла, совпадающий с параметром ***fmt***
FormatVersi on	Строк а или число	Версия формата
Width	Число	Ширина изображения в пикселях
Height	Число	Высота изображения в пикселях
BitDepth	Число	Глубина изображения в битах на пиксель
ColorType	Строк а	Тип изображения: □□ ***'truecolor'*** или ***'RGB'*** для полноцветных изображений; □□ ***'grayscale'***для полутоновых; □□ ***'indexed'***для палитровых

Пример. Необходимо произвести операцию сдвига над треугольном с координатами вершин: 1 1; 2 1;1 2. В матрице преобразования S1=0 и S2=5. Результат сохранить в GIF – файл с анимацией пошагового преобразования с табуляцией S2 с шагом 0.5.

```
s= 0.5:0.5:5;
Tr=[1 1;2 1;1 2;1 1]; plot(Tr(:,1),Tr(:,2)); axis( [ 0, 3, 0, 12] );
f = getframe;
[im,map] = rgb2ind(f.cdata,256); im(1,1,1,10) = 0;
for k = 1:10 Sh=[1 s(k);0 1];
```

*Trs=Tr*Sh plot(Trs(:,1),Trs(:,2));*

axis([0, 3, 0, 12]); f = getframe;

im(:,:,1,k) = rgb2ind(f.cdata,map); end

imwrite(im,map,'test.gif','DelayTime',0,'LoopCount',inf)

Список литературы

1. Никулин Е. А. Компьютерная геометрия и алгоритмы машинной графики. – СПб.: БХВ-Петербург, 2003. – 560 с.: ил.
2. Потемкин В. Г. Справочник по MATLAB. Графические команды и функции //Интернет–ресурс: http:// matlab.exponenta.ru/ imageprocess/ index.php (Дата обращения: 24.08.2016).
3. Р. Гонсалес, Р. Вудс, С. Эддинс. Цифровая обработка изображений в среде MATLAB.– Москва:Техносфера, 2006. – 616 с.
4. Фисенко В.Т., Фисенко Т.Ю. Компьютерная обработка и распознавание изображений: учеб. пособие. - СПб: СПбГУ ИТМО, 2008. – 192 с.
5. Р. Гонсалес, Р. Вудс. Цифровая обработка изображений. – Москва: Техносфера, 2012. – 1104 с.

Приложение 1

Алгоритм построения многоугольника со случайным количеством вершин (не менее 3-х)

Рассмотрим алгоритм генерирования случайного многоугольника с вершинами (не менее трех) последовательно соединенными друг с другом не пересекающейся замкнутой ломанной линией со случайным направлением обхода:

1. Из заданной точки О как из центра проводим лучи по $0^0 \le \varphi \le 360^0$ углами к оси ОХ. Начальное значения угла $f_1=0^0$, а последующие углы рассчитываем путем приращения $f_{n+1} = f_n + \Delta f_n$ на случайное значение $\Delta f_n = B\ rand()$. Максимальное приращения углов $120^0 \le B \le 180^0$ выбираем с таким расчетом, чтобы минимальное число лучей было равно трем.

2. Вдоль лучей откладываем расстояния r_n=a+(b-a)*rand(), генерируемые как случайные числа в диапазоне от a до b. В итоге, получаем вершины многоугольника:

$$p_n = 0 + r[\cos(\varphi_n)\sin(\varphi_n)],\ n = 1,\ 2,\ \ldots .$$

3. Генерируем $d=2\ rand()$. При $d \ge 1$нумеруем вершины P в обратном порядке.

4. Возврат P.

Приложение 2

Данные с сайта http://cfo.gov.ru – официальный сайт ЦФО.

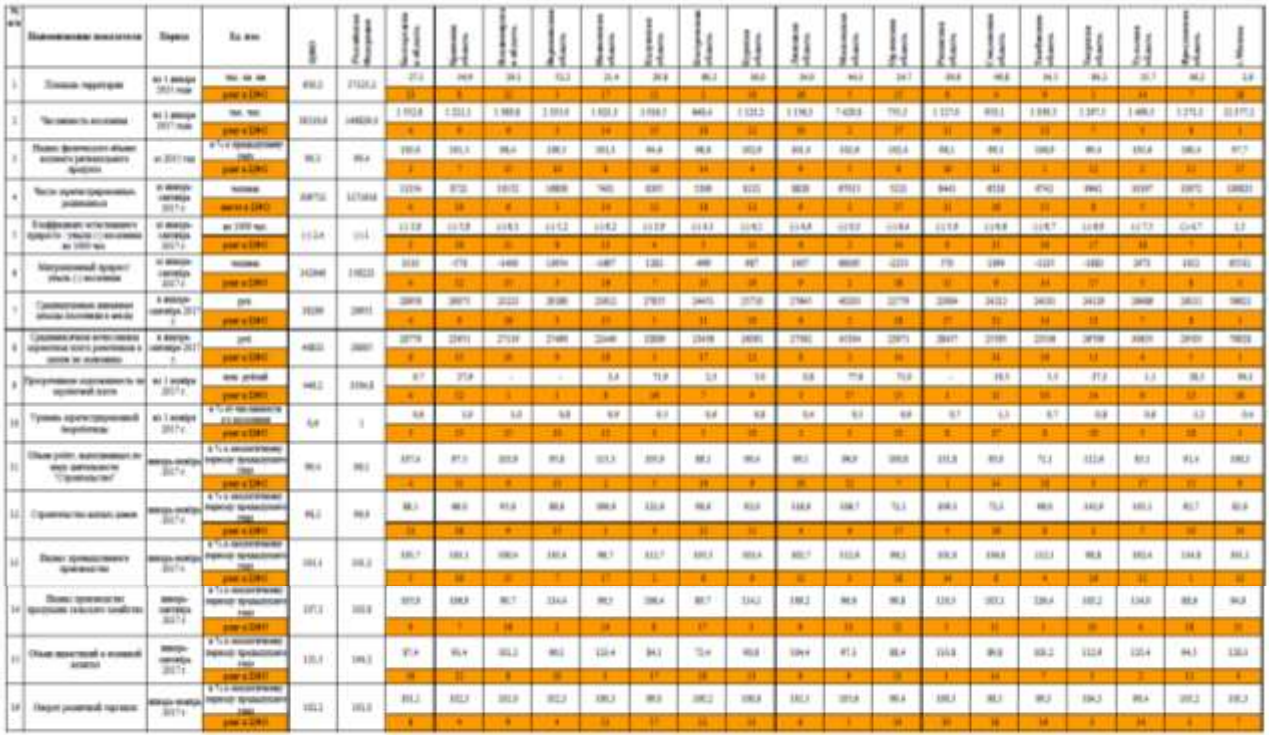


Рисунок 1 – Таблица основных социально-экономических показателей субъектов Российской Федерации, находящихся в пределах Центрального федерального округа