Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ» (ТУСУР)**

Кафедра автоматизации обработки информации (АОИ)

АНАЛИЗ ДАННЫХ

Методические указания к лабораторным работам
и организации самостоятельной работы
для студентов направления «Бизнес-информатика»
(уровень бакалавриата)

2018

Лепихина Зинаида Павловна
Анализ данных: Методические указания к лабораторным работам и организации самостоятельной работы для студентов направления «Бизнес-информатика» (уровень бакалавриата) / З.П.Лепихина. – Томск, 2018. – 65 с.

Оглавление

1 Введение

Цель изучения дисциплины «Анализ данных» — формирование у студентов теоретических представлений об основных современных методах анализа данных, основных типах задач, решаемых методами многомерного анализа данных, развитие навыков использования современных технологий обработки данных для решения исследовательских и практических задач.

Задачи изучения дисциплины:

– развить навыки и способности студентов к применению современных теоретических и эмпирических моделей для решения конкретных задач анализа данных;

– сформировать навыки сбора необходимой информации и использования соответствующего математического аппарата и инструментальных средств для обработки, анализа и систематизации информации по теме исследования.

В данных Методических указаниях содержится:

- кратное изложения теоретического материала по теме лабораторной работы, варианты заданий и порядок их выполнения;
- рекомендации по организации самостоятельной работы.

Лабораторные работы выполняются с использованием табличного процессора MS Excel (LibreOffice Calc, OpenOffice Calc). Форма контроля выполнения лабораторной работы: демонстрация преподавателю расчетов и результатов анализа, собеседование, ответы на вопросы, выполнение дополнительных заданий.

При подготовке к лабораторным и при выполнении заданий в рамках самостоятельной работы студенту следует повторить теоретический материал по конспекту лекций и источникам, приведенным в разделе «Рекомендуемая литература», а также пользоваться информацией, представленной в статистических сборниках, в научной литературе и Интернете

2 Методические указания к проведению лабораторных работ

2.1 Лабораторная работа «Первичный анализ данных на компьютере»

Цель работы

Получение практических навыков вычисления статистических показателей в среде табличного процессора MS Excel, анализа и представления результатов расчетов.

Форма проведения

Выполнение индивидуального задания.

Форма отчетности

Устный опрос, демонстрация расчетов, выполнение дополнительных заданий.

Теоретические основы

Табличный процессор Microsoft Excel (LibreOffice Calc, OpenOffice Calc) представляет собой визуальную среду с большим набором библиотечных функций, позволяющую выполнять вычисления различного характера при анализе социально-экономических данных.

Абсолютными величинами в статистике называются величины, характеризующие размеры (уровни, объемы) общественных явлений в конкретных условиях места и времени.

Индивидуальными называют абсолютные статистические величины, характеризующие размеры признака у отдельных единиц совокупности (например, размер заработной платы отдельного работника, вклада гражданина в определенном банке и т.д.)

Суммарные абсолютные статистические величины характеризуют итоговое значение признака по определенной совокупности объектов. Они являются суммой количества единиц совокупности (численность совокупности) или суммой значений варьирующего признака всех единиц совокупности (объем варьирующего признака).

Относительная величина – это обобщающий показатель, который представляет собой частное от деления одного показателя на другой и дает числовую меру соотношений между ними.

Величина, с которой производится сравнение (знаменатель дроби), обычно называется базой сравнения или основанием. Относительные величины измеряются в «разах» или в процентах (%), промилле (‰) т.п.

Относительная величина динамики характеризует изменение уровня какого-либо явления во времени. Относительные величины динамики называются коэффициентами роста (показывают во сколько раз значение показателя в момент времени *t1* больше того же показателя в момент времени *t0*) или темпами роста (показывают сколько процентов составляет значение показателя в момент времени *t1* по сравнению с тем же показателем в момент времени *t0*).

$$ОВД = \frac{П_{t1}}{П_{t0}}$$

Относительными величинами структуры называются показатели, характеризующие долю отдельных частей изучаемой совокупности во всем ее объеме.

$$ОВС = \frac{y_i}{\sum_{i=1}^{k} y_i}$$

Yi –объем *i-й* части совокупности, *i=1,2,…,k*

k – число частей, на которое поделена совокупность

Относительными величинами координации называют показатели, характеризующие соотношение отдельных частей целого между собой.

$$ОВК = \frac{y_i}{y_j}$$

Yi, Yj –объем *i-й* и *j-й* частей совокупности, *i,j=1,2,…,k*

k – число частей, на которое поделена совокупность

Относительными величинами наглядности (сравнения) называют показатели, представляющие собой частное от деления значений одного и того же статистического показателя, характеризующих разные объекты А и Б (предприятия, фирмы, районы, области, страны и т.д.) и относящихся к одному и тому же периоду времени.

$$ОВН = \frac{П_А}{П_Б}$$

Группировка – это распределение единиц по группам в соответствии со следующим принципом: различия между единицами, отнесенными к одной группе, должны быть меньше, чем между единицами, отнесенными к разным группам.

Группировку необходимо проводить, если совокупность неоднородна. Однородность совокупности оценивается коэффициентом вариа-

ции: $K_{\text{var}} = \bar{x}/\sigma$. Если коэффициент вариации более 30%, то совокупность считается неоднородной.

Типологическая группировка служит для выделения социально-экономических типов.

Структурная группировка характеризует структуру совокупности по какому-либо одному признаку, ее элементами являются относительные величины структуры.

Структурная группировка позволяет изучать динамику структуры совокупности.

Пусть w_{i0} и w_{i1} - доли *i*-ой группы в период «0» и «1». Показатели среднего абсолютного изменения структуры:

$$d_{w_1 - w_0} = \frac{\sum_{j=1}^{k} \left| w_{j_1} - w_{j_0} \right|}{k},$$

где *k* — число групп.

Средний квадратический показатель структурных сдвигов строится на основе формулы стандартного отклонения:

$$S_{w_1 - w_0} = \sqrt{\frac{\sum_{j=1}^{k} (w_{j_1} - w_{j_0})^2}{k}}.$$

При отсутствии структурных сдвигов эти показатели равны нулю; их величина тем больше, чем значительнее абсолютные изменения удельных весов групп.

Результаты статистического исследования *представляются в виде статистических таблиц и графиков.*

Статистическая таблица — система строк и столбцов, в которых в определенной последовательности и связи излагается статистическая информация о социально-экономических явлениях.

Различают подлежащее и сказуемое статистической таблицы.

В подлежащем указывается характеризуемых объект — либо единица совокупности, либо группы единиц, либо совокупность в целом. В сказуемом дается характеристика подлежащего, обычно в количественной форме в виде системы показателей.

По характеру подлежащего статистические таблицы подразделяются на простые, групповые и комбинационные.

В подлежащем простой таблицы объект изучения не подразделяется на группы, а дается либо перечень всех единиц совокупности, либо указывается совокупность в целом. Единицы упорядочиваются (по алфа-

виту, по возрастанию, по убыванию). В подлежащем групповой таблицы совокупность подразделяется на группы по одному признаку.

В сказуемом указываются число единиц в группах (абсолютное и/или в процентах к итогу) и сводные показатели по группам. В подлежащем комбинационной таблицы совокупность подразделяется по группам не по одному, а по нескольким признакам.

По характеру сказуемого статистические таблицы делятся на таблицы с простой разработкой сказуемого и таблицы со сложной разработкой сказуемого.

В таблицах с простой разработкой сказуемого показатели, характеризующие подлежащее, получаются путем простого суммирования значений по каждому признаку независимо друг от друга. Сложная разработка сказуемого предполагает деление признака на группы.

При оформлении таблиц необходимо соблюдать *следующие правила*.

Обязателен заголовок таблицы, в котором указывается, к какой категории и к какому времени относится таблица. В таблице не должно быть лишних линий. Может быть горизонтальная черта, отделяющая итоговую строку. Вертикальные линии могут быть, а могут отсутствовать. Заголовки граф содержат названия показателей без сокращения слов и единиц измерения. Общие единицы измерения могут быть вынесены в заголовок таблицы. Итоговая строка завершает таблицу и располагается внизу таблицы. Иногда итоговая строка бывает первой, в этом случае второй строкой идет строка «в том числе» или «из них». Цифровые сведения записываются в пределах каждой графы с одной и той же степенью точности.

Статистические графики представляют собой условные изображения числовых величин и их соотношений посредством линий, геометрических фигур, рисунков или географических карт-схем.

Графики обязательно сопровождаются заголовками, в которых указывается, какой показатель изображен, в каких единицах измерения, по какой территории и за какое время он определен. На графике должен быть указан масштаб — мера перевода числовой величины в графическую.

По способу построения статистические графики делятся на диаграммы (линейные, объемные, плоскостные, радиальные, точечные, фигурные), картограммы и картодиаграммы.

Среди плоскостных диаграмм часто используются столбиковые диаграммы, на которых величина столбика соответствует значению показателя. Линейные графики обычно используются для представления динамики показателя. Пример линейного графика приведен на рис.1.

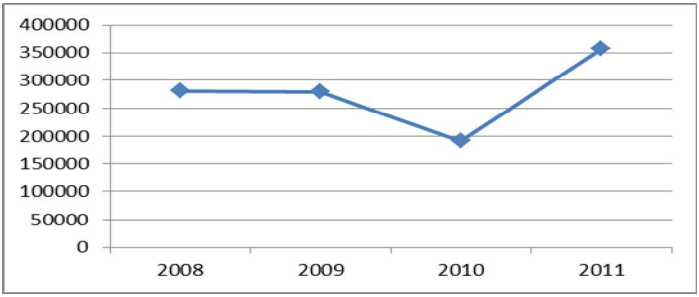


Рисунок 1. Динамика прибывших в РФ (чел.)

Для иллюстрации структуры совокупности используется секторная диаграмма. Вся совокупность принимается за 100 процентов, ей соответствует вся площадь круга, а площади секторов соответствуют частям совокупности.

Варианты заданий

Исходные данные о международной миграции приведены в Приложении 1.

Вариант	***Год 1***	***Год 2***	***Вариант***	***Год 1***	***Год 2***
Вариант 1	2005	2010	Вариант 6	2005	2006
Вариант 2	2006	2011	Вариант 7	2006	2007
Вариант 3	2007	2012	Вариант 8	2007	2008
Вариант 4	2008	2013	Вариант 9	2008	2009
Вариант 5	2009	2014	Вариант 10	2009	2010

Порядок выполнения работы

1) В соответствии с вариантом выбрать из исходной таблицы Приложения 1 данные за два указанных в варианте года и представить их в виде **рабочей табл.1**.

Таблица 1. Исходные данные

Название страны	**Число прибывших в РФ (чел.)**	
	Год1	***Год2***
Страна_1	a_{11}	a_{12}
…	…	…
Страна_i	a_{i1}	a_{i2}
…	…	…
Страна_n	a_{n1}	a_{n2}

где a_{ij} – число мигрантов из страны i в j-м году

2) По каждому году

a) Представить графически (столбиковая диаграмма) значения показателя у стран.

b) Пользуясь статистическими процедурами Excel, определить:
   - итоговое (суммарное) значение (СУММ):
   - максимальное и минимальное значение признака(МАКС, МИН);
   - среднее значение (СРЗНАЧ);
   - медиану (МЕДИАНА);
   - моду (МОДА);
   - дисперсию (ДИСПР);
   - среднее квадратическое отклонение (СТАНДОТКЛОН).

c) Вычислить коэффициенты вариации.

d) Сделать вывод об однородности совокупности.

3) На основе **рабочей табл.1**провести группировку стран, объединив данные в 3 группы (типа): «из европейской части бывшего СССР», «из азиатской части бывшего СССР», «из стран дальнего зарубежья». Данные оформить в виде **рабочей табл.2** (графы 1,2,3).

Таблица 2 Показатели численности, структуры и динамики миграции

Группы стран	*Численность прибывших в РФ (чел.)*		*Структура прибывших в РФ (в процентах)*		*Темп роста (в процентах)*
	Год 1	*Год 2*	*Год 1*	*Год 2*	
1	***2***	***3***	***4***	***5***	***6***
из европейской части бывшего СССР					
из азиатской части бывшего СССР					
из стран дальнего зарубежья					
Итого					

4) Провести расчеты структуры прибывших по каждому году и занести результаты в табл.2 (графы 4,5).

5) Рассчитать линейный и средний квадратический показатели изменения структуры (структурных сдвигов).

6) Рассчитать относительные величины динамики (темпы роста) численности мигрантов и занести результаты в графу 6 табл.2.

7) По данным табл. 2 построить для абсолютных показателей столбиковую диаграмму.

8) Для относительных величин, характеризующих структуру статистической совокупности, построить секторные диаграммы (по каждому году отдельно).
9) Рассчитать относительные величины координации.
10) Провести ***содержательный анализ*** полученных результатов расчетов.

Контрольные вопросы и задания

1) Что является подлежащим таблицы, что – сказуемым?
2) Какие виды относительных величин были использованы?
3) Какие виды группировок были использованы?
4) Удельный вес какой группы мигрантов наибольший?
5) Какие показатели характеризуют структурные сдвиги?
6) Как вычисляются относительные величины динамики и координации, что они характеризуют?
7) Какие содержательные выводы можно сделать о динамике абсолютных показателей и структурных сдвигах?

2.2 Лабораторная работа «Анализ взаимосвязи признаков»

Цель работы

Закрепление теоретического материала и получение практических навыков расчета и анализа показателей взаимосвязи нечисловых данных.

Форма проведения

Выполнение индивидуального задания.

Форма отчетности

Устный опрос, демонстрация расчетов, выполнение дополнительных заданий.

Теоретические основы

Анализ четырехклеточных таблиц.

Рассмотрим генеральную совокупность из n объектов (индивидов) в которой классификация произведена на основании наличия или отсутствия двух дихотомических (бинарных) признаков A и B.

Тогда количества попаданий в 4 возможные подгруппы могут быть представлены таблицей 2×2 (четырехклеточная) и ее записывают в виде:

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

Суммы *a+b, a+c, c+d, b+d* называются маргинальными суммами. Если между A и B не существует никакой связи, т.е. если обладание признаком A не связано с обладанием признаком B, то доля индивидов с признаком A среди индивидов, обладающих признаком B, должна быть равна доле индивидов с признаком A среди индивидов, не обладающих признаком B.

Таким образом, *по определению признаки* ***независимы*** в данной совокупности из n наблюдений, если

$$\frac{a}{a+c}=\frac{b}{b+d}=\frac{a+b}{n}. \qquad (2.2.1)$$

Соотношение (2.2.1) можно переписать в виде:

$$a=\frac{(a+b)(a+c)}{n}$$

Это теоретическая частота в предположении независимости.

Теперь, если для какой-либо таблицы выполнено неравенство $a>\frac{(a+b)(a+c)}{n}$, означающее, что доля A среди B больше, чем среди не B, то A и B называют *положительно связанными* или просто связанными.

Если имеем противоположное неравенство, то есть $a<\frac{(a+b)(a+c)}{n}$, то A и B *отрицательно связаны.*

Меры связи в таблицах 2x2

Коэффициент ассоциации

$$Ka=\frac{ad-bc}{ad+bc}.$$

Коэффициент контингенции:

$$Kk=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

Коэффициенты изменяются в диапазоне

$$-1\le Ka, Kk\le +1$$

Если признаки независимы, то *Ka, Kk*=0;

Ka, K k = +1 в случае полной положительной связи,

Ka, Kk = -1, если полная отрицательная связь между признаками.

Анализ таблиц $r \times c$

Рассмотрим совокупность из *n* объектов, каждый из которых описан двумя признаками, измеренными в шкале наименований. Первых признак имеет ***r*** градаций (располагается по строкам таблицы сопряженности), второй признак имеет ***c*** градаций (располагается в столбцах таблицы сопряженности). Задача анализа взаимосвязи заключается в установлении при заданном значении *уровня значимости* α наличия или отсутствия статистической связи (проверка гипотезы Но) и измерении силы связи.

Классическим тестом, использующимся в тех случаях, когда данные расклассифицированы в таблице с двумя входами, является χ^2 - тест К.Пирсона:

$$\chi^2 = \sum \sum \frac{(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{n})^2}{\frac{n_{i.} \cdot n_{.j}}{n}},$$

где n_{ij} - наблюдаемая частота (число объектов) в ячейке (i, j) таблицы;

$e_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$ - теоретически ожидаемая частота (по H_0 - предположение о статистической независимости рассматриваемых переменных) в этой ячейке;

$i = 1, 2, \ldots, r$ – r - число строк;

$j = 1, 2, \ldots, c$ – c - число столбцов;

$n_{i.} = \sum_{j=1}^{c} n_{ij}$ - сумма по строке;

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ - сумма по столбцу;

$n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$ - общее число объектов или объем выборки.

Число степеней свободы для таблицы сопряженности $r \times c$ равно $df = (r-1)(c-1)$. Заметим, что для таблицы 2×2 $df = 1$.

Установление статистической связи между признаками проводится по правилам проверки гипотез.

Шаг 1. Выдвигается гипотеза H_0: статистическая связь между признаками отсутствует, то есть обладание первым признаком никак не связано с обладанием вторым признаком.

Шаг 2. По исходной таблице сопряженности вычисляем по формуле (1.2) фактическое значение χ^2 .

Шаг 3. Вычисляем число степеней свободы df . Задаем уровень значимости α . В таблице распределения статистики $\chi^2_{df,\alpha}$ отыскиваем значения этой величины при заданном уровне значимости α и при вычисленном числе степеней свободы df . Например, при уровне $\alpha = 0{,}05$ для $df = 1$ $\chi^2 = 3.84$.

Шаг 4. Сравниваем фактическое и табличное значения χ^2 . Если $\chi^2_{ф} \geq \chi^2_{df,\alpha}$, то гипотеза H_0 на данном уровне значимости α может быть отвергнута. То есть можно утверждать, что с вероятностью 1-α статистическая связь между признаками существует.

Вероятность того, что, отвергая H_0 , мы совершаем ошибку (ошибка первого рода) численно равна уровню значимости α , задаваемому при проверке гипотезы.

В социально-экономических исследованиях часто принимают уровень значимости α = 0,05 (α = 5%). Таблица значений χ^2 при α = 0,05 для различных степенях свободы *df* приведена в Приложении 2.

Меры связи для таблиц $r \times c$

1.Средняя квадратичная сопряженность (коэффициент Фи)

$$\Phi = \sqrt{\frac{\chi^2}{n}}, \quad 0 \leq \Phi^2 \leq \min(r-1, c-1).$$

2. Коэффициент Крамера $C = \sqrt{C^2}$,

где величина

$$C^2 = \frac{\chi^2}{n \min(r-1, c-1)} = \frac{\Phi^2}{\min(r-1, c-1)},$$

верхним пределом которой является 1, не зависимо от того, равны ли r и c .

3. Коэффициент Чупрова $T = \sqrt{T^2}$, где

$$T^2 = \frac{\chi^2}{n\sqrt{(r-1)(c-1)}} = \frac{\Phi^2}{\sqrt{(r-1)(c-1)}}$$

.

Верхним пределом изменения коэффициента Чупрова является 1, которая достигается, однако, при полной связи между переменными только в том случае, если $r = c$. Во всех других случаях, даже при полной связи $T^2 < 1$.

Анализ связи между порядковыми переменными

В порядковой шкале результатом измерения является приписывание каждому объекту некоторой *условной числовой метки*, обозначающей место этого объекта в ряду из всех n анализируемых объектов, упорядоченном по возрастанию (убыванию) степени проявления в них k - го изучаемого свойства.

Под *ранговой корреляцией* понимается статистическая связь между порядковыми переменными.

Методы анализа ранговых корреляций часто используются в экспертных обследованиях для оценки согласованности мнений экспертов и построения интегральной (совокупной) оценки признака.

*Случай двух ранжировок (экспертов***)**

Пусть 2 эксперта проранжировали *n* объектов по какому-либо свойству, то есть пусть заданы две ранжировки

$$x^{(1)} = (x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)})$$

$$x^{(2)} = (x_1^{(2)}, x_2^{(2)}, \cdots, x_n^{(2)}),$$

где n – число объектов;

$x_i^{(j)}$ – ранг, присвоенный *i*-му объекту *j*- м экспертом по какому-либо признаку.

$i = 1,2,\ldots,n;\ \ j=1,2.$

Ранговый коэффициент корреляции Спирмена.

$$\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(x_i^{(1)} - x_i^{(2)})^2 . \qquad (2.2.2)$$

Очевидно, для совпадающих ранжировок $\rho = 1$, для противоположных $\rho = -1$. Таким образом, $-1 \le \rho \le 1$.

Формула (2.2.2) пригодна для случая отсутствия объединенных рангов. В общем случае,

$$R = \frac{\frac{1}{6}(n^3 - n) - \sum_{i=1}^{n}(x_i^{(1)} - x_i^{(2)})^2 - T^{(1)} - T^{(2)}}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T^{(1)}\right]\left[\frac{1}{6}(n^3 - n) - 2T^{(2)}\right]}}, \qquad (2.2.3)$$

где $T^{(l)} = \frac{1}{12}\sum_{t=1}^{m^{(l)}}\left[(n_t^l)^3 - n_t^l\right]$. (2.2.4)

Здесь $m^{(l)}$ - число групп неразличимых рангов, $n_t^{(l)}$ - число элементов (рангов), входящих в t -ю группу неразличимых рангов.

Ранговый коэффициент корреляции Кендалла определяется по формуле

$$\tau = 1 - \frac{4v(x^{(1)}, x^{(2)})}{n(n-1)}, \qquad (2.2.5)$$

где $v(x^{(1)}, x^{(2)})$ - минимальное число обменов соседних элементов последовательности $x^{(2)}$, необходимое для приведения ее к упорядочению $x^{(1)}$.

При совпадающих ранжировках $\tau = 1$, для противоположных $\tau = -1$. Таким образом, $-1 \le \tau \le 1$.

Если существуют связки – объединенные ранги, то коэффициент будет иметь вид:

$$T = \frac{\tau - \frac{2(U^{(1)} + U^{(2)})}{n(n-1)}}{\sqrt{\left(1 - \frac{2U^{(1)}}{n(n-1)}\right)\left(1 - \frac{2U^{(2)}}{n(n-1)}\right)}}, \qquad (2.2.6)$$

при котором τ_{kj} вычисляется по формуле (2.2.6), «поправочные» величины

$$U^{(l)} = \frac{1}{2}\sum_{t=1}^{m^{(l)}} n_t^{(l)}(n_t^{(l)} - 1), \quad l = 1,2, \qquad (2.2.8)$$

где $m^{(l)}$ и $n_t^{(l)}$ - тот же смысл, что и в формуле выше (2.2.4).

Случай двух и более ранжировок (экспертов)

При решении задач часто необходимо измерить связь между *несколькими* (более чем двумя) переменными (экспертами). С этой целью Кендаллом предложен *коэффициент конкордации* (согласованности):

$$W(m)=\frac{12}{m^2(n^3-n)}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}x_i^{(j)}-\frac{m(n+1)}{2}\right)^2, \qquad (2.2.9)$$

где m - число порядковых переменных (число ранжировок или число экспертов);

n - число объектов или длина ранжировки (объем выборки);

j=1,2,...,m- номера отобранных для анализа порядковых переменных (или номера экспертов).

Коэффициент конкордации обладает свойствами:

а) $0 \le W \le 1$;

б) $W=1$ при совпадении всех m ранжировок.

Если имеются объединенные ранги, то формула (2.2.9) должна быть модифицирована.

$$W(m)=\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m}x_i^{(j)}-\frac{m(n+1)}{2}\right)^2}{\frac{1}{12}m^2(n^3-n)-m\sum_{j=1}^{m}T^{(j)}},$$

где $T^{(j)}$ - как в (2.2.4).

Показано, что при $n>7$ величина $m(n-1)\cdot W(m)$ распределена приближенно по χ^2 распределению при отсутствии связи. Если окажется, что $m(n-1)\cdot W(m)>\chi^2_{\alpha(n-1)}$, то гипотеза об отсутствии связи должна быть отвергнута с уровнем α. Это означает, что коэффициент конкордации значим.

Варианты заданий и порядок выполнения

Исходные данные приведены в Приложении 3.

Задание 1.В Табл.1 Приложения 3 приведены данные 50 респондентов о предпочитаемых напитках. При этом данные закодированы следующим образом: первая переменная ПОЛ (1- мужской, 2-женский), НАПИТОК (1-pepsi, 2-cola).

Порядок выполнения работы

1) В соответствии с вариантом отобрать из исходных данных Табл.1 Приложения 3 для анализа 20 респондентов (строк).

№ варианта	1	2	3	4	5
Строки	2-31	5-34	10-39	15-44	20-49
№ варианта	**6**	**7**	**8**	**9**	**10**
Строки	3-32	6-35	11-40	16-45	21-50

2) Построить четырехклеточную таблицу вида

ПОЛ	НАПИТОК	
	1-pepsi	2-cola
Мужской (1)	***a***	***b***
Женский (2)	***c***	***d***

3) Для удобства подсчета частот провести сортировку данных по первой переменной, а затем по второй.

4) Провести одномерный анализ данных: рассчитать
   - **a)** частоты (число мужчин, число женщин, число любителей пепси, число любителей колы),
   - **b)** относительные частоты по каждому признаку,
   - **c)** построить диаграммы различных видов (гистограммы, секторные; линейные и др.

5) Провести анализ таблицы 2x2:
   - **a)** установить **наличие** связи,
   - **b)** рассчитать коэффициент ассоциации
   - **c)** рассчитать коэффициент контингенции

6) Объяснить полученные результаты

Задание 2. В Табл.2 Приложения 3 приведены данные социологического опроса студентов. Приведены ответы 30 респондентов на 11 вопросов.

Текст вопросов и Кодировка ответов:

1.С КАКИМ НАСТРОЕНИЕМ ВЫ СМОТРИТЕ В БУДУЩЕЕ(*1-оптимистично, 2-спокойно 3-пессимистично)*

2. НУЖНО ЛИ МОЛОДЫМ ЛЮДЯМ ЗНАТЬ ИСТОРИЮ СВОИХ ПРЕДКОВ? *(1-обязательно, 2-необязательно)*

3 КАК СКЛАДЫВАЮТСЯ ВЗАИМООТНОШЕНИЯ СО СВЕРСНИКАМИ- ЮНОШАМИ *(1-легко, 2-спокойно 3-трудно)*

4 КАК СКЛАДЫВАЮТСЯ ВЗАИМООТНОШЕНИЯ СО СВЕРСНИКАМИ- ДЕВУШКАМИ *(1-легко, 2-спокойно 3-трудно)*

5. ХОТЕЛИ БЫ ВЫ ИМЕТЬ ЛИЧНОЕ ОГНЕСТРЕЛЬНОЕ ОРУЖИЕ?(*1-да, 0-нет)*

6 НАСКОЛЬКО ВЫ СОГЛАСНЫ, ЧТО ЖИЗНЕННЫЙ УСПЕХ – ЭТО ВЫСОКОЕ СЛУЖЕБНОЕ ПОЛОЖЕНИЕ, КАРЬЕРА *(1-нет, 2-скорее нет 3-скорее да 4 - да)*
7 НАСКОЛЬКО ВЫ СОГЛАСНЫ, ЧТО ЖИЗНЕННЫЙ УСПЕХ – ЭТО БОГАТСТВО, ФИНАНСОВОЕ БЛАГОСОСТОЯНИЕ *(1-нет, 2-скорее нет 3-скорее да 4 - да)*
8 НАСКОЛЬКО ВЫ СОГЛАСНЫ, ЧТО ЖИЗНЕННЫЙ УСПЕХ – ЭТО Удачный Брак (*1-скорее «нет», 2- скорее «да»)*
9. ПОЛ (*1-м, 2-ж*)
10. ВОЗРАСТ
11. ВУЗ

Порядок выполнения работы

1)В соответствии с вариантом отобрать для анализа 2 вопроса (столбца).

№ варианта	1	2	3	4	5	6	7	8	9	10
Вопрос разреза	9	10	11	9	10	11	9	10	11	9
Вопрос анализа	1	2	6	4	5	7	8	3	2	8

2) Провести сортировку данных по первой переменной (вопрос разреза), а затем по второй (вопрос анализа).
3) Построить таблицу сопряженности $r \times c$.
4) Провести одномерный анализ данных: рассчитать
    a. частоты по вопросу разреза;
    b. частоты по вопросу анализа;
    c. относительные частоты по каждому признаку;
    d. построить диаграммы распределения респондентов по значениям признаков (гистограммы, секторные; линейные и др.)
5) Провести анализ таблицы $r \times c$:
    a. рассчитать значение Хи-квадрат;
    b. установить наличие связи,
    c. рассчитать коэффициент Фи;
    d. рассчитать коэффициент Крамера;
    e. рассчитать коэффициент Чупрова.
6) Дать содержательную интерпретацию полученным результатам.

Задание 3 Рассчитать оценку согласованности мнений экспертов согласно индивидуальному заданию. Индивидуальное задание (задачи) выдает преподаватель непосредственно на занятии. Пример типового задания приведен в Приложении 4.

Порядок выполнения задания

- Повторить теоретические положения.
- Определить возможные оценки согласованности (коэффициенты Спирмена, Кендалла, конкордации).
- Вычислите средний рейтинг объекта.
- Дать содержательную интерпретацию полученным результатам.

Контрольные вопросы и задания

1) Что такое «дихотомические» признаки? Приведите пример.
2) Какие показатели характеризуют силу связи в четырехклеточных таблицах?
3) Если связь между признаками отсутствует, какое значение принимает коэффициент ассоциации?
4) Как определить теоретическое значение Хи-квадрат критерия?
5) В каком случае значения коэффициентов Крамера и Чупрова совпадают?
6) Какие коэффициенты следует вычислять, если исследуется связь двух переменных?
7) Можно ли «доверять» средней оценке рейтинга?

2.3 Лабораторная работа «Построение и анализ типологии объектов»

Цель работы

Закрепление теоретического материала и получение практических навыков реализации методов группировок при решении задач типологизации. Анализ типологии и взаимосвязи количественных признаков.

Форма проведения

Выполнение индивидуального задания.

Форма отчетности

Устный опрос, демонстрация расчетов, выполнение дополнительных заданий.

Теоретические основы

Группировка проводится с целью установления статистических связей и закономерностей, построения описания объекта, выявления структуры изучаемой совокупности. В зависимости от размерности признакового пространства (числа признаков) можно выделить простые (моноте-

тические) и сложные (комбинационные, многомерные, политетические) группировки. Основной принцип группировки – различия между объектами, отнесенными к одной группе, должны быть меньше, чем между единицами, отнесенными к разным группам.

Типологическая группировка служит для выделения социально-экономических типов. Последовательность ее построения следующая:

1) называются те типы явлений, которые могут быть выделены;

2) выбирается группировочный признак, формирующий описание типов;

3) устанавливаются границы интервалов группировочного признака;

4) группировка оформляется в таблицу, определяется численность каждой группы, рассчитываются сводные показатели по группам (групповые средние, показатели вариации).

Оценка качества группировки делается на основе вычисления коэффициента детерминации R^2, характеризующего долю межгрупповой дисперсии в полной.

Коэффициент детерминации R^2 вычисляется по формуле:

Полная дисперсия признака вычисляется по несгруппированным данным по всей совокупности по формуле:

$$R^2 = \frac{\sigma^2_{м.гр.}}{\sigma^2}.$$

Здесь полная дисперсия признака:

$$\sigma^2 = \frac{\sum_{l=1}^{n}(x_l - \bar{x})^2}{n},$$

где n – число объектов в совокупности

x_l – значение признака у l-го объекта, $l=1,2,\ldots, n$

$\bar{x}$ — среднее значение признака в совокупности.

Обозначим $\overline{x_j}$ — среднее значение признака в группе j; f_j — число наблюдений в группе j.

Межгрупповая дисперсия вычисляется по формуле

$$\sigma^2_{м.гр.} = \frac{\sum(\overline{x_j} - \bar{x})^2 f_j}{\sum f_j},,$$

Коэффициент детерминации изменяется от 0 до 1. Если значение R^2 близко к 1, то группировка построена «правильно».

Аналитическая группировка характеризует взаимосвязь между двумя и более признаками, один из которых рассматривается как результат, другой (другие) – как фактор (факторы). Группировка строится *по признаку - фактору,* а оценивается по признаку-результату.

Задача состоит в том, чтобы увидеть, есть связь между признаками или нет; прямая связь или обратная; линейная или нелинейная.

Если среднее значение результата изменяется от группы к группе, то связь между признаками есть. Причем, если при увеличении фактора значение результата увеличивается, то связь прямая.

Проводится сопоставление изменения средних значений результата с изменениями фактора. Чтобы эти изменения были сравнимыми надо делать группировку с равными интервалами или рассчитывать изменения результата на единицу изменения фактора. Рассчитаем величины

$$b_{xy} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1};\ b_{xy} = \frac{\bar{y}_3 - \bar{y}_2}{\bar{x}_3 - \bar{x}_2} \text{ и т.д}$$

Полученные значения показывают величину изменения результата на единицу изменения фактора. Величина b_{xy} равна тангенсу угла наклона отрезка прямой к оси *x*. Если $b_{xy} \neq const$, то связь нелинейная. b_{xy} - показатели силы связи, характеризует прирост результата на единицу изменения фактора.

Для оценки силы связи проводится расчет коэффициента детерминации R^2 по результативному признаку и эмпирического корреляционного отношения *r*.

$$r = \sqrt{R^2} = \sqrt{\frac{\delta^2_{м.гр.}}{\sigma^2}}.$$

Коэффициент детерминации изменяется от 0 до 1. Если значение R^2 близко к 1, то связь между результативным и факторным признаком существует.

Эмпирическое корреляционное соотношение варьирует от -1 до 1.

При $r = 0$ связи нет, при $r = 1$ — связь прямая полная, $r = -1$ — связь обратная полная.

Для исследования взаимосвязи переменных, измеренных в метрических шкалах, применяется *коэффициент корреляции Пирсона (r-Пирсона).*

Корреля́ция (от лат. correlatio «соотношение, взаимосвязь») *или корреляционная зависимость* — это статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения

значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Коэффициент корреляции r-Пирсона характеризует существование ***линейной связи*** между двумя величинами. Если связь криволинейная, то он не будет работать.

Формула расчет коэффициента корреляции Пирсона следующая:

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

Если r_{xy} =0, то связь отсутствует; если r_{xy} = 1, то связь – функциональная; если < 0, то связь – обратная; $r_{xy} > 0$, то связь – прямая.

Для качественной оценки показателей тесноты связи часто применяется шкала Чеддока:

Количественная мера тесноты связи	*Качественная характеристика силы связи*
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая
0,9 - 0,99	Весьма высокая

Методы многомерной классификации (группировки) позволяют проводить разбиение совокупности на основе множества признаков. В общей постановке задача классификации объектов заключается в том, чтобы некоторую совокупность n объектов, статистически представленную в виде матрицы X, разбить на сравнительно небольшое число k (заранее известно или нет) однородных в определенном смысле групп (классов, типов, кластеров, таксонов).

Одним из методов многомерной классификации является кластер-анализ (англ. The cluster – группа, пучок, куст, т.е. объединение каких-то однородных объектов, явлений).

Каждый объект является точкой в признаковом пространстве, которое представляет собой область варьирования всех признаков совокупности изучаемых явлений. Расстояния между точками определяет «схожесть» объектов: чем ближе точки, тем более похожи (однородны) объекты по своим характеристикам. В качестве расстояния будем рассматривать евклидово расстояние:

$$d_E(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ki} - x_{kj})^2} ;$$

Принцип работы *иерархических агломеративных процедур* состоит в последовательном объединении групп элементов сначала самых близких, а затем все более отдаленных друг от друга.

В агломеративно – иерархических алгоритмах процесс объединения объектов в группы совершается последовательно за $n-1$ шагов (если объединяются все n объектов).

На первом шаге в матрице расстояний (различий) D находится *минимальный элемент* d_{ij} и объекты i и j объединяются в один кластер $i+j$, состоящий из двух единиц – объектов. После этого матрица различий изменяется. Из нее выбрасываются две строки и два столбца, содержащие расстояния от i и j до остальных объектов, но добавляется одна строка и один столбец с расстоянием от кластера $i+j$ до остальных объектов. При пересчете матрицы расстояние $d_{i+j,k}$ между объединенным кластером $(i+j)$ и любым из остальных кластеров k вычисляются по определенному правилу, которое определяет *алгоритм.* В социально-экономических исследованиях применяются также алгоритмы минимальной (одной) связи или «ближайшего соседа»: «дальнего соседа», среднего связывания, Варда, центроидной и другие иерархические алгоритмы, реализованные в статистических пакетах.

Далее, на каждом шаге процедура повторяется, т.е. находится минимальный элемент в матрице расстояний и соответствующие кластеры объединяются в один. Итогом работы алгоритма является *иерархическое дерево (дендрограмма*), отражающая последовательность создания вариантов кластеризации на *n, (n-1), ..., 2, 1* групп.

Последовательные процедуры кластер-анализа рассмотрим на примере метода k - средних. В отличие от иерархических алгоритмов в последовательных процедурах на каждом шаге обрабатываются одно наблюдение.

Пусть наблюдения $X_1, X_2, \ldots, X_n$ надо разбить на k $(k << n)$ классов, однородных в смысле некоторой метрики.

Смысл алгоритма состоит в последовательном уточнении эталонных точек – центров классов $E^{(\nu)} = \{e_1^{(\nu)}, e_2^{(\nu)}, \ldots, e_k^{(\nu)}\}$, ν - номер итерации. При этом эталонным точкам приписываются «веса» $\Omega^{(\nu)} = \{\omega_1^{(\nu)}, \omega_2^{(\nu)}, \ldots, \omega_k^{(\nu)}\}$, которые пересчитываются на каждом шаге.

Реализация алгоритма происходит в два этапа. *На первом этапе* пересчитываются эталонные точки, на *втором этапе* производится разбиение объектов на k классов по числу эталонных точек.

Этап 1.

В качестве *нулевого* приближения примем первые k точек (объектов) исходной совокупности:

$e_i^{(0)} = X_i$, $\omega_i^{(0)} = 1$, $i = \overline{1,k}$.

На первом шаге «извлекается» точка X_{k+1} и выясняется, к какому из эталонов она ближе, то есть рассчитываются расстояние от точки X_{k+1} до каждого эталона. Этот ближайший эталон заменяется новым эталоном – центром тяжести старого и присоединенной точки – с увеличением веса, а остальные эталоны не изменяются. Затем появляется следующая точка, и опять выясняется, к какому из эталонов она ближе и т.д.

На ν -ом шаге извлекается $X_{k+\nu}$ и алгоритм пересчета эталонных точек следующий

$$e_i^{(\nu)} = \begin{cases} \dfrac{\omega_i^{(\nu-1)} e_i^{(\nu-1)} + X_{k+\nu}}{\omega^{(\nu-1)} + 1}, \textit{если } d(X_{k+\nu}, e_i^{(\nu-1)}) = \min\limits_{1 \le j \le k} d(X_{k+\nu}, e_j^{(\nu-1)}), \\ e_i^{(\nu-1)}, \quad \textit{иначе.} \end{cases}$$

$$\omega_i^{(\nu)} = \begin{cases} \omega_i^{(\nu-1)} + 1, \textit{если } d(X_{k+\nu}, e_i^{(\nu-1)}) = \min\limits_{1 \le j \le k} d(X_{k+\nu}, e_j^{(\nu-1)}), \\ \omega_i^{(\nu-1)}, \quad \textit{иначе.} \end{cases}$$

$i = \overline{1,k}$.

Максимальное число итераций - $n - k$. Пересчет эталонов заканчивается, если задано число итераций, либо когда эталоны перестают «колебаться», то есть $\max d(e_k^{\nu}, e_k^{\nu-1}) \le \varepsilon$.

Этап 2.

Процесс разбиения исходной совокупности объектов на классы следующий:

извлекается точка X_i $(i = \overline{1,n})$ и вычисляются расстояния от нее до всех e_j $(j = \overline{1,k})$ эталонов. Если $d(X_i, e_S) = \min\limits_{1 \le j \le k} d(X_i, e_j)$, то точка

X_i включается в класс, образованный эталоном e_S. В результате последовательного просмотра все точки все точки будут разбиты на заданное число классов. Результатом является алфавит классификации, то есть списки объектов, входящих в каждый кластер.

Порядок выполнения работы и варианты заданий

Задание 1.

1) В соответствии с номером варианта определить номера двух показателей субъектов федерации Сибирского федерального округа, представленные в таблице «Основные социально-экономические показатели в 2016 г.» Приложения 5

Вариант	*X*- признак-фактор (номер показателя)	*У*- признак-результат (ВРП)
1.	4	1
2.	5	1
3.	6	1
4.	7	1
5.	8	1
6.	9	1
7.	10	1
8.	11	1
9.	12	1
10.	13	1
11.	14	1
12.	15	1

2) Сформировать рабочую таблицу, содержащую названия регионов и указанные в варианте показатели социально-экономического развития регионов СФО.

Название региона	Показатель (фактор)- ***X***	ВРП (результат) ***У***
1.		
…	…	…
12.		

2) Провести сортировку регионов по значению **фактора – *x***
3) Провести по всей совокупности *для каждого признака* расчет
   - среднего значения (функция СРЗНАЧ)
   - дисперсии (функция ДИСПР)
   - стандартного отклонения (функции СТАНДОТКЛОНП, КОРЕНЬ)
   - коэффициента вариации

4) Построить точечную диаграмму зависимости результата от фактора.
5) Провести группировку районов (городов) по *значению* ***фактора*****,** выделив 3 группы: «Малые», «Средние», «Крупные». Границы группировочного показателя задать самостоятельно.
   Для каждой группы определить и занести в табл.1:
   - частоту группы,
   - групповые средние значения показателей ***x*** и ***y***
   - групповые дисперсии показателя ***y***,
   - групповые коэффициенты вариации показателей ***x*** и ***y*** .
6) Вычислить по факторному признаку межгрупповую дисперсию по формуле.
7) Вычислить по факторному признаку коэффициент детерминации и сделать вывод о качестве построенной группировки

Таблица 1. Статистические характеристики группировки

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

8) Рассчитать величины b_{yx}.. и сделать вывод о линейности связи.
9) Рассчитать межгрупповую дисперсию по признаку-**результату**
10) Рассчитать коэффициент детерминации R^2 для группировки по признаку-**результату** и сделать вывод о качестве группировки и силе связи между признаками
11) Вычислить значение эмпирического корреляционного отношения по признаку – результату;
12) Рассчитать коэффициент корреляции r_{xy} (функция КОРРЕЛ или ПИРСОН)
14) Сделать общие выводы о качестве группировок и силе связи между признаками.

Задание 2.

1) В соответствии с номером варианта определить номера таблицы исходных данных, номера двух показателей субъектов федерации Сибирского федерального округа, представленные в таблицах Приложения 6, и исследуемый иерархический алгоритм.

Вариант	*Номер таблицы в Приложении 9*	*x*-признак-фактор (номер показателя)	*У*-признак-результат-ВРП	*Иерархический алгоритм*
1.	1	2	1	«ближайшего соседа»
2.	2	2	1	«дальнего соседа»
3.	1	3	1	«медианной связи»
4.	2	3	1	«ближайшего соседа»
5.	1	4	1	«дальнего соседа»
6.	2	4	1	«медианной связи»
7.	1	5	1	«ближайшего соседа»
8.	2	5	1	«дальнего соседа»
9.	1	6	1	«медианной связи»
10.	2	6	1	«ближайшего соседа»
11.	1	7	1	«дальнего соседа»
12.	2	7	1	«медианной связи»

2) Сформировать рабочую таблицу, содержащую названия регионов и указанные в варианте показатели социально-экономического развития регионов СФО, следующего вида

Название региона	Показатель (фактор) *X*	ВРП (результат) *У*
1.		
…	…	…
5.		

5) Провести сортировку регионов по значению **фактора** – ***x***

6) Провести по всей совокупности *для каждого признака* расчет
   - среднего значения (функция СРЗНАЧ)
   - дисперсии (функция ДИСПР)
   - стандартного отклонения (функции СТАНДОТКЛОНП или КОРЕНЬ из дисперсии)
   - коэффициента вариации.

5) Построить точечную диаграмму в пространстве 2-х признаков.

6) Провести кластерный анализ, используя указанный в варианте *иерархический алгоритм*. Определить вариант разбиения **на 3 класса** и результаты разбиения записать в табл.2. Построить дендрограмму.
7) Провести кластерный анализ **на 3 класса**, используя метод **К-средних.** Эталонные точка задать самостоятельно. Сделать максимальное число итераций**.** Результаты разбиения записать в табл.2.

Таблица 2.Варианты разбиения регионов на классы

	Названия регионов, входящих в классы		
Метод	Класс № 1	Класс №2	Класс №3
Иерархический метод			
k-средних			

8) Сравнить результаты, полученные двумя алгоритмами кластер-анализа.

9) По результатам метода К-средних:

a. внести в рабочую таблицу дополнительный столбец с номером кластера и провести сортировку по номеру кластера;
b. построить точечную диаграмму в пространстве двух признаков с изображением принадлежности регионов кластерам (выделить разным цветом);
c. рассчитать групповые средние значения по классификационным признакам и записать в табл.3;
d. рассчитать коэффициенты детерминации по классификационным признакам и записать в табл.3.

Таблица 3 Показатели в группах

Признак	Средние значения признаков в кластерах			Коэффициент детерминации
	№ 1	№ 2	№ 3	
Название признака-1(x)				R_x^2
Название признака-2 (y)				R_y^2

10) Дать содержательную интерпретацию результатов кластер-анализа (оценить: уровень развития регионов по классификационным признакам, наличие «естественного расслоения», различие средних значений, однородность групп, взаимосвязь признаков)

Контрольные вопросы и задания.

1) Сформулируйте основной принцип группировки.
2) Какой показатель оценивает однородность совокупности объектов?
3) Можно ли назвать совокупность регионов однородной по исследуемым признакам?

4) Объясните выбор значений границ группировочного признака-фактора.
5) Запишите формулу вычисления межгрупповой дисперсии.
6) Можно ли признать построенную группировку «правильной»?
7) Охарактеризуйте влияние признака-фактора на ВРП.
8) Объясните, почему коэффициент детерминации в аналитической группировке характеризует силу связи между признаками?
9) Охарактеризуйте группы, полученные в результате применения алгоритма к-средних.
10) Сравните результаты типологической группировки и кластер-анализа.

2.4 Лабораторная работа «Анализ и прогнозирование временных рядов»

Цель работы

Закрепление теоретического материала и получение практических навыков вычисления показателей динамики. Построение моделей временного ряда.

Форма проведения

Выполнение индивидуального задания.

Форма отчетности

Устный опрос, демонстрация расчетов, выполнение дополнительных заданий.

Теоретические основы

Временной (динамический) ряд - ряд расположенных в хронологической последовательности значений статистических показателей. Каждый временной ряд включает два элемента: момент или период времени и конкретное значение показателя (уровень ряда). Уровни ряда обычно обозначают латинской буквой ***y***, а моменты или периоды времени, к которым они относятся, - буквой ***t***.

Пусть *n* — число уровней ряда и нумерация ряда начинается с 1 (единицы).

Базисный абсолютный прирост $\Delta y_{\text{б}}$ исчисляется как разность между сравниваемым уровнем y_i и уровнем, принятым за постоянную базу сравнения y_1:

$$\Delta y_{\text{бi}} = y_i - y_1 .$$

Цепной абсолютный прирост $\Delta y_{\text{ц}}$ — разность между сравниваемым уровнем y_i и уровнем, который ему предшествует y_{i-1}:

$$\Delta y_{\text{ці}} = y_i - y_{i-1}$$

Темп роста базисный (в процентах) определяется по формуле

$$\text{Тр}_{\text{б}_i} = (y_i : y_1) \cdot 100$$

Темп роста цепной (в процентах) определяется по формуле

$$\text{Тр}_{\text{ц}_i} = (y_i : y_{i-1}) \cdot 100 .$$

Темп прироста базисный (в процентах) определяется по формуле

$$\text{ТПр}_{\text{бі}} = \text{Тр}_{\text{бі}} - 100$$

Средний уровень ряда ($\overline{y}$) динамики характеризует типическую величину абсолютных уровней. Метод расчета среднего уровня ряда динамики зависит от вида временного ряда.

Для *интервального* временного ряда абсолютных показателей с равными периодами времени средний уровень ряда $\overline{y}$ рассчитывается по формуле простой арифметической:

$$\overline{y} = \frac{\sum y_i}{n} = \frac{y_1 + y_2 + \ldots + y_n}{n} .$$

где *n* — число уровней ряда.

В *моментном* ряду динамики с равностоящими датами времени средний уровень определяется по формуле средней хронологической

$$\overline{y} = \frac{\frac{1}{2} y_1 + y_2 + \ldots + \frac{1}{2} y_n}{n-1} .$$

Показатель *среднего абсолютного прироста* можно определить по формуле

$$\Delta\overline{y} = \frac{y_n - y_1}{n-1} .$$

Средний темп роста можно определить по абсолютным уровням ряда динамики по формуле

$$\overline{\text{Т}}\text{р} = \sqrt[n-1]{y_n : y_1} \times 100\%$$

Для получения *средних темпов прироста* $\overline{\text{Т}}\text{п}$ в процентах используется зависимость:

$$\overline{\text{Т}}\text{п} = \overline{\text{Т}}\text{р} - 100 .$$

Прогнозирование на следующий (n+1) период времени осуществляется по формуле:

$$\hat{y}_{n+1} = y_n + \overline{\Delta y},$$

При этом предполагается, что период упреждения составляет 1-2 временных интервала, в течение которых тенденция развития сохраняется.

Динамический ряд теоретически может быть представлен в виде совокупности *трех составляющих*:

1) *тренд* — основная тенденция развития динамического ряда (тенденция к росту или к снижению);

2) циклические (периодические) колебания, в том числе сезонные;

3) случайные колебания.

На практике для непосредственного выявления и изучения тренда в рядах динамики используются три основных метода:

метод укрупнения интервалов;

метод скользящей средней;

метод аналитического выравнивания.

Метод укрупнения интервалов заключается в том, что исходные уровни ряда заменяются средними значениями, вычисленными на более длинных временных интервалах. Например, переходим от помесячных данных к поквартальным или от годичных данных к пятилетним и т.д.

В методе *трехзвенной скользящей средней* сглаженные уровня ряда вычисляются последовательно по формуле

$$\overline{y}_i = \frac{y_{i-1} + y_i + y_{i+1}}{3};$$

При ***аналитическом выравнивании*** ряда динамики фактический уровень изучаемого показателя оценивается как функция времени (трендовая модель, уравнение регрессии)

$$y = f(t) + \varepsilon,$$

где $f(t) = \hat{y}$ — уровень, определяемый тенденцией развития;

ε — случайное или циклическое отклонение от тенденции.

Подбор адекватной функции осуществляется методом наименьших квадратов — минимальностью отклонений суммы квадратов между теоретическими $\hat{y}_i$ и эмпирическими y_i уровнями:

$$\sum (\hat{y}_i - y_i)^2 = \min .$$

В простейшем случае динамический ряд характеризуется *равномерным развитием*. Для этого типа динамики характерны постоянные цепные абсолютные приросты:

$$\Delta y_i = const .$$

Основная тенденция развития в рядах динамики со стабильными приростами отображается уравнением линейной функции:

$$\hat{y}_t = a_0 + a_1 t ,$$

где a_0 и a_1 — параметры уравнения;

t — обозначение времени

При вычисления параметров функции на основе требований метода наименьших квадратов получаются следующие формулы расчета коэффициентов

$$a_0 = \frac{\sum y \sum t^2 - \sum ty \sum t}{n\sum t^2 - \sum t \sum t};$$

$$a_1 = \frac{n\sum ty - \sum t \sum y}{n\sum t^2 - \sum t \sum t}.$$

Построив уравнение регрессии, проводят оценку его надежности. Это делается посредством критерия Фишера (F). Фактический уровень ($F_{\text{факт}}$) сравнивается с теоретическим (табличным) значением:

$$F_{\text{факт}} = \frac{\sigma_{\text{f}}^2 (n-k)}{\sigma_{\text{ост}}^2 (k-1)},$$

где k — число параметров функции, описывающей тенденцию;

n — число уровней ряда.

σ_{f}^2 - факторная дисперсия, которая вычисляется по формуле

$$\sigma_{\text{f}}^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{n}$$

Остаточная дисперсия $\sigma_{\text{ост}}^2$ определяется по формуле

$$\sigma_{\text{ост}}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n}.$$

По правилу сложения дисперсий общая дисперсия

$$\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \sigma_{\text{f}}^2 + \sigma_{\text{ост}}^2 .$$

$F_{\text{факт}}$ сравнивается с $F_{\text{теор}}$ при $\nu_1 = (k-1)$, $\nu_2 = (n-k)$ степенях свободы и уровне значимости α (обычно $\alpha = 0{,}05$). Таблица теоретических значений приведена в Приложении 7. Если $F_{\text{факт}} > F_{\text{теор}}$, то уравнение регрессии значимо, т.е. основная модель адекватна фактической временной тенденции.

Для оценки точности модели вычисляют коэффициент детерминации:

$$R^2 = \frac{\sigma_{\text{f}}^2}{\sigma_{\text{y}}^2}, \quad 0 \le R^2 \le 1 .$$

Если значение коэффициента детерминации R^2 близко к 1, то модель близка к реальному процессу.

Подставляя в полученное уравнение модели значение времени $t_k (k \notin i = 1,2...,n)$, получаем **точечный прогноз (оценку прогнозного)** значения

$$\hat{y}_k = a_0 + a_1 \cdot t_k .$$

Порядок выполнения работы и варианты заданий

Исходные данные о прибывших в Россию из стран дальнего зарубежья представлены в таблице Приложения 8.

В соответствии с номером варианта определить из таблицы исходных данных страну

№ варианта	Страна	№ варианта	Страна
Вариант 1	Австралия	Вариант 6	США
Вариант 2	Австрия	Вариант 7	Турция
Вариант 3	Вьетнам	Вариант 8	Финляндия
Вариант 4	Германия	Вариант 9	Франция
Вариант 5	Греция	Вариант 10	Израиль

Порядок выполнения работы

Задание 1.

1) Вычислить
   - базисные и цепные абсолютные приросты,

- базисные и цепные темпы роста,
- базисные и цепные темпы прироста мигрантов

2) Вычислить средние показатели
   - среднегодовую численность мигрантов
   - среднегодовой абсолютный прирост,
   - среднегодовой темп роста,
   - среднегодовой темп прироста

Задание 2

1) Вычислить прогнозное значение на следующий 2017 год, используя средний абсолютный прирост

Задание 3

1) Провести выравнивание временного ряда методами
   - укрупнения интервалов (перейти к трехгодичным периодам);
   - трехзвенной скользящей средней.

2) Построить графики исходных и выровненных значений

Задание 4

1) построить линейную модель для выбранных данных, вычисляя коэффициенты ***a0*** и ***a1***
2) провести оценку модели по критерию Фишера,
3) рассчитать коэффициент детерминации.
4) построить график динамики исходных и выровненных значений.
5) вычислить прогнозное значение на следующий год.

Задание 5

Пользуясь средствами MS Excel (Приложение 2), провести исследование различного вида моделей тренда.

1) Для каждого случая на графике исходных данных добавить линию тренда и поместить на график уравнение тренда и значение коэффициента детерминации.
2) Уравнение тренда и значение коэффициента детерминации занести в таблицу 1, макет которой приведен ниже.
3) Сделать прогноз на следующие периоды (2 периода) для разных трендов.
4) В строку 7 поместить уравнение и коэффициент детерминации, рассчитанные самостоятельно в Задании 2.
5) Сравнить уравнения, полученные самостоятельно (строка 7) и средствами MS Excel (строка 1). Сделать вывод о совпадении/не совпадении уравнений.
6) Определить по Таблице, какая модель является наиболее точной.

Таблица 1

№	Наименование модели	Уравнение тренда	Коэффициент детерминации
1.	Линейная		
2.	Полиномиальная 2-й степени		
3.	Полиномиальная 3-й степени		
4.	Логарифмическая		
5.	Степенная		
6.	Экспоненциальная		
7.	Линейная *(самостоятельно рассчитанная)*		

Контрольные вопросы и задания.

1) Дайте определения показателей динамики.
2) Объясните выбор формул для расчета показателей динамики.
3) Укажите на графике абсолютные цепные и базисные приросты
4) Объясните результаты построения линейных моделей.
5) Какая из полученных средствами MS Excel моделей точнее?
6) Запишите формулу пятизвенной скользящей средней
7) Каким показателем оценивается точность модели?.

.

3 Методические указания к организации самостоятельной работы

3.1 Общие положения

Цель самостоятельной работы по дисциплине – закрепление и углубление теоретических знаний; формирование умения работать с научной и технической литературой и осуществлять самостоятельный поиск информации; развитие научно-исследовательских и творческих способностей; приобретение навыков расчётно-аналитической работы.

Самостоятельная работа студента по дисциплине «Анализ данных» включает следующие виды его активности:

1. проработка лекционного материала;
2. изучение тем теоретической части дисциплины, вынесенных для самостоятельной проработки;
3. выполнение домашнего задания;
4. выполнение индивидуального задания;
5. подготовка к контрольным работам;
6. подготовка к лабораторным работам;
7. подготовка к экзамену.

3.2 Проработка лекционного материала

При проработке лекционного материала по каждой теме студент должен внимательно ознакомиться с конспектом лекций, а затем для углубленного изучения материала следует обратиться к литературным источникам (учебникам, учебным пособиям, монографиям, статьям, статистическим сборникам), а также материалам, размещенным в сети Интернет. Для закрепления материала темы необходимо ответить на предлагаемые в пособиях вопросы и прорешать задачи по теме.

При изучении каждой темы целесообразно:

1) ознакомиться с методическим обеспечением изучаемой дисциплины, включающей тематический план и программу курса;

2) руководствоваться рекомендованной нормативной базой и учебной литературой, которая имеется в фондах библиотеки;

3) использовать возможности сайта библиотеки университета и другие информационные ресурсы Интернета;

4) прочитать соответствующую теме главу учебника;

5) доработать конспект лекции.

При изучении учебного материала темы студенту необходимо, прежде всего, разобраться в основанных понятиях и терминах данной темы. Для этого рекомендуется использовать различные источники информации, в том числе учебные пособия, монографии, периодические издания, статистические материалы, а также труды зарубежных авторов.

Изучение рекомендованной литературы следует начинать с основных рекомендованных преподавателем глав и разделов учебников и учебных пособий, а затем переходить к нормативно-правовым актам, научным монографиям и материалам периодических изданий. При этом полезно делать выписки и конспекты наиболее интересных материалов, что способствует более глубокому осмыслению материала и лучшему его запоминанию. Такая практика учит отделять в тексте главное от второстепенного, а также позволяет проводить систематизацию и сравнительный анализ изучаемой информации, что важно в условиях большого количества разнообразных по качеству и содержанию сведений.

Проработка пройденного лекционного материала Проработка пройденного лекционного материала является наиболее важным видом самостоятельной работы. Чем глубже и полнее проработан материал, тем легче при выполнении других видов самостоятельной работы. Систематическая, регулярная работа над пройденным лекционным материалом, начиная с первого занятия, является необходимым условием для понимания материалов последующих лекций и усвоения материалов практических и лабораторных занятий.

3.3 Самостоятельное изучение тем теоретической части курса

3.3.1 Тема: Факторный анализ как метод снижения размерности

Перечень вопросов, подлежащих изучению

1. Сущность факторного анализа как метода снижения размерности.
2. Основная модель факторного анализа.
3. Проблемы факторного анализа и схема решения задач.
4. Основные методы факторного анализа.

Методические рекомендации по изучению

Факторный анализ - это метод многомерного статистического анализа, позволяющий на основе экспериментального наблюдения признаков объекта выделить группу переменных, определяющих корреляционную взаимосвязь между признаками.

При изучении темы студенту необходимо повторить математические понятия матрицы, обратной матрицы, собственные числа и векторы, а

также понятия статистической связи, определения ковариации и корреляции, основные положения регрессионного анализа.

Важно понимать, что с помощью метода факторного анализа можно решить четыре основные задачи:

1) отыскание скрытых, но объективно существующих закономерностей, которые определяются воздействием внутренних и внешних причин на изучаемый процесс;

2) сжатие информации путем описания процесса при помощи общих факторов, число которых значительно меньше количества первоначально взятых признаков;

3) выявление и изучение статистической связи признаков с факторами;

4) прогнозирование хода развития процесса на основе уравнения регрессии; уравнения регрессии, построенные при помощи результатов, полученных в факторном анализе, обладают значительными преимуществами перед классическим регрессионным анализом.

Студент должен уяснить основное предположение факторного анализа, которое заключается в том, что корреляционные связи между большим числом наблюдаемых переменных определяются существованием меньшего числа гипотетических наблюдаемых переменных или факторов.

Целью метода факторного анализа является представление величины z_{ij}, то есть элемента матрицы Z, в виде линейной комбинации нескольких факторов (гипотетических величин). Таким образом, значение z_{ij} может быть выражено в виде линейной комбинации r факторов

$$z_{ij} = a_{i1} \cdot p_{1j} + a_{i2} \cdot p_{2j} + ... + a_{ir} \cdot p_{rj},$$

где z_{ij} - стандартизированное значение *i*-ой переменной для *j*-го индивидуума, то есть

$$z_{ij} = \frac{y_{ij} - \bar{y}_i}{\sigma_i}$$

a_{il} - вычисленные факторные нагрузки (постоянные коэффициенты, которые следует определить);

$p_{1j} - p_{rj}$ - значения фактора r у *j*-го объекта.

Это равенство отражает основную модель факторного анализа. В матричном виде модель для всех z_{ij} запишется

$Z = A \cdot P,$

где Z является матрицей порядка $m \times n$ стандартизированных переменных - исходных данных. $A = (a_{il})$ является матрицей порядка $m \times r$, которую можно определить. Она называется *факторным отображением*, а ее коэффициенты - *факторными нагрузками*. $P = (p_{lj})$ - матрицей порядка $r \times n$ значений всех факторов у всех объектов.

Таким образом, матрица Z представляет собой произведение двух матриц: A и P. При этом матрица A отражает связи переменных с факторами, а P описывает отдельные объекты.

Умножив обе части на A^{-1} (обратная матрица), можно методом множественного регрессионного анализа получить значения факторов для каждого объекта

$P = A^{-1}Z$

Студенту следует изучить основные проблемы факторного анализа (проблему общности, проблему факторов, проблему вращения, проблему значений факторов) и идею методов их решения.

Вычислительные процедуры, отражающие содержание этих методов, реализованы в стандартных программах которые входят в большинство пакетов статистического анализа данных.

Рекомендуемые источники

Методы и средства комплексного анализа данных [Электронный ресурс] /Кулаичев А.П., 4-е изд., перераб. и доп. - М.: НИЦ ИНФРА-М, 2016. – с.315-349 с. — Режим доступа: http://znanium.com/catalog/product/548836

При необходимости рекомендуется ознакомиться с другими источниками, приведенными в разделе «Рекомендуемые источники».

3.3.2 Тема: Основные положения дискриминантного анализа

Перечень вопросов, подлежащих изучению

1) Основные понятия и сущность дискриминантного анализа.
2) Типы задач, решаемы методами дискриминантного анализа .
3) Основные методы дискриминантного анализа.

Методические рекомендации по изучению

Дискримина́нтный ана́лиз представляет набор методов статистического анализа для решения задач распознавания образов. Он используется для принятия решения о том, какие переменные разделяют (т.е. «дис-

криминируют») возникающие наборы данных (так называемые «группы»). В отличие от кластерного анализа в дискриминантном анализе группы известны априори.

Студенту необходимо понять, что основная дискриминантного анализа состоит 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.

Методы дискриминантного анализа связаны с получением одной или нескольких функций, обеспечивающих возможность отнесения данного объекта к одной из групп. Эти функции называются дискриминантными (классифицирующими) и зависят от значений переменных таким образом, что появляется возможность отнести каждый объект к одной из групп. Вид получаемой дискриминантной функции не отличается от уравнения регрессии:

$$D = b_0 + b_1x_1 + b_2x_2 + .. + b_kx_k..$$

В качестве зависимой переменной выступает номинальная переменная, идентифицирующая принадлежность объектов к одной из нескольких групп. Независимые переменные ($x_1, x_2 .. x_k$) могут быть количественные и качественные.

Основной задачей дискриминантного анализа является исследование групповых различий - различие (дискриминация) объектов по определенным признакам. Дискриминантный анализ позволяет выяснить, действительно ли группы различаются между собой, и если да, то каким образом (какие переменные вносят наибольший вклад в имеющиеся различия).

Студенту рекомендуется изучить идею методов определения дискриминантных переменных (переменных, входящих в дискриминантную функцию). Например, пошаговый (*stepwise*) дискриминантный анализ, при котором переменные вводятся последовательно, исходя из их способности различить (дискриминировать) группы.

Студент должен иметь в виду, что вычислительные процедуры, отражающие содержание этих методов, реализованы в стандартных программах которые входят в большинство пакетов статистического анализа данных.

Рекомендуемые источники

Методы и средства комплексного анализа данных [Электронный ресурс] /Кулаичев А.П., 4-е изд., перераб. и доп. - М.: НИЦ ИНФРА-М, 2016. – с.365-370 с. — Режим доступа:
http://znanium.com/catalog/product/548836

При необходимости рекомендуется ознакомиться с другими источниками, приведенными в разделе «Рекомендуемые источники».

3.4 Домашнее задание

Тема: Анализ периодической составляющей временного ряда

Цель домашнего задания

Получение практических навыков построения и анализа модели временного ряда при наличии периодических колебаний

Порядок выполнения и содержание работ

Если в анализируемой временной последовательности наблюдаются устойчивые отклонения от тенденции (как в большую, так и в меньшую сторону), то можно предположить *наличие* в ряду динамики некоторых (одного или нескольких) *колебательных процессов*, то есть *циклические (периодические) колебания*

Моделирование временного ряда при наличии периодической составляющей осуществляется с помощью гармонического анализа.

Для анализа внутригодовой динамики социально-экономических явлений могут применяться ***гармоники ряда Фурье***.

При аналитическом выражении изменений уровней ряда динамики используется формула

$$\hat{y}_t = a_0 + \sum (a_k \cos kt + b_k \sin kt), \qquad (3.4.1)$$

где k определяет номер гармоники, который используется с различной степенью точности (обычно от 1 до 4).

При решении уравнения (3.4.1) параметры определяются на основе положений метода наименьших квадратов, в результате получают систему нормальных уравнений, параметры которых вычисляются по формулам:

$$a_0 = \frac{\sum y_i}{n}; \qquad (3.4.2)$$

$$a_k = \frac{2}{n} \sum y_i \cos kt_i; \qquad (3.4.3)$$

$$b_k = \frac{2}{n} \sum y_i \sin kt_i. \qquad (3.4.4)$$

При анализе ряда внутригодовой динамики по месяцам значение k принимается за 12. Подставляя месячные периода как части окружности, ряд внутригодовой динамики можно записать в таком виде:

Периоды (t_i)	**Уровни (y_i)**
0	y_1
$\frac{1}{6}\pi$	y_2
$\frac{1}{3}\pi$	y_3
$\frac{1}{2}\pi$	y_4
$\frac{2}{3}\pi$	y_5
$\frac{5}{6}\pi$	y_6
π	y_7
$\frac{7}{6}\pi$	y_8
$\frac{4}{3}\pi$	y_9
$\frac{3}{2}\pi$	y_{10}
$\frac{5}{3}\pi$	y_{11}
$\frac{11}{6}\pi$	y_{12}

Если применяется одна **первая гармоника** ряда Фурье *(k=1)*, то параметры в формулах (3.4.2), (3.4.3) и (3.4.4) принимают вид:

$$a_0 = \frac{\sum y_i}{n}$$

$$a_1 = \frac{2}{n}\sum y_i \cos t_i$$

$$b_1 = \frac{2}{n}\sum y_i \sin t_i$$

По полученным параметрам синтезируется математическая модель в виде:

$$\hat{y}_t = a_0 + a_1 \cos t + b_1 \sin t \qquad (3.4.5)$$

Правильность вычислений проверяется: $\sum y = \sum \hat{y}$.

Для оценки точности модели вычисляют коэффициент детерминации R^2.

При выполнении домашнего задания студенту необходимо изучить теорию, выполнить задание, представить в письменном виде результаты расчетов, уметь ответить на вопросы преподавателя:

1. Перечислите составляющие временного ряда.
2. Приведите примеры процессов, имеющих сезонную составляющую.
3. Какой метод используется для определения параметров модели? определяются параметров модели?
4. Запишите общий вид модели с использованием гармоник.
5. Запишите вид модели с первой гармоникой.
6. Объясните полученные результаты.

Индивидуальное Задание и исходные данные студенту выдает преподаватель. Пример выполнения задания приведен в Приложении 10

Рекомендуемые источники

Методы и средства комплексного анализа данных [Электронный ресурс] /Кулаичев А.П., 4-е изд., перераб. и доп. - М.: НИЦ ИНФРА-М, 2016. – с.315-349. — Режим доступа: http://znanium.com/catalog/product/548836

Лепихина З.П. Статистика: Учебное пособие/ З. П. Лепихина; Федеральное агентство по образованию, Томский государственный университет систем управления и радиоэлектроники. - Томск: ТУСУР, 2005. – с.161-167.

При необходимости рекомендуется ознакомиться с другими источниками, приведенными в разделе «Рекомендуемые источники».

3.4. Индивидуальное задание «Инструментальные средства статистического анализа данных»

Цель индивидуального задания

Изучение и анализ рынка программных продуктов по статистическому анализу данных.

Исходные данные к работе

Рынок программных продуктов по статистическому анализу данных. Вид исследования — «кабинетное». Студенту необходимо сделать обзор программных средств анализа данных на основе информации из различных источников, в том числе в глобальных компьютерных сетях.

Порядок выполнения и содержание работы

Рекомендуется начать работу с изучения официальных статистических данных о состоянии данной сферы. Например, ознакомиться с разделом 19. «ИНФОРМАЦИОННЫЕ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ» в статистических сборниках «Регионы России. Социально-экономические показатели» на официальном сайте Федеральной службы государственной статистики. Далее следует изучить статьи в компьютерных журналах — периодических изданиях, основной темой которых являются информационные технологии, программное и аппаратное обеспечение. Задача студента — составить общее представление о спектре имеющихся на рынке программных продуктов. По согласованию с преподавателем студент отбирает для сравнительного анализа 3-4 продукта и проводит углубленное исследование, используя специальные журналы и интернет.

Объём обзора, как правило, от 7 до 10 машинописных страниц. Студент разрабатывает и оформляет отчет в соответствии с требованиями ОС ТУСУР 02-2013. Перед началом работы над рефератом следует наметить план и подобрать литературу. Прежде всего, следует пользоваться литературой, рекомендованной учебной программой, а затем расширить список источников, включая и использование специальных журналов и Интернет.

Рекомендуется следующая структура отчета:

- Титульный лист.
- Оглавление.
- Введение (дается постановка вопроса, объясняется значимость и актуальность темы, указываются цель и задачи работы, даётся характеристика используемой литературы).

- Основная часть (может состоять из разделов, которые раскрывают отдельную проблему или одну из её сторон и логически являются продолжением друг друга).
- Заключение (подводятся итоги и даются обобщённые основные выводы по теме работы, делаются рекомендации).
- Список использованных источников. В списке должно быть не менее 6-8 различных источников.

Приветствуется включение таблиц, графиков, схем, как в основном тексте, так и в качестве приложений.

Содержание основной части отчета должно разрабатываться в направлениях:

- рассмотрение понятий «программная система», «инструментальные средства», «пакет прикладных программ», «информационная технология» и.т.д., в том числе в нормативных документах;
- классификация рассматриваемых программных продуктов;
- выбор показателей для сравнения программных продуктов (функционал, стоимость и т.д.);
- сравнительный анализ рассматриваемых программных продуктов по выбранным показателям (критериям).

При изучении источников студенту следует обратить внимание на возможности использования программного продукта для решения практических задач, степень открытости кода программ, репутацию разработчика.

При оценке работы учитываются: глубина проработки материала; правильность и полнота использования источников; владение терминологией и культурой речи; оформление отчета.

Рекомендуемые источники

Компьютерные технологии анализа данных в эконометрике [Электронный ресурс]/ Д.М. Дайитбегов. - 2-е изд., испр. и доп. - М.: Вузовский учебник: ИНФРА-М, 2010. – с511-530: — Режим доступа: http://znanium.com/catalog/product/251791

При необходимости рекомендуется ознакомиться с другими источниками, приведенными в разделе «Рекомендуемые источники».

3.6 Подготовка к контрольным работам

Контрольная работа – одна из форм проверки и оценки усвоенных знаний, получения информации о характере познавательной деятельности, уровня самостоятельности и активности студентов в учебном процессе.

При подготовке к выполнению контрольной работы необходимо повторить теоретический материал по теме, основные формулы и методы решения задач на данную тему. Следует вновь просмотреть примеры и задачи, разобранные в учебниках, на лекции и при выполнении лабораторных работ.

Важно понять, что если студент систематически работает над пройденным материалом, начиная с первой лекции, то подготовка к контрольной работе не вызовет затруднений и много времени на нее не понадобится.

3.7 Подготовка к лабораторным работам

Лабораторные занятия являются связующим звеном теории и практики. Они позволяют углубить и закрепить теоретические знания, получаемые на лекциях, проверить теоретические положения экспериментальным путем, выработать у студентов практические умения и навыки работы с реальной статистической информацией. Одновременно они являются базой для аналитической исследовательской работы студентов.

Содержание лабораторных работ и порядок выполнения определены в разделе 2 настоящих указаний. Следует помнить, что в начале методических указаний на выполнение каждой лабораторной работы приводится краткое изложение теоретических положений, поэтому студент должен заранее самостоятельно подготовиться к лабораторной работе с использованием указанной преподавателем литературы: Подготовить ответы на контрольные вопросы, предложенные преподавателем к данной лабораторной работе.

Каждая лабораторная работа выполняется по определенной теме с указанием цели её выполнения. Студенту необходимо уяснить цель работы и при подготовке к работе, при выполнении работы и анализе результатов следовать ей.

4 Рекомендуемые источники

1) Методы и средства комплексного анализа данных [Электронный ресурс] /Кулаичев А.П., 4-е изд., перераб. и доп. - М.: НИЦ ИНФРА-М, 2016. – с.365-370 с. — Режим доступа: http://znanium.com/catalog/product/548836
2) Годин, А.М. Статистика [Электронный ресурс] : учебник / А.М. Годин. — Электрон. дан. — Москва : Дашков и К, 2017. — 412 с. — Режим доступа: https://e.lanbook.com/book/93468.
3) Лепихина З.П. Статистика: Учебное пособие/ З. П. Лепихина; Федеральное агентство по образованию, Томский государственный университет систем управления и радиоэлектроники. - Томск: ТУСУР, 2005. – 284 с.
4) Лепихина, З.П. Основы социального прогнозирования: Учебное пособие/ З. П. Лепихина; Федеральное агентство по образованию, Томский государственный университет систем управления и радиоэлектроники, Кафедра автоматизации обработки информации. - Томск: ТМЦДО, 2006. – 112 с.
5) Статистический анализ данных в MS Excel: Учебное пособие [Электронный ресурс] / А.Ю. Козлов, В.С. Мхитарян, В.Ф. Шишов. - М.: ИНФРА-М, 2012. - 320с. — Режим доступа: http://znanium.com/catalog/product/238654
6) Компьютерные технологии анализа данных в эконометрике [Электронный ресурс] / Д.М. Дайитбегов. - 2-е изд., испр. и доп. - М.: Вузовский учебник: ИНФРА-М, 2010. - 578 с.: — Режим доступа: http://znanium.com/catalog/product/251791
7) Статистические методы анализа данных: Учебник [Электронный ресурс] / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга [и др.]; под общ. ред. д-ра экон. наук, проф. Л.И. Ниворожкиной. — М.: РИОР: ИНФРА-М, 2016. — 333 с. — Режим доступа: http://znanium.com/catalog/product/556760
8) Регионы России. Социально-экономические показатели. 2017: [Электронный ресурс] : Р32 Стат. сб. / Росстат. – М., 2017. – 1402 с. — Режим доступа: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/publications/catalog/doc_1138623506156

ПРИЛОЖЕНИЕ 1

Таблица Международная миграция (человек)

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Прибыло в Российскую Федерацию										
Азербайджан	4600	8900	20968	23331	22874	14500	22316	22287	23453	26323
Армения	7581	12949	30751	35216	35753	19890	32747	36978	42361	46515
Беларусь	6797	5619	6030	5865	5517	4894	10182	16564	15748	17878
Грузия	51945	38606	40258	39964	38830	27862	36474	45506	51958	59096
Казахстан	15592	15669	24731	24014	23265	20901	41562	34597	30388	28539
Киргизия	6569	8649	14090	15519	16433	11814	19578	23594	28666	32030
Молдова	4717	6523	17309	20717	27028	18188	35087	41674	51011	54636
Таджикистан	4104	4089	4846	3962	3336	2283	4524	5442	5986	6033
Туркмения	30436	37126	52802	43518	42539	24100	64493	87902	118130	130906
Узбекистан	30760	32721	51492	49064	45920	27508	43586	49411	55037	115524
Украина	4600	8900	20968	23331	22874	14500	22316	22287	23453	26323
из стран дальнего зарубежья	7581	12949	30751	35216	35753	19890	32747	36978	42361	46515

ПРИЛОЖЕНИЕ 2

Таблица - Теоретические значения χ^2 при $\alpha = 0,05$

df	1	2	3	4	5	6	7	8	9	10
χ^2	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92	18,31

ПРИЛОЖЕНИЕ 3

Таблица -Данные 50 респондентов о предпочитаемых напитках

№ респон-дента	ПОЛ	НАПИ-ТОК	№ респон-дента	ПОЛ	НАПИ-ТОК	№ респон-дента	ПОЛ	НАПИ-ТОК
1	1	1	**18**	2	1	**35**	1	2
2	1	2	**19**	2	2	**36**	1	2
3	2	2	**20**	1	2	**37**	2	1
4	1	1	**21**	1	1	**38**	2	1
5	1	1	**22**	2	1	**39**	2	1
6	2	2	**23**	1	1	**40**	1	2
7	2	2	**24**	2	2	**41**	1	1
8	1	1	**25**	1	2	**42**	1	2
9	2	1	**26**	1	2	**43**	2	1
10	1	1	**27**	1	1	**44**	1	1
11	2	1	**28**	2	1	**45**	1	2
12	1	2	**29**	2	1	**46**	1	1
13	1	1	**30**	1	2	**47**	2	2
14	1	2	**31**	1	1	**48**	2	1
15	2	1	**32**	2	2	**49**	2	2
16	2	1	**33**	2	1	**50**	1	2
17	1	2	**34**	1	2			

Таблица 2 - Ответы 30 респондентов на 11 вопросов

	№ вопроса										
№ анкеты	**1**	**2**	**3.**	**4.**	**5**	**6.**	**7.**	**8**	**9**	**10**	**11**
1	1	1	1	1	1	4	3	4	1	19	ТПУ
2	1	2	2	2	1	4	4	4	1	20	ТПУ
3	2	0	1	2	0	4	3	4	1	22	ТПУ
4	2	0	2	1	0	3	3	4	1	19	ТПУ
5	1	1	2	2	1	3	3	3	1	19	ТПУ
6	1	1	2	3	1	1	4	1	1	20	ТПУ
7	1	2	3	3	1	4	4	2	1	19	ТУСУР
8	1	2	1	1	0	2	3	1	1	19	ТУСУР
9	3	1	2	2	1	3	3	4	1	20	ТПУ
10	2	2	1	1	1	4	3	4	1	19	ТПУ
11	1	1	2	2	1	4	4	4	2	20	ТПУ
12	1	1	1	2	1	1	1	2	1	18	ТПУ
13	1	1	1	1	0	4	4	4	1	19	ТПУ
14	1	0	1	3	0	2	4	4	2	19	ТПУ
15	2	2	2	2	0	4	4	2	2	19	ТУСУР
16	1	2	2	1	1	2	2	2	1	19	ТУСУР
17	1	2	1	1	1	3	3	4	2	20	ТУСУР
18	3	2	2	3	1	4	3	4	2	20	ТУСУР

Продолжение таблицы 2

	№ вопроса										
№ анкеты	**1**	**2**	**3.**	**4.**	**5**	**6.**	**7.**	**8**	**9**	**10**	**11**
19	1	2	2	2	0	4	4	4	2	22	ТПУ
20	1	2	2	2	1	4	4	4	2	20	ТУСУР
21	0	2	2	2	1	1	1	1	2	20	ТУСУР
22	3	2	3	2	1	4	4	4	2	20	ТУСУР
23	3	2	1	2	0	4	4	4	2	19	ТУСУР
24	1	1	1	1	0	4	4	4	2	19	ТПУ
25	1	2	1	1	0	4	4	4	2	19	ТПУ
26	1	1	1	1	1	4	4	4	2	20	ТПУ
27	2	2	2	2	1	4	4	4	2	19	ТПУ
28	1	1	2	2	1	3	3	4	2	20	ТПУ
29	3	2	2	2	0	4	4	4	2	20	ТУСУР
30	1	1	1	2	1	4	4	4	2	22	ТУСУР

ПРИЛОЖЕНИЕ 4

Пример типового задания к лабораторной работе 2 по расчету оценки согласованности мнений экспертов.

а) Два эксперта проранжировали 10 предприятий с точки зрения конкурентоспособности.

Эксперт 1	1	2	3	4	5	6	7	8	9	10
Эксперт 2	2	3	1	4	6	5	9	7	8	10

б) Десять однородных предприятий проранжированы по степени

- прогрессивности оргструктур - $x^{(1)}$;

- по эффективности $x^{(2)}$.

$x^{(1)}$	1	2,5	2,5	4,5	4,5	6,5	6,5	8	9,5	9,5
$x^{(2)}$	1	2	4,5	4,5	4,5	4,5	8	8	8	10

в) Три эксперта Э1, Э2и Э3 упорядочили 10 объектов по конкурентным преимуществам.

	Э1	Э2	Э3
1	1	2,5	2
2	4,5	1	1
3	2	2,5	4,5
4	4,5	4,5	4,5
5	3	4,5	4,5
6	7,5	8	4,5
7	6	9	8
8	9	6,5	8
9	7,5	10	8

ПРИЛОЖЕНИЕ 5

Таблица - Основные показатели развития регионов Сибирского федерального округа

	Валовой региональный продукт в 2015 г. , млн. руб.	Площадь территории[1], тыс. км^2	Численность населения на 1 января 2017 г., тыс. человек	Среднегодовая численность занятых в экономике, тыс. человек	Среднедушевые денежные доходы (в месяц), руб.	Потребительские расходы в среднем на душу населения (в месяц), руб.	Среднемесячная номинальная начисленная заработная плата работников, руб.	Основные фонды в экономике (по полной учетной стоимости; на конец года)[2], млн. руб.
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Республика Алтай	41776,8	92,9	217,0	91,7	13836,9	7179,0	15632,4	61628
Республика Бурятия	204156,2	351,3	984,1	417,4	15715,5	11340,0	19924,0	430210
Республика Тыва	47287,3	168,6	318,6	106,0	10962,8	4944,6	19163,1	47409
Республика Хакасия	171663,9	61,6	537,7	239,2	14222,8	9680,5	20689,5	292915
Алтайский край	492138,9	168,0	2365,7	1075,6	12499,9	9765,7	13822,6	757632
Забайкальский край	248847,6	431,9	1079,0	489,4	15968,8	10572,7	21099,6	650405
Красноярский край	1618166,0	2366,8	2875,3	1437,5	20145,5	14105,7	25658,6	1815754
Иркутская область	1013542,3	774,8	2408,9	1121,7	16017,2	10580,2	22647,7	1975486
Кемеровская область	842618,9	95,7	2708,8	1302,0	16666,0	11237,2	20478,8	1406912
Новосибирская область	980850,5	177,8	2779,5	1305,1	18244,1	14898,1	20308,5	1229181
Омская область	617184,4	141,1	1972,7	945,5	17247,9	12663,1	19087,8	725451
Томская область	473693,1	314,4	1078,9	487,5	16516,0	11199,4	24001,0	863117

Продолжение Таблицы

Объем отгруженных товаров собственного производства, выполненных работ и услуг собственными силами по видам экономической деятельности, млн. руб.			Продукция сельского хозяйства - всего, млн. руб.	Ввод в действие общей площади жилых домов, тыс. $м^2$	Оборот розничной торговли, млн. руб.	Инвестиции в основной капитал, млн. руб.	
добыча полезных ископаемых	обрабатывающие производства	Производство и распределение электроэнергии, газа и воды					
9	**10**	**11**	**12**	**13**	**14**	**15**	
862	1296	1654	8020	76,6	14312	11802	Республика Алтай
12808	51115	19826	13044	304,4	100938	41017	Республика Бурятия
3376	514	2868	4648	52,4	13742	7033	Республика Тыва
26536	56595	21462	9371	156,2	46034	38064	Республика Хакасия
6041	189279	31991	93784	663,2	218077	70833	Алтайский край
40377	14365	18311	15154	276,9	106366	51557	Забайкальский край
266636	628113	95432	68598	1047,1	361607	303885	Красноярский край
129795	299406	81275	43610	755,2	225846	137995	Иркутская область
507993	385413	85949	38044	1082,6	287279	225131	Кемеровская область
19674	249816	61592	60425	1505,2	368292	142078	Новосибирская область
4411	529355	36013	66911	836,7	228595	83342	Омская область
137513	100598	25617	19420	457,6	93050	101927	Томская область

ПРИЛОЖЕНИЕ 6

Таблица 1- Основные показатели развития регионов Сибирского федерального округа (округленные значения)

	Валовой регио-наль-ный продукт в 2014 г. , млрд. руб.	Оборот розничной торговли, млрд. руб.	Добыча полезных ископае-мых млрд. руб.	Инвести-ции в основ-ной капитал, млрд. руб.	Продук-ция сельского хозяйства - всего, млрд. руб.	Средне-месячная номи-наль-ная на-чис-ленная заработ-ная плата работ-ников, тыс.руб.	обраба-ты-вающие произ-водства млрд. руб.
	1	**2**	**3**	**4**	**5**	**6**	**7**
1. Республика Алтай	39	14	1	12	8	15	1
2. Республика Бурятия	185	100	13	41	13	20	51
3. Республика Тыва	46	13	3	7	5	19	1
4. Республика Хакасия	160	46	26	38	9	21	56
5. Алтайский край	448	218	6	71	94	14	189

Таблица 2- Основные показатели развития регионов Сибирского федерального округа (округленные значения)

	Валовой регио-наль-ный продукт в 2014 г. , млрд. руб.	Оборот розничной торговли, млрд. руб.	Добыча полезных ископае-мых млрд. руб.	Инвести-ции в основ-ной капитал, млрд. руб.	Продук-ция сельского хозяйства - всего, млрд. руб.	Средне-месячная номи-наль-ная на-чис-ленная заработ-ная плата работ-ников, тыс.руб.	обраба-ты-вающие произ-водства млрд. руб.
	1	**2**	**3**	**4**	5	**6**	**7**
1. Красноярский край	1423	3617	267	304	68	26	630
2. Иркутская область	907	225	130	138	44	23	300
3. Кемеровская область	747	287	508	225	38	20	380
4. Новосибирская область	895	368	20	142	60	20	250
5. Томская область	428	93	138	102	19	24	100

ПРИЛОЖЕНИЕ 7

Значения F-распределения Фишера

Уровень значимости $\alpha=0,05$

Число степеней свободы $v_1 = k-1=1$, $v_2 = n-k = n-2$,

где k – число параметров функции, описывающей тенденцию (для линейной функции k=2); n – число уровней ряда

$v_2 = n-2$	F_{tab}	$v_2 = n-2$	F_{tab}	$v_2 = n-2$	F_{tab}
1	162,4	11	4,84	21	4,32
2	18,51	12	4,75	22	4,30
3	10,13	13	4,67	23	4,28
4	7,71	14	4,60	24	4,26
5	6,61	15	4,55	25	4,24
6	5,99	16	4,51	26	4,22
7	5,59	17	4,45	27	4,21
8	5,32	18	4,41	28	4,19
9	5,12	19	4,38	29	4,18
10	4,96	20	4,35	30	4,17

ПРИЛОЖЕНИЕ 8

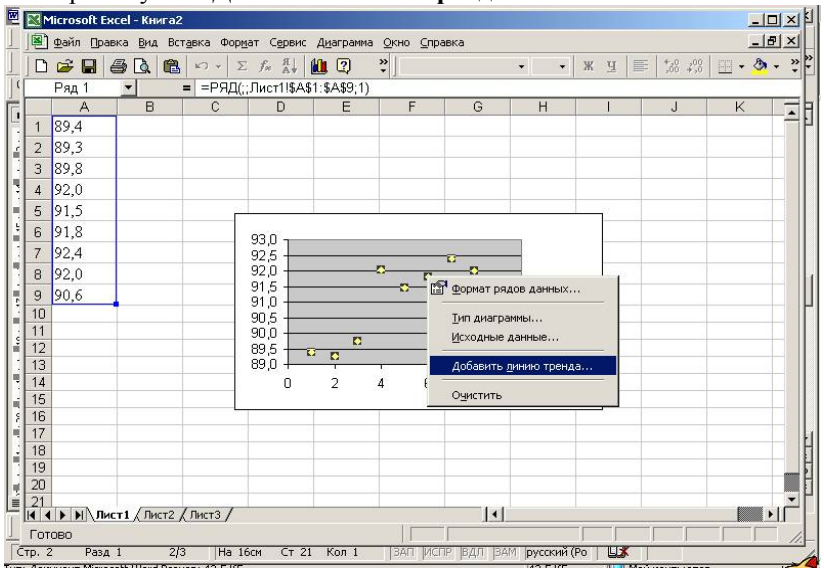
Таблица – Международная миграция (человек)

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Прибыло в РФ – из стран:												
Австралия	30	28	38	31	39	49	83	78	113	71	89	82
Австрия	24	53	50	35	37	45	60	70	84	81	122	83
Вьетнам	114	157	921	714	950	921	3294	3653	3852	3854	4012	3735
Германия	3025	2900	3164	3134	2585	2621	4520	4239	4166	3743	3976	4153
Греция	200	176	260	289	240	298	614	835	995	694	557	450
Израиль	1004	1053	1094	1002	861	814	1240	1091	1132	1139	1077	900
Индия	54	72	107	66	72	110	1390	1068	1451	1850	2894	4768
Канада	99	77	118	105	98	110	192	207	226	171	189	193
Китай	432	499	1687	1177	770	1380	7063	8547	8149	10563	9043	8027
Польша	55	48	96	100	97	105	187	200	217	199	194	181
Великобритания	40	34	100	80	92	125	166	182	221	185	273	226
США	396	411	578	551	575	653	947	1122	954	1000	1084	1137
Турция	86	172	315	373	443	562	1832	2252	2755	2631	2091	1626
Финляндия	129	137	172	174	141	178	266	342	429	468	401	393
Франция	40	54	144	72	96	150	322	326	352	351	360	303

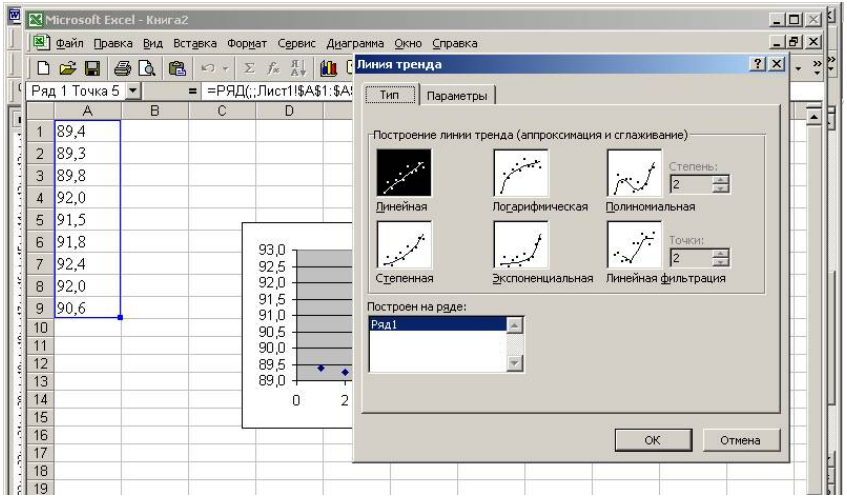
ПРИЛОЖЕНИЕ 9

ПРИМЕР ПОСТРОЕНИЯ ТРЕНДА В MS EXCEL.

1. При помощи средства **Мастер диаграмм** построить график исходных данных.
2. Подвести курсор к графику. Щелчком правой кнопки мыши вызвать выплывающее контекстное меню и выбрать пункт **<Добавить линию тренда>.**

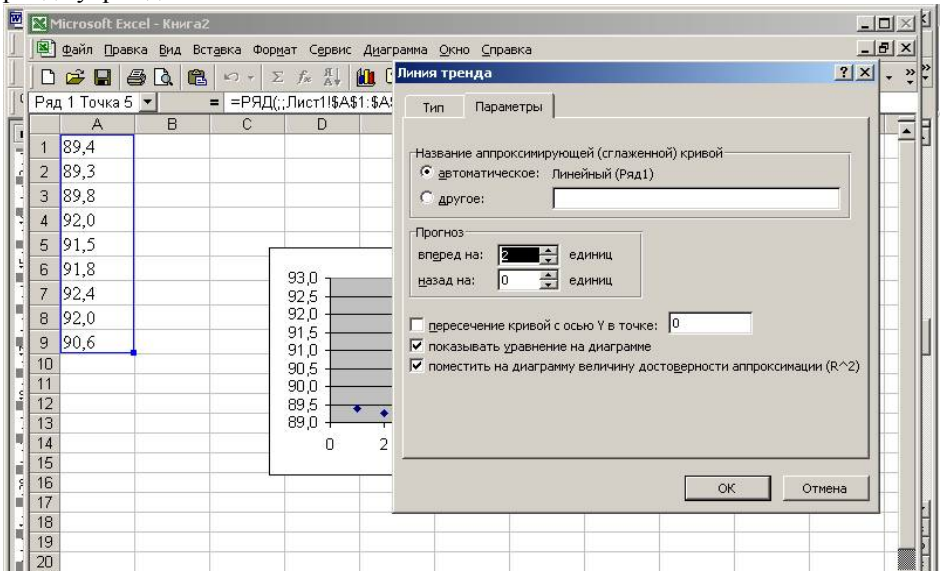


3. В появившемся диалоговом окне <Линия тренда> на вкладке <Тип> выбирается вид функции.

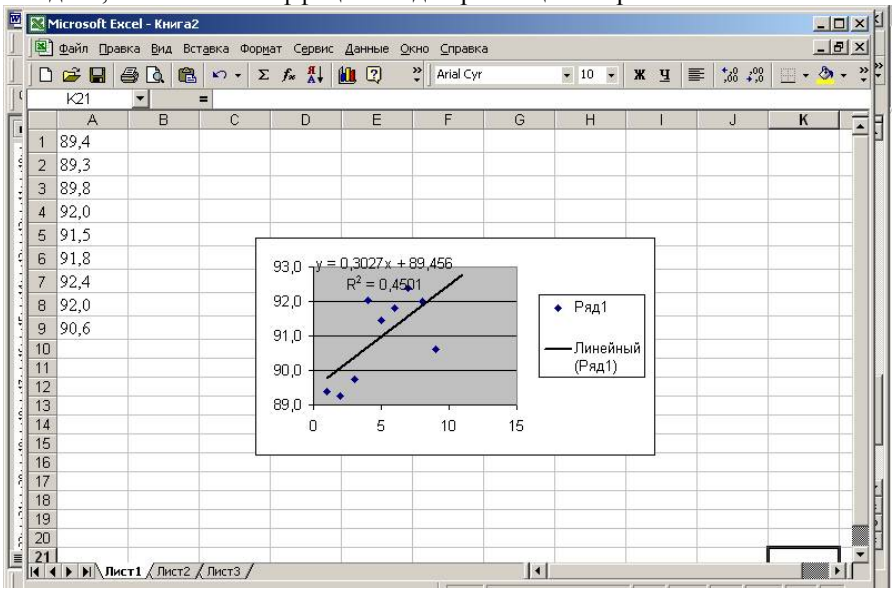


4. На вкладке <Параметры> задаются дополнительные параметры: отмечаем поля «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2)».

5. Для расчета прогнозного значения на вкладке <Параметры> в поле <Прогноз> задается значение периода упреждения.



6. По нажатию кнопки <ОК> получаем график с нанесенными линиями выбранного типа тренда, уравнением модели, значением коэффициента детерминации и прогнозным значением.



ПРИЛОЖЕНИЕ 10

Пример построения модели внутригодовой динамики по первой гармонике ряда Фурье на данных о розничном товарообороте по месяцам (таблица).

Таблица - Розничный товарооборот по месяцам

Месяц	t_i	Объем розничного товарооборота, млрд. руб.	$\cos t_i$	$\sin t_i$	$y_i \cos t_i$ (для формулы 3)	$y_i \sin t_i$ (для формулы 4)	$\hat{y}_{t_i}$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Январь	0	27,3	1,0	0,0	27,3	0,0	30,1
Февраль	(1:6)π	28,0	0,866	0,5	24,2	14,0	29,5
Март	(1:3)π	31,2	0,5	0,866	15,6	27,0	29,2
Апрель	(1:2)π	30,1	0,0	1,0	0,0	30,1	29,2
Май	(2:3)π	29,2	–0,5	0,866	–14,6	25,3	29,6
Июнь	(5:6)π	30,0	–0,866	0,5	–26,6	15,0	30,2
Июль	π	30,1	–1,0	0,0	–30,1	0,0	30,9
Август	(7:6)π	32,0	–0,866	–0,5	–27,7	–16,0	31,5
Сентябрь	(4:3)π	31,4	–0,5	–0,866	–15,7	–27,2	31,8
Октябрь	(3:2)π	32,3	0,0	–1,0	0,0	–32,3	31,8
Ноябрь	(5:3)π	31,2	0,5	–0,866	15,6	–27,0	31,4
Декабрь	(11:6)π	33,5	0,866	–0,5	29,0	–16,7	30,8
	×	366,4	×	×	–2,4	–7,8	366,0

Применяя **первую гармонику** ряда Фурье *(k=1)*, определяются параметры уравнения:

$$a_0 = \frac{\sum y_i}{n}$$

с учетом итогового значения графы 4, подставляем численное значение $a_0 = \frac{366,4}{12} = 30,5$;

$$a_1 = \frac{2}{n}\sum y_i \cos t_i$$

с учетом итогового значения графы 6 подставляем численное значение $a_1 = \frac{2 \cdot (-2,4)}{12} = -0,4$;

$$b_1 = \frac{2}{n}\sum y_i \sin t_i$$

с учетом итогового значения графы 7 подставляем численное значение $b_1 = \frac{2 \cdot (-7,8)}{12} = -1,3$.

По полученным параметрам синтезируется математическая модель:

$$\hat{y}_i = 30,5 - 0,4\cos t_i - 1,3\sin t_i . \qquad (1)$$

На основе модели (1) определяются для каждого месяца расчетные уровни $\hat{y}_t$

для января $\hat{y}_t = 30,5 - 0,4 \cdot 1,0 - 1,3 \cdot 0 = 30,1$ млрд. руб.;

…….

для декабря $\hat{y}_t = 30,5 - 0,4 \cdot 0,866 - 1,3 \cdot (-0,5) = 30,8$ млрд. руб.

Вычисленные для января и декабря теоретические уровни $\hat{y}_t$ записаны в гр. 8 таблицы.