

Министерство образования и науки РФ

Федеральное государственное бюджетное образовательное
учреждение высшего образования

**ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И
РАДИОЭЛЕКТРОНИКИ» (ТУСУР)**

Кафедра автоматизации обработки информации (АОИ)

УТВЕРЖДАЮ
Зав.кафедрой АОИ,
профессор

_____ Ю.П. Ехлаков
" ____ " _____ 2017 г.

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ
ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНЫХ РАБОТ,
К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ
И ОРГАНИЗАЦИИ САМОСТОЯТЕЛЬНОЙ РАБОТЫ
ПО ДИСЦИПЛИНЕ «ПРИКЛАДНАЯ СТАТИСТИКА»
для студентов направления подготовки
«Бизнес-информатика» (уровень бакалавриата)**

Разработчик:
доцент каф. АОИ, к.т.н.
_____ З.П. Лепихина

СОДЕРЖАНИЕ

	Введение.....	3
1.	Методические указания по выполнению лабораторных работ.....	3
2.	Лабораторная работа 1.....	3
3.	Лабораторная работа 2.....	8
4.	Лабораторная работа 3.....	13
5.	Лабораторная работа 4.....	20
6.	Методические указания к практическим занятиям	24
7.	Методические указания по организации самостоятельной работы.....	31
8.	Рекомендуемая литература	36
9.	Приложения.....	37

ВВЕДЕНИЕ

Дисциплина «Прикладная статистика» в соответствии с Государственным образовательным стандартом **относится к базовой части образовательных дисциплин. Цель дисциплины** — дать студентам представление о содержании статистики как научной дисциплины, познакомить с ее основными понятиями, методами получения статистической информации о социально-экономических явлениях и ее обобщения, научить использовать современные методы и технологии обработки данных для решения задач анализа данных и выявления тенденций в развитии социально-экономических процессов.

Целью выполнения лабораторных работ, практических занятий и самостоятельной работы является закрепление и углубление знаний теоретической части дисциплины; выбора соответствующей методики исчисления статистических показателей, приобретение навыков самостоятельного выполнения расчетов и анализа полученных результатов.

В «Методических указаниях» представлены краткое изложение теоретического материала и задания, необходимые для выполнения лабораторного исследования, примеры решения задач из разделов статистики, а также даны рекомендации по изучению материала и задания для самостоятельного выполнения. При самостоятельной работе и подготовке к лекционным, лабораторным и практическим занятиям студенту следует повторить теоретический материал по конспекту лекций, источникам, приведенным в разделе «Рекомендуемая литература».

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНЫХ РАБОТ

ЛАБОРАТОРНАЯ РАБОТА 1

Тема работы.

Создание информационной базы данных социально-экономических процессов. Первичный анализ данных на компьютере в среде Microsoft Excel. Анализ рядов динамики.

Краткие сведения из теории

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 .

Для количественной оценки динамики социально-экономических явлений применяются статистические показатели: абсолютные приросты, темпы роста и прироста.

Базисный абсолютный прирост $\Delta y_{\text{б}}$ исчисляется как разность между сравниваемым уровнем y_i и уровнем, принятым за постоянную базу сравнения y_0 :

$$\Delta y_{\text{б}_i} = y_i - y_0.$$

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

$$\Delta y_{\text{ц}_i} = y_i - y_{i-1}.$$

Темп роста базисный (в процентах) определяется по формуле

$$\text{Тр}_{\text{б}_i} = Kp_{\text{б}_i} \cdot 100 = (y_i : y_0) \cdot 100$$

Темп роста цепной (в процентах) определяется по формуле

$$Tp_{ц_i} = Kp_{ц_i} \cdot 100 = (y_i : y_{i-1}) \cdot 100.$$

Темп прироста базисный (в процентах) определяется по формуле

$$ТПр_{бi} = TP_{бi} - 100.$$

Темп прироста цепной (в процентах) определяется по формуле

$$ТПр_{цi} = TP_{цi} - 100.$$

Средний уровень ряда ($\bar{y}$) 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 $\bar{y}$ рассчитывается по формуле простой арифметической:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{y_1 + y_2 + \dots + y_n}{n}.$$

где n — число уровней ряда.

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

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

Показатель *среднего абсолютного прироста* можно определить по формуле

$$\Delta \bar{y} = \frac{\Delta y_{\bar{y}_n}}{n - 1}.$$

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

$$\bar{Tp} = \sqrt[n]{y_n : y_0} \times 100\%$$

Для получения *средних темпов прироста* $\bar{ТП}$ в процентах используется зависимость:

$$\bar{ТП} = \bar{Tp} - 100.$$

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.

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

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

где $f(t)$ — 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 абсолютные приросты:

$$\Delta y_n = \text{const} .$$

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

$$\hat{y}_t = a_0 + a_1 t , \quad (1.1)$$

где a_0 и a_1 — параметры уравнения;

t — обозначение времени.

Для вычисления параметров функции (1.1) на основе требований метода наименьших квадратов составляется система нормальных уравнений:

$$\begin{aligned} na_0 + a_1 \sum t &= \sum y ; \\ a_0 \sum t + a_1 \sum t^2 &= \sum t \cdot y . \end{aligned} \quad (1.2)$$

Решение системы уравнений (1.2) дает следующие формулы расчета коэффициентов

$$a_0 = \frac{\sum y \sum t^2 - \sum ty \sum t}{n \sum t^2 - \sum t \sum t} ; \quad (1.3)$$

$$a_1 = \frac{n \sum ty - \sum t \sum y}{n \sum t^2 - \sum t \sum t} . \quad (1.4)$$

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

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

где k — число параметров функции, описывающей тенденцию;

n — число уровней ряда.

$\sigma_{\text{факт}}^2$ - факторная дисперсия

$\sigma_{\text{ост}}^2$ - остаточная дисперсия

Остаточная дисперсия определяется по формуле

$$\begin{aligned} \sigma_{\text{ост}}^2 &= \frac{\sum (y_i - \hat{y}_i)^2}{n} , \\ \sigma_{\text{факт}}^2 &= \frac{\sum (\hat{y}_i - \bar{y})^2}{n} , \end{aligned}$$

По правилу сложения дисперсий общая дисперсия

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

$F_{\text{факт}}$ 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 1. 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:

$$R^2 = \frac{\sigma_{\text{факт}}^2}{\sigma_y^2}, \quad 0 \leq R^2 \leq 1.$$

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

Подставляя в полученное уравнение модели значение времени t_k ($k \in i = 1, 2, \dots, n$), получаем **оценку прогнозного** (предсказанного) значения

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

Построение моделей временного ряда в MS Excel

В MS Excel существует специальный аппарат для аппроксимации экспериментальных данных с последующим подбором аппроксимирующей функции (линии тренда) $y=f(t)$.

Возможны следующие варианты функций.

1. Линейная $y=a_0 + a_1t$. Обычно применяется в случаях, когда динамический ряд возрастает или убывает с постоянной скоростью.

2. Полиномиальная $y=a_0 + a_1t + a_2t^2 + \dots + a_nt^n$, где $n \leq 6$. Используется для динамического ряда, попеременно возрастающего и убывающего. Степень полинома определяется количеством экстремумов (максимумов или минимумов) кривой. Полином второй степени может описать только один максимум или минимум, полином третьей степени может иметь один или два экстремума и т.д.

3. Логарифмическая $y=a_0 + a_1 \ln(t)$. Функция применяется для описания динамических рядов, которые вначале быстро растут или убывают, а затем постепенно стабилизируются.

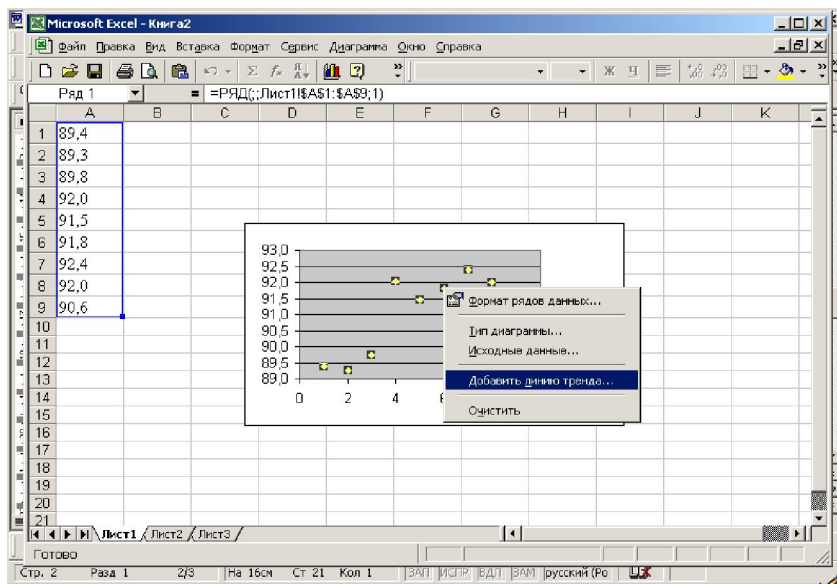
4. Степенная $y=bt^n$. Функция используется для динамических рядов в постоянно увеличивающейся (или убывающей) скоростью роста.

5. Экспоненциальная $y=ae^t$. Применяется для описания динамических рядов, которые вначале быстро растут или убывают, а затем постепенно стабилизируются.

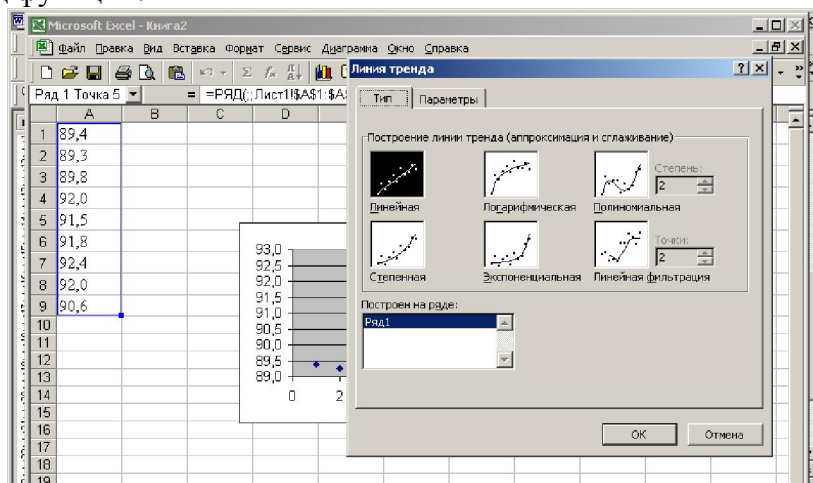
Степень близости аппроксимации экспериментальных данных выбранной функции оценивается коэффициентом детерминации R^2 . По построенной модели тренда MS Excel позволяет проводить точечный прогноз на заданное число шагов. Линиями тренда можно дополнить ряды данных, представленные на линейных диаграммах, гистограммах, графиках, точечных и пузырьковых графиках.

Для осуществления аппроксимации экспериментальных данных (подбора модели, построения линии тренда) необходимо выполнить следующие действия.

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

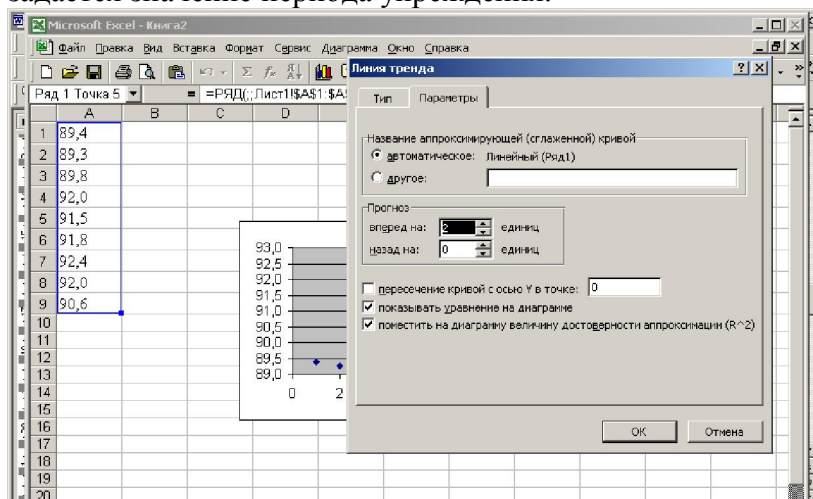


3. В появившемся диалоговом окне <Линия тренда> на вкладке <Тип> выбирается вид функции.

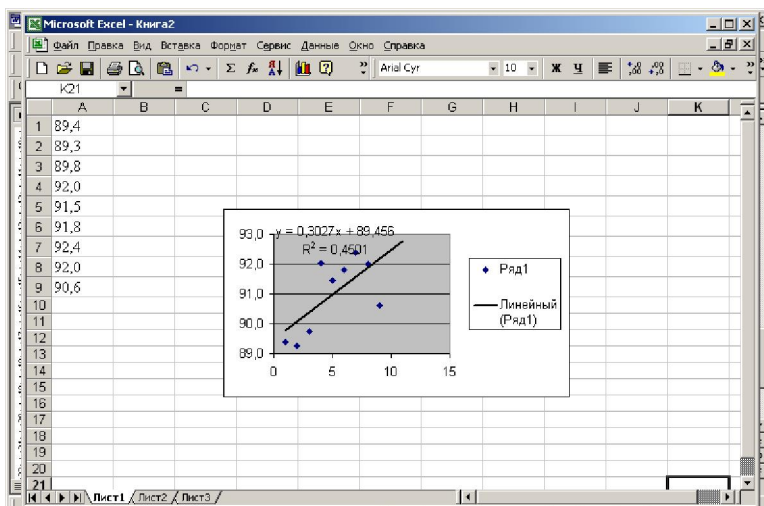


4. На вкладке <Параметры> задаются дополнительные параметры: отмечаем поля «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2)».

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



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



Исходные данные.

В соответствии с номером варианта необходимо выбрать данные из таблицы Приложения 2 исходных данных о динамике численности персонала.

Порядок выполнения работы

- 1) вычислить цепные абсолютные приросты и цепные темпы роста;
- 2) построить линейную модель для выбранных данных
- 3) провести оценку модели по критерию Фишера, рассчитать коэффициент детерминации.
- 4) Построить график динамики исходных и выровненных значений.
- 5) Рассчитать точечный прогноз на следующий период.
- 6) Пользуясь средствами MS Excel, провести исследования различного вида моделей тренда. Рассмотреть полиномы 2-й и т.д. степеней, помещая на график уравнение тренда и значения коэффициента детерминации. Сделать прогноз на следующие периоды для разных трендов.
- 7) Оформить отчет в Word

Вопросы для защиты работы.

- 1) Дать определения рядов динамики.
- 2) Объяснить результаты построения линейных моделей.
- 3) Объяснить возможности применения разных моделей трендов для прогнозирования.

ЛАБОРАТОРНАЯ РАБОТА 2

Тема работы.

Анализ связи между номинальными признаками

Краткие сведения из теории

Связь в таблицах 2x2

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 (четырёхклеточная) 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.

По определению признаки независимы в данной совокупности из n наблюдений, если

$$a = \frac{(a+b)(a+c)}{n} \text{ - частота в предположении независимости.}$$

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 будем говорить, что A и B *отрицательно связаны*.

Меры связи в таблицах 2х2

В 1900 Юл предложил коэффициент связи

$$Ka = \frac{ad - bc}{ad + bc}. \text{ Это коэффициент ассоциации Юла.}$$

Также часто используется коэффициент контингенции:

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

$Ka, Kk = 0$, если признаки независимы, $Ka, Kk = +1$ в случае полной положительной связи, $Ka, Kk = -1$, если полная отрицательная связь между признаками.

В общем случае, когда число градаций номинальных признаков более двух, применяются другие статистические тесты.

Связь в таблицах гхс

Рассмотрим совокупность из n объектов, каждый из которых описан двумя признаками, измеренными в шкале наименований. Первый 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 H_0) и измерении силы связи.

Классическим тестом, используемым 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;

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

$j = 1, 2, \dots, c$ - 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 (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 $\alpha = 0,001$ для $df = 1$ $\chi^2 = 10,827$.

Шаг 4. Сравниваем фактическое и табличное значения χ^2 . Если $\chi^2 \geq \chi^2_{df, \alpha}$, то гипотеза H_0 на данном уровне значимости α может быть отвергнута. То есть можно утверждать, что с вероятностью $1-\alpha$ 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 $\alpha = 0,05$ ($\alpha = 5\%$). Таблица значений χ^2 при $\alpha = 0,05$ для различных степеней свободы df приведена ниже.

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

Установив наличие связи, исследователь должен измерить ее силу с тем, чтобы иметь возможность сравнивать взаимосвязи между различными характеристиками, выделять наиболее сильные и т.д.

Для измерения силы связи используются коэффициенты:

Средняя квадратичная сопряженность

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

Учитывая тот факт, что верхний предел Φ^2 не больше $\min(r-1, s-1)$ Крамер предложил использовать в качестве меры связи величину $C = \sqrt{C^2}$,

где

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

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

Коэффициент Крамера равен Φ^2 для таблиц 2×2 .

А.А.Чупров предложил использовать в качестве меры связи величину

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

или квадратный корень из этой величины.

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

Исходная информация. Часть 1

В файле ДляЛАБ_ТАБЛИЦЫ.xls 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).

В соответствии с вариантом отобрать для анализа 20 анкет (строк).

№ варианта	1	2	3	4	5	6	7	8
Строки	2-21	5-24	10-29	15-34	20-39	25-44	30-49	32-51
№ варианта	9	10	11	12	13	14	15	
Строки	3-22	6-25	11-30	16-35	21-40	26-45	31-50	

Порядок выполнения работы

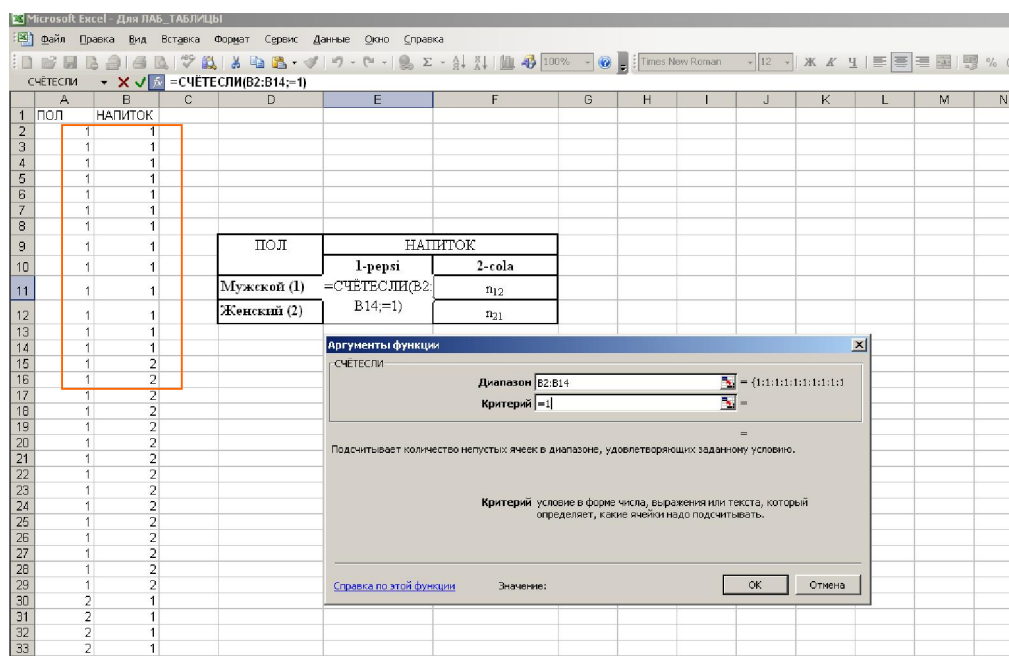
1. Построить четырехклеточную таблицу вида

ПОЛ	НАПИТОК	
	1-pepsi	2-cola
Мужской (1)	n_{11}	n_{12}
Женский (2)	n_{21}	n_{22}

Для удобства подсчета n_{11} , n_{12} , n_{21} , n_{22} провести сортировку данных по первой переменной, а затем по второй.

Заполнение таблицы 2×2 можно провести «вручную» или воспользоваться функцией СЧЁТЕСЛИ.

Пример использования функции СЧЁТЕСЛИ:



В клетку n_{11} занесется число объектов, имеющих значение «1» по первому признаку и значение «1» по второму признаку.

2. Провести одномерный анализ данных: рассчитать частоты, относительные частоты по каждому признаку, построить диаграммы различных видов (гистограммы, секторные; линейные и др.
3. провести анализ таблицы 2×2 : установить наличие связи, рассчитать коэффициенты ассоциации, контингенции
4. Объяснить полученные результаты

Исходная информация. Часть 2

Обработка и анализ данных социологического исследования с использованием программной системы СТАТИСТИКА. Исследование взаимосвязи признаков.

ИСХОДНАЯ ИНФОРМАЦИЯ. В файле Для_Лаб_30анкет.xls приведены ответы 30 респондентов на 10 вопросов

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

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

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

2. Насколько широк КРУГ вашего общения? (1-много друзей, 2-чувствую одиноким)

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

4. Считаете ли вы, что удачный брак - фактор жизненного успеха (1-скорее да, 2-не влияет)

5. ПЛАНИРУЕТЕ РАБОТАТЬ ПО СПЕЦИАЛЬНОСТИ ПОСЛЕ ОКОНЧАНИЯ ВУЗА? (1-ДА, 2-НЕОБЯЗАТЕЛЬНО)

6. ПОЛ (1-м, 2-ж)

7. ВОЗРАСТ

8. ВУЗ

9. ПРОФИЛЬ ОБРАЗОВАНИЯ, КОТОРОЕ ВЫ ПОЛУЧАЕТЕ: (1-технический, 2-гуманитарный)

10. КУРС

2) Запустить модуль базовых характеристик STA_BAS.EXE.

3) Создать файл с исходными данными:

в стандартном файле 10x10 добавить 20 строк и удалить 8 столбцов

скопировать выбранные 2 столбца из файла Для_Лаб_50анкет.xls с данными анкет и вставить их в подготовленную таблицу

4) Провести анализ таблицы сопряженностей 2x2:

Средствами СТАТИСТИКА построить таблицу сопряженности 2x2, получить Хи-квадрат, сравнить его с табличным, получить коэффициенты Φ , Крамера, контингенции

5) Оформить отчет: записать таблицы сопряженностей, значения Хи-квадрат фактическое и теоретическое, значения коэффициентов. Объяснить полученные результаты.

ЛАБОРАТОРНАЯ РАБОТА 3.

Тема работы.

Группировки. Типологические группировки. Многомерная классификация. Кластер-анализ.

Краткие сведения из теории

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.

Типологическая группировка служит для выделения социально-экономических типов.

Правильность проведения типологической группировки проверяется. Для этого рассчитываются сводные показатели по группам (средние, показатели вариации), коэффициент детерминации.

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

Коэффициент детерминации R^2 на основе **правила сложения дисперсий**. Если совокупность разбита на группы, то полная (общая) дисперсия признака σ^2 может быть определена как сумма межгрупповой дисперсии $\delta_{м.гр.}^2$ и средней из групповых дисперсий $\overline{\sigma_i^2}$:

$$\sigma^2 = \overline{\sigma_i^2} + \delta_{м.гр.}^2$$

Полная дисперсия признака вычисляется по обычной формуле:

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

где n – число объектов в совокупности

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

Межгрупповая дисперсия вычисляется по формуле

$$\delta_{м.гр.}^2 = \frac{\sum (\bar{x}_j - \bar{x})^2 f_j}{\sum f_j},$$

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

$\bar{x}_j$ — среднее значение признака в группе j ;

f_j — число наблюдений в группе j .

Коэффициент детерминации R^2 — отношение межгрупповой дисперсии к общей:

$$R^2 = \frac{\delta_{м.гп.}^2}{\sigma^2},$$

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

Методы многомерной классификации 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

$$X = \{x_1, x_2, \dots, x_n\} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

разбить на сравнительно небольшое число (заранее известно или нет) однородных в определенном смысле групп (классов, типов, кластеров, таксонов).

Каждый объект является точкой в признаковом пространстве, расстояния между точками определяют «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.

На каждом шаге объединяются только два объекта (объект + объект, объект + кластер, кластер + кластер). Поэтому 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, которые можно считать также кластерами, состоящими их одного объекта.

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

Если на некотором шаге объединились кластеры i и j с числом объектов n_i и n_j в каждом из них, то *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*. Например, в алгоритме минимальной (одной) связи или «ближайшего соседа» расстояния пересчитываются по формуле:

$$d_{i+j,k} = \min (d_{ik}, d_{jk});$$

В социально-экономических исследованиях применяются также алгоритмы «дальнего соседа», среднего связывания, Варда, центроидной и другие иерархические алгоритмы, реализованные в статистических пакетах.

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 $X_1, X_2, \dots, X_n$ надо разбить на k ($k \ll n$) классов, однородных в смысле некоторой метрики.

Смысл алгоритма состоит в последовательном уточнении эталонных точек – центров классов $E^{(v)} = \{e_1^{(v)}, e_2^{(v)}, \dots, e_k^{(v)}\}$, v - номер итерации. При этом эталонным точкам приписываются «веса» $\Omega^{(v)} = \{\omega_1^{(v)}, \omega_2^{(v)}, \dots, \omega_k^{(v)}\}$, 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 расстояния 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.

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

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 $d(X_i, e_s) = \min_{1 \leq j \leq k} d(X_i, e_j)$, 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.

Кластер-анализ в STATISTICA

В программе STATISTICA реализованы следующие методы кластеризации — агломеративные методы: **joining (tree-clustering)**, two-way joining, а также метод k -средних — **k-means clustering**.

В STATISTICA можно выбрать следующие правила *иерархического объединения кластеров*:

Single linkage — метод одиночной связи,

Complete linkage — метод полной связи,

Unweighted pair group average — невзвешенный метод «средней связи»,

Weighted pair group average — взвешенный метод «средней связи»,

Weighted centroid pair group (median) — взвешенный центроидный метод,

Ward method — метод Уорда (Варда).

Данные алгоритмы различаются правилами объединения объектов в кластеры. Например, в *методе одиночной связи* (Single linkage) на первом шаге объединяются два объекта, имеющие между собой *максимальную* меру сходства, то есть находящиеся на *минимальном расстоянии* друг от друга. На следующем шаге к ним присоединяется объект с максимальной мерой сходства с *одним* из объектов кластера.

В *методе полных связей* мера сходства между объектом — кандидатом на включение в кластер и всеми членами кластера не может быть меньше некоторого порогового значения. В *методе средней связи* мера сходства между кандидатом и членами кластера усредняется, например, берется просто среднее арифметическое мер сходства.

Идея еще одного агломеративного метода — *метода Уорда* (Ward method) состоит в том, чтобы проводить объединение, дающее минимальное приращение внутригрупповой суммы квадратов отклонений..

Результатом работы иерархических алгоритмов является дендрограмма (иерархическое дерево).

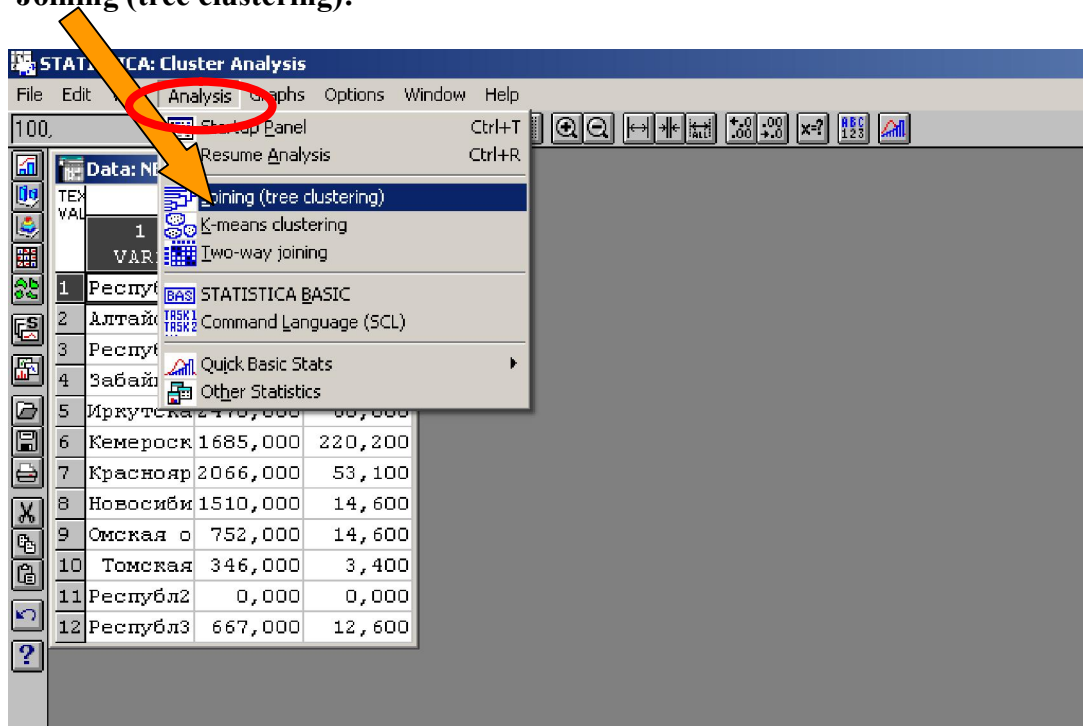
Последовательный (итеративный) метод группировки ***k-средних*** — **k-means clustering** работает непосредственно с *объектами*, а не с матрицей расстояний. В методе ***k-средних*** объект относится к тому классу, расстояние до центра тяжести (*эталонной точки*) которого минимально. В результате работы алгоритма определяется принадлежность точек к кластеру - «алфавит классификации» (*Members of each cluster*), а также рассчитываются межклассовые расстояния.

Пример работы с модулем Cluster Analysis в программе STATISTICA

Необходимо провести классификацию объектов-регионов в пространстве числовых признаков.

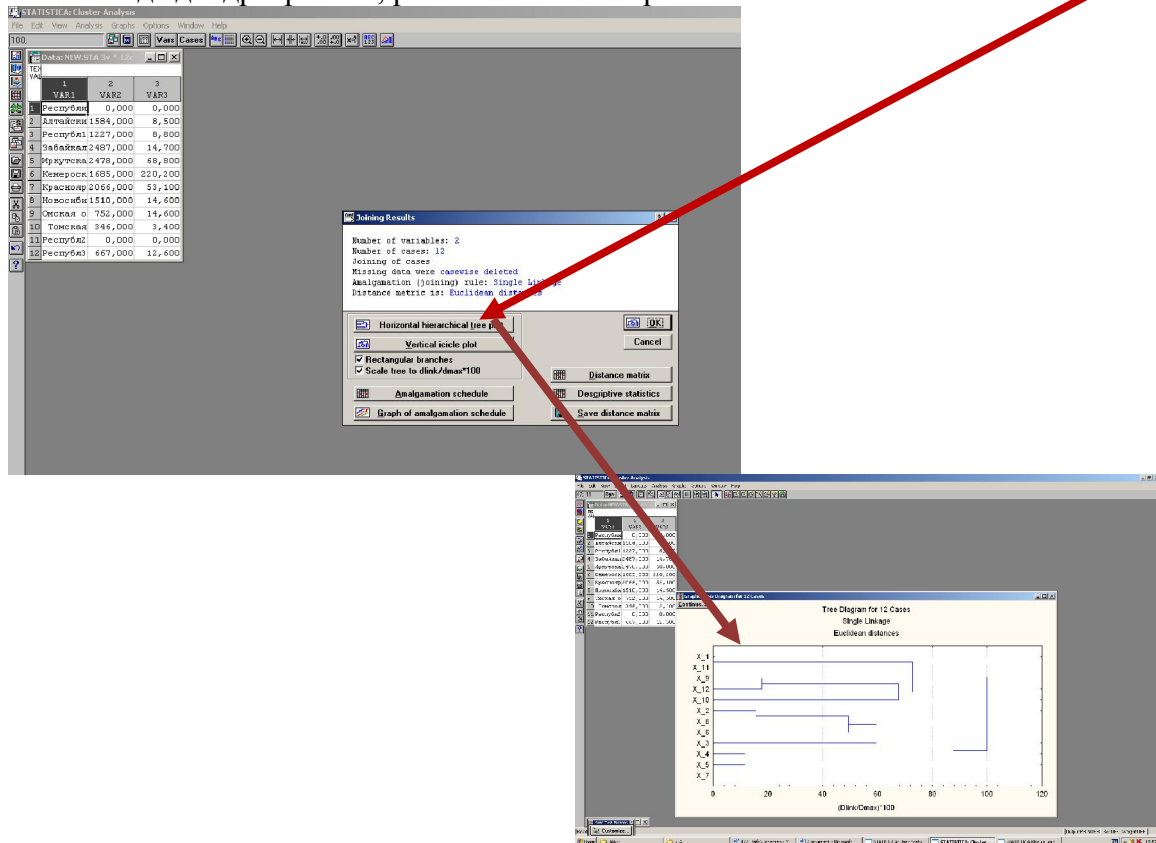
- 1) запустить модуль кластерного анализа STA_CLU.EXE и создать таблицу с данными; Если вы работали в модуле базовых характеристик STA_BAS.EXE, то ***переключитесь*** на модуль кластерного анализа
- 2) Для проведения кластерного анализа с использованием ***иерархических алгоритмов*** на панели кластерного анализа выбрать **Analysis** и в выпадающем окне выбрать

Joining (tree clustering):

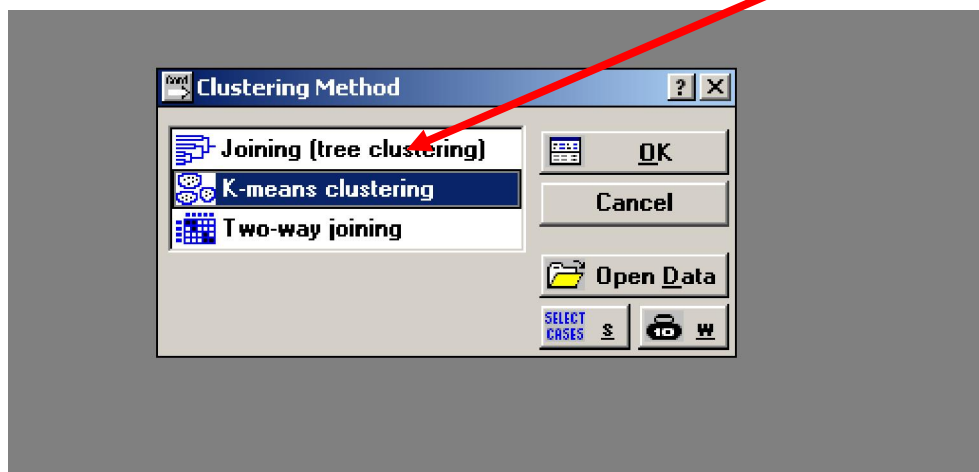


3) В выпадающем окне необходимо переменные кластеризации, объекты кластеризации и метод единственной связи (**Single Linkage**); алгоритм Варда (**Ward**).

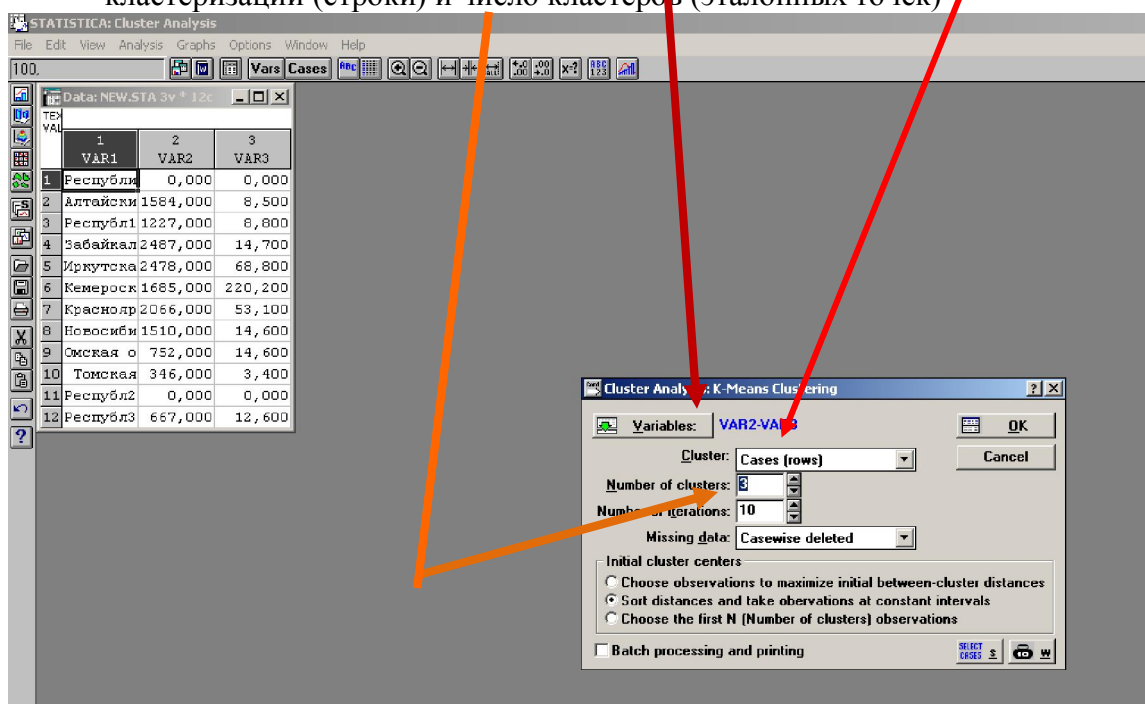
4) Результаты кластеризации могут быть представлены в разных формах, например, в виде дендрограммы, расположенной горизонтально



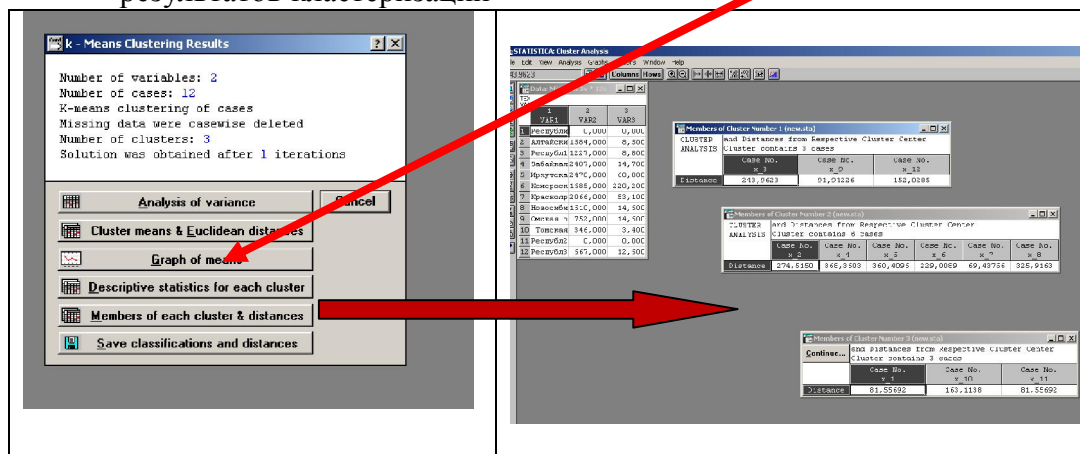
5) Для проведения кластерного анализа с использованием *алгоритма k-средних* на панели кластерного анализа выбрать **Analysis** и в выпадающем окне выбрать **k-means clustering**:



- 6) В выпадающем окне необходимо переменные кластеризации, объекты кластеризации (строки) и число кластеров (эталонных точек)



- 7) Алфавит классификации можно увидеть, выбрав кнопку на панели результатов кластеризации



Задание на лабораторную работу. Часть 1.

Изучить основные положения работы с программной системой STATISTICA (файл О системе STATISTICA 5.0):

- Запуск системы
- Создание файла в системе
- Добавление, удаление, копирование переменных и объектов
- Первичный анализ данных в системе
- Переключение модулей,

Задание на лабораторную работу. Часть 2

Определить вариант лабораторной работы

Вариант	Номер показателя - 1	Номер показателя - 2	Вариант	Номер показателя - 1	Номер показателя - 2	Вариант	Номер показателя - 1	Номер показателя - 2

1.	2	1	6.	7	1	11.	12	1
2.	3	1	7.	8	1	12.	13	1
3.	4	1	8.	9	1	13.	14	1
4.	5	1	9.	10	1	14.	15	1
5.	6	1	10.	11	1	15.	16	1

Исходные данные

Сформировать в **Excel** исходную таблицу, содержащую названия регионов и указанные в варианте показатели социально-экономического развития регионов СФО (ПРИЛОЖЕНИЕ 1).

Регион	Показатель 1	Показатель 2 (ВРП)
...	...	...

Порядок выполнения работы

1). Запустить модуль **STA_BAS.EXE**.

- Скопировать файл исходных данных из EXCEL в систему STATISTICA, предварительно увеличив число строк (случаев – *case*) до нужного количества.
- Уметь средствами STATISTICA определять описательные статистики показателей (среднее, дисперсия и др.), коэффициент корреляции, строить диаграммы.

2) Переключиться на модуль кластерного анализа **Cluster Analysis**.

- 2.1. **Провести кластерный анализ**, используя *иерархический алгоритм* (Joining (tree clustering) единственной связи (**Single Linkage**)). Определить вариант разбиения на 3 класса и результаты разбиения записать в табл.1
- 2.2. **Провести кластерный анализ**, используя *иерархический алгоритм* (Joining (tree clustering) Варда (**Ward**)). Определить вариант разбиения на 3 класса и результаты разбиения записать в табл.1.
- 2.3. **Провести кластерный анализ** на 3 класса, используя метод К-средних (**k-means clustering**) и результаты разбиения (из алфавита классификации) записать в табл.1.

Таблица 1. Варианты разбиения регионов на классы

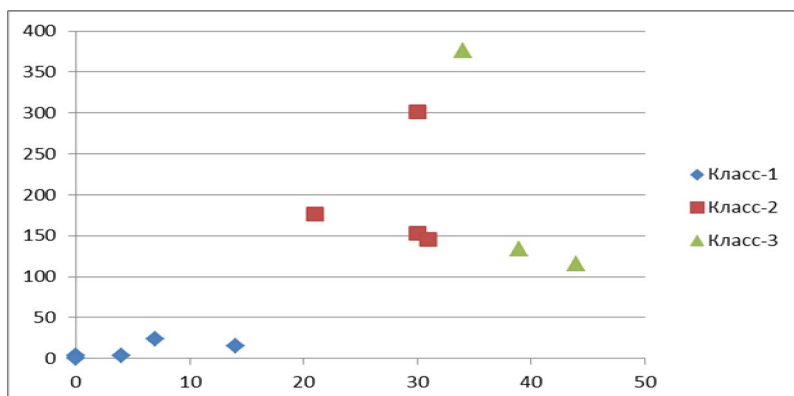
	Названия регионов, входящих в классы		
	Класс № 1 (малые)	Класс №2 (средние)	Класс №3 (крупные)
Single Linkage			
Ward			
k-means			

3) Сравнить результаты, полученные тремя алгоритмами кластер-анализа (схожие результаты?).

4. По результатам метода К-средних:

4.1. внести в исходный файл (EXCEL) дополнительный столбец с номером кластера, то есть каждому региону приписать номер кластера, в который он попал при кластеризации методом К-средних;

4.2. Построить точечную диаграмму в пространстве двух признаков с изображением принадлежности регионов кластерам. Для удобства можно провести сортировку по переменной «номер кластера». Например,



4.3. Рассчитать коэффициенты детерминации R^2 по каждому из двух классификационных признаков.

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

ЛАБОРАТОРНАЯ РАБОТА 4

Тема работы.

Исследование однородности статистической совокупности.

Краткие сведения из теории

Средняя величина должна вычисляться по однородной совокупности. Если совокупность неоднородна, то **общие средние**, рассчитанные по всей совокупности, должны подкрепляться **групповыми средними**, дающими характеристику размера явления, складывающуюся в конкретных условиях данной группы.

Оценка однородности совокупности производится на основе вычисления средних величин и показателей вариации.

Средние величины делятся на две большие категории:

- степенные средние (к ним относятся средняя арифметическая, средняя гармоническая, средняя геометрическая, средняя квадратическая);
- структурные средние (мода и медиана).

Средней арифметической величиной называется такое среднее значение признака, при вычислении которого общий объем признака в совокупности сохраняется неизменным. При ее вычислении общий объем признака мысленно распределяется поровну между всеми единицами совокупности.

Если данные представлены в виде списка, то среднее значение вычисляется по формуле простой средней

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

. Если данные сгруппированы и представлены в виде дискретного или интервального вариационного ряда, то средняя величина должна рассчитываться по формуле взвешенной средней

$$\bar{X} = \frac{\sum x_i f_i}{\sum f_i}$$

Структурные средние — **медиана** и **мода** — характеризуют величину значения признака, занимающего определенное значение в ранжированном вариационном ряду.

Мода (M_o) — наиболее часто встречающееся значение признака в совокупности. В дискретном ряду мода определяется без вычислений как значение признака с наибольшей частотой.

В интервальном ряду модальным является интервал с наибольшей частотой. Внутри него определяется **точечная мода**

$$M_o = X_0 + h \cdot \frac{f_{M_o} - f_{M_o-1}}{(f_{M_o} - f_{M_o-1}) + (f_{M_o} - f_{M_o+1})},$$

где X_0 — нижняя граница модального интервала;

f_{M_o} — частота в модальном интервале;

f_{M_o-1} — частота в предыдущем интервале;

f_{M_o+1} — частота в следующем за модальным интервале;

h — величина модального интервала.

Медиана (Me) — значение признака, которое делит **упорядоченную** последовательность его значений на две равные по численности части. В итоге у одной половины единиц совокупности значение признака не превышает медианного уровня, а у другой — не меньше его.

Положение медианы определяется ее номером $N_{Me} = (n + 1) / 2$, где n — число единиц в совокупности.

В интервальном вариационном ряду для нахождения медианы применяется формула

$$Me = X_0 + h \cdot \frac{\left(\sum_{i=1}^k f_i \right) / 2 - f'_{Me-1}}{f_{Me}},$$

где Me — медиана;

X_0 — нижняя граница интервала, в котором находится медиана;

h — величина интервала;

k — число интервалов;

f'_{Me-1} — накопленная частота в интервале, предшествующем медианному;

f_{Me} — частота в медианном интервале.

Аналогично медиане вычисляются значения признака, делящие совокупность на четыре равные части — **квартили** Q . Второй квартиль совпадает с медианой, а первый и третий рассчитываются по формулам:

$$Q_1 = X_0 + h \cdot \frac{\left(\sum_{i=1}^k f_i \right) / 4 - f'_{Q_1-1}}{f_{Q_1}} ;$$

$$Q_3 = X_0 + h \cdot \frac{\left(3 \sum_{i=1}^k f_i \right) / 4 - f'_{Q_3-1}}{f_{Q_3}} .$$

Значения признака, делящие ряд **на пять равных частей называются квинтилями**, на десять частей — децилями, на сто частей — перцентилями.

Для измерения вариации признака применяются различные абсолютные и относительные показатели.

Размах или **амплитуда вариации** — абсолютная разность между максимальным и минимальным значениями признака изучаемой совокупности

$$R = X_{\max} - X_{\min} .$$

Точнее характеризуют вариацию показатели, учитывающие колеблемость всех значений признака.

Среднее линейное отклонение $d = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} \quad i = 1, 2, \dots, n.$

Если данные сгруппированы $d_{\text{гр}} = \frac{\sum_{j=1}^k |x_j - \bar{x}| f_j}{\sum_{j=1}^k f_j} \quad j = 1, 2, \dots, k,$

где k — число групп;

f_j — число единиц совокупности j -й группы.

Дисперсия $D_x = \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ и $D_{x_{\text{гр}}} = \sigma^2 = \frac{\sum_{j=1}^k (x_j - \bar{x})^2 f_j}{\sum f_j} .$

Стандартное отклонение (среднее квадратическое отклонение) равно квадратному корню из дисперсии, т.е. $\sigma = \sqrt{D_x} .$

Для оценки интенсивности вариации и для сравнения ее в различных совокупностях и для разных признаков **необходимы относительные показатели вариации**. Они вычисляются как отношения абсолютных показателей силы вариации к средней арифметической величине признака:

- относительный размах вариации ρ : $\rho = R : \bar{x}$;
- относительное линейное отклонение: $m = d : \bar{x}$;
- коэффициент вариации v : $v = \sigma : \bar{x}$.

Принято считать, что если коэффициент вариации более 30%, то совокупность неоднородна и ее надо делить на группы.

Порядок выполнения работы

Исходными данными являются данные о заработной плате работников (таблица к заданию 2 лабораторной работы 1)

1) . На основе **исходных** данных определить:

- а) среднее значение показателя, моду и медиану

- б) размах вариации, среднее линейное отклонение, дисперсию, стандартное отклонение, коэффициент вариации
- 2). 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
- 3). 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

Вопросы для защиты работы.

- 1) Какие формулы для вычисления средних значений использовались.
- 2) Как изменятся среднее значение, дисперсия, стандартное отклонение и коэффициент вариации при увеличении значения признака у всех единиц совокупности в 3 раза?.
- 3) Какова размерность показателей вариации?
- 4) Однородна ли исследуемая совокупность?

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

Тема «Вариационные ряды. Представление статистических данных»

Задание. 1). На основе исходных данных построить дискретный вариационный ряд; представить его в виде статистической таблицы и статистических графиков. 2). На основе исходных данных построить интервальный вариационный ряд с равными интервалами. Число интервалов выбрать самостоятельно и объяснить этот выбор. Представить полученный вариационный ряд их в виде статистической таблицы и статистических графиков. Указать виды примененных таблиц и графиков.

Для выполнения расчетов и построения графиков рекомендуется использовать Excel.

Пример выполнения задания

Известны размеры месячной заработной платы рабочих бригады за май 2003 года:

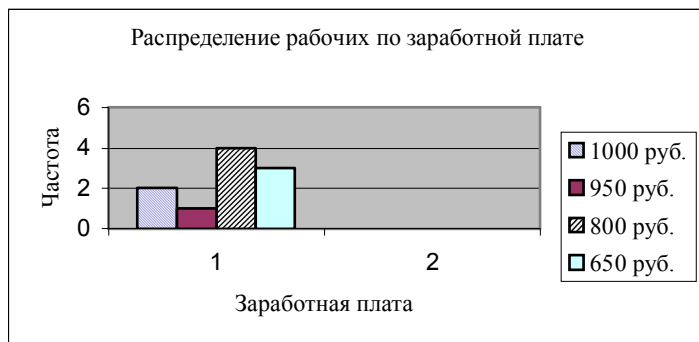
№ п/п	Месячная зарплата, руб.	№ п/п	Месячная зарплата, руб.
1.	800	6.	800
2.	650	7.	950
3.	800	8.	650
4.	800	9.	650
5.	1000	10.	1000

Сгруппируем исходные данные. Получим дискретный вариационный ряд и оформим его в виде таблицы.

Таблица 1. Распределение рабочих по заработной плате

Месячная зарплата, руб.	1000	950	800	650	Всего
Число рабочих	2	1	4	3	10

Построим столбиковую диаграмму.



Построим интервальный вариационный ряд. Число равных интервалов определяем по формуле Старджесса

$$k = 1 + 3,32 \lg n = 1 + 3,32 = 4,32. \text{ Округляем } k = 4.$$

Вычисляем длину интервала.

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{1000 - 650}{4} = 87,5.$$

Определяем границы интервалов:

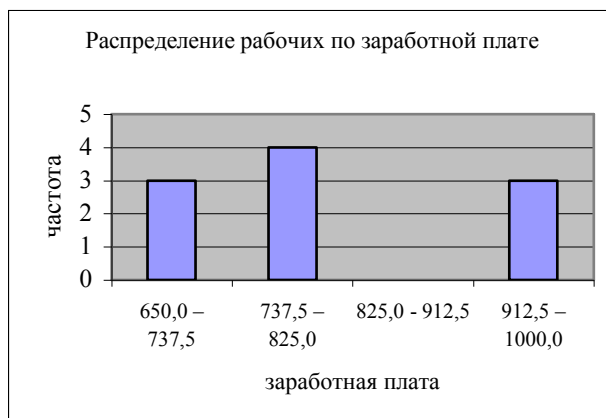
$$\begin{aligned} x_{1\min} &= 650; x_{1\max} = x_{1\min} + h = 650 + 87,5 = 737,5; \\ x_{2\min} &= x_{1\max} = 737,5; x_{2\max} = x_{2\min} + h = 737,5 + 87,5 = 825; \\ x_{3\min} &= x_{2\max} = 825; x_{3\max} = x_{3\min} + h = 825 + 87,5 = 912,5; \\ x_{4\min} &= x_{3\max} = 912,5; x_{4\max} = x_{4\min} + h = 912,5 + 87,5 = 1000. \end{aligned}$$

Вариационный ряд оформляем в виде таблицы.

Таблица 2. Распределение рабочих по заработной плате

Интервалы зарплатной платы, руб.	Число рабочих
650,0 – 737,5	3
737,5 – 825,0	4
825,0 – 912,5	0
912,5 – 1000,0	3
Всего	10

Строим столбиковую диаграмму



Задачи для самостоятельного выполнения.

Задача 1.

Швейная фабрика выпустила партию 50 мужских костюмов следующих размеров:

54, 52, 40, 44, 48, 50, 52, 48, 50, 50, 44, 52, 48, 44, 56, 58, 40, 42, 54, 52, 60, 48, 52, 52, 52, 48, 56, 56, 50, 52, 48, 52, 48, 48, 52, 60, 40, 44, 46, 46, 60, 56, 52, 48, 46, 52, 58, 58, 60, 52.

Задача 2.

Зарплатная плата работников предприятия (руб.):

15000, 26400, 95000, 95000, 56000, 15000, 26400, 44000, 8000, 9500, 56000, 26400, 44000, 56000, 56000, 9500, 44000, 56000, 8000, 26400, 44000, 56000, 44000, 44000, 44000, 80000, 15000, 26400, 44000, 8000.

Задача 3.

Урожайность зерновых (ц/га) в хозяйствах области:

11, 20, 22, 25, 12, 24, 25, 16, 20, 20, 14, 26, 21, 27, 17, 27, 15, 23, 22, 13, 20, 14, 26, 21, 27, 17, 27, 15, 23, 22, 13, 11, 20, 22, 25, 12, 24, 25, 16, 20.

Тема «Средние величины».

Пример выполнения задания.

На основании следующих данных по двум сельскохозяйственным кооперативам необходимо определить, в каком из них и насколько выше средняя урожайность зерновых культур:

Культура	СХК «Заря»		СХК «Луговое»	
	Валовый сбор, ц	Урожайность, ц/га	Посевная площадь, га	Урожайность, ц/га
Пшеница	32500	25	1540	20
Рожь	1620	18	120	19
Ячмень	13640	22	460	18
Просо	1650	15	80	13
ИТОГО	49410	—	2200	—

Для СХК «Заря» среднюю урожайность определяем как среднюю гармоническую взвешенную

$$\bar{Y}_1 = \frac{\sum_{i=1}^4 BC_i}{\sum_{i=1}^4 \frac{BC_i}{Y_i}} = \frac{32500 + 1620 + 13640 + 1650}{\frac{32500}{25} + \frac{1620}{18} + \frac{13640}{22} + \frac{1650}{15}} = \frac{32500 + 1620 + 13640 + 1650}{1300 + 90 + 620 + 110} = \frac{49410}{2120} = 23,31 \text{ ц/га};$$

Здесь слагаемые знаменателя — посевные площади, занятые под различные культуры.

Для СХК «Луговое» средняя урожайность рассчитывается как средняя арифметическая взвешенная, причем в качестве весов берутся посевные площади:

$$\bar{Y}_2 = \frac{\sum_{i=1}^4 Y_i \Pi\Pi_i}{\sum_{i=1}^4 \Pi\Pi_i} = \frac{20 \cdot 1540 + 19 \cdot 120 + 18 \cdot 460 + 13 \cdot 80}{1540 + 120 + 460 + 80} = \frac{42400}{2200} = 19,24 \text{ ц/га};$$

Таким образом, средняя урожайность зерновых культур в СХК «Заря» по сравнению с СХК «Луговое» выше на 4,04 ц/га (или на 21%).

Задачи для самостоятельного выполнения

Задача 1.

Заполнить пробелы и рассчитать среднюю заработную плату в целом по всем предприятиям в случаях

	Численность ППП, чел.	Средняя заработная плата, руб.	Месячный фонд заработной платы, тыс.руб
А	1	2	3
ООО «Фиалка»	100	?	500
ООО «Кактус»	?	5000	100
ООО «Роза»	10	?	100
ООО «Ландыш»	70	3000	210

а) если известны данные столбцов 1 и 2

б) если известны данные столбцов 1 и 3

в) если известны данные столбцов 2 и 3

Какой вид средней величины применяется в каждом случае?

Задача 2.

Определить Среднее, Медианное и Модальное значение распределения по размеру обуви, проданной фирмой

Размер	34	35	36	37	38	39	40	41	42	43	44 и более
Количество проданных пар	3	5	7	9	10	13	15	14	20	3	1

Задача 3.

Имеются следующие данные о распределении населения по душевым доходам

Группы населения по душевому доходу, руб.	Численность населения, тыс.чел.
До 1000	50
1000-5000	36
5000-10000	10
10000 и более	4

Определить среднее значение, моду и медиану признака «Душевой доход».

Задача 4.

Расстояние между городами А и Б 600 километров, автобус прошел его со скоростью 60 км/час. Расстояние между городами Б и В 100 километров, автобус прошел его со скоростью 50 км/час, а при возвращении из В в А автобус ехал со скоростью 70 км/час. Какова средняя скорость автобуса за весь рейс? Какой вид средней величины применяется?

Задача 5.

Удельный вес шоколадных конфет во всем выпуске кондитерских изделий на четырех кондитерских фабриках области составляет: на первой - 25%, на второй - 12%, на третьей - 8%, на четвертой - 15%. Объем выпуска всей продукции соответственно равен: 500, 300, 100 и 200 млн.руб.

Определить среднюю долю выпуска шоколадных конфет по всем кондитерским фабрикам области в целом. Какой вид средней величины применяется?

Задача 6. В таблице приведены данные о выпуске продукции по годам

	2000г.	2001г.	2002г.	2003г.
Выпуск приборов	2	6	8	16

Определить, во сколько раз в среднем ежегодно возрастал выпуск приборов с 2000 по 2003 год. Какой вид средней величины применяется?

Задача 7.

Определить среднее значение, моду и медиану признака «Число студентов в группах университета» :

25, 27, 30, 17, 25, 25, 17, 30, 20, 20, 20, 17, 25, 27, 20, 25, 27, 25, 30, 20, 17, 27, 20, 20, 17, 25, 30, 27, 20, 25, 25, 17, 30, 20, 20, 20, 17, 25, 27, 25, 25, 17, 30, 20, 20, 20, 17, 25, 14, 20.

Тема «Показатели вариации».

Пример выполнения задания

В таблице приведены данные социологического исследования - распределение посетителей театра по возрасту (графы 1 и 2).

Группы посетителей по возрасту, лет x_i	Число посетителей f_i	Расчетные графы		
		Середина интервала а, лет x'_i	$x'_i f_i$	$(x_j - \bar{x})^2 f_j$
1	2	3	4	5
10–15	6	12,5	75,0	1893,0
15–20	9	17,5	157,5	1465,9
20–25	20	22,5	450,0	1205,0
25–30	41	27,5	1127,5	312,8
30–35	26	32,5	845,0	130,2
35–40	21	37,5	787,5	1100,1
40–45	14	42,5	595,0	2096,7
45–50	5	47,5	237,5	1485,7
50–55	1	52,5	52,5	494,5
ИТОГО	143		4327,5	10183,9

Определить, однородна ли совокупность посетителей по признаку «возраст».

Решение.

Показателем однородности совокупности может служить коэффициент вариации

$$v: v = \sigma : \bar{x}.$$

Вычислим среднее значение признака «возраст» по формуле средней арифметической взвешенной. Для удобства вычислений дополним исходную таблицу столбцами 3-5, в которых будем помещать промежуточные расчеты.

Так как исходный ряд – интервальный, то при вычислении среднего значения используется срединное значение интервала, равное полусумме нижней и верхней границ интервала:

$$(10 + 15)/2 = 12,5; (15 + 20)/2 = 17,5 \text{ и т.д.}$$

Вычисленные значения середины интервала записываем в графу 3. В графе 4 запишем значения произведения $x'_i f_i$. Тогда итог графы 4 равен числителю в формуле среднего.

Среднее значение возраста равно

$$\bar{X} = \frac{\sum x_i f_i}{\sum f_i} = \frac{4327,5}{143} = 30,3 \text{ лет.}$$

Для определения дисперсии вычислим и поместим в графу 5 значения выражения $(x_j - \bar{x})^2 f_j$. Дисперсия равна

$$D = \frac{\sum_{j=1}^k (x_j - \bar{x})^2 f_j}{\sum f_j} = \frac{10183,9}{143} = 71,2.$$

Среднее квадратическое отклонение равно квадратному корню из дисперсии, т.е. $\sigma = \sqrt{D} = 8,44$ лет.

Находим коэффициент вариации

$$v = \frac{\sigma}{\bar{x}} = \frac{8,44}{30,3} = 0,279 \text{ или } v = 27,9\%.$$

Так как коэффициент вариации менее 30%, то можно говорить об однородности опрошенной совокупности посетителей театра.

Задачи для самостоятельного выполнения

Задача 1. В таблице приведено распределение земельных участков по размеру.

Размер земельного участка, га	2	4	6	7	8	9	10	12
Количество участков	5	10	25	30	10	10	7	5

Определить размах вариации и среднее квадратическое отклонение показателя «размер земельного участка»

Задача 2.

В таблице приведены показатели объема товарооборота торговых предприятий.

Группы предприятий по объему товарооборота, млн.руб.	Число предприятий
90-100	30
100-110	50
110-120	24
120-130	4
130-150	2
150-200	10

Определить показатели вариации показателя «объем товарооборота».

Задача 3.

Показать, что при увеличении значения признака у всех единиц совокупности в 3 раза, дисперсия признака увеличится в 9 раз, а коэффициент вариации не изменится.

Задача 4.

В таблице приведены показатели объема товарооборота торговых предприятий двух регионов

Регион 1		Регион 2	
Группы предприятий по объему товарооборота, млн.руб.	Число предприятий	Группы предприятий по объему товарооборота, млн.руб.	Число предприятий
90-100	26	60-80	13
100-110	40	80-100	27
110-120	24	100-120	20

В каком регионе коэффициент вариации показателя «объем товарооборота» меньше?

Задача 5.

Жилая площадь квартир (кв.м), выставленных на продажу агентством недвижимости: 45, 36, 98, 36, 56, 56, 36, 42, 36, 42, 45, 90, 42, 45, 56, 36, 42, 42, 45, 56.

Однородна ли совокупность квартир? Каким показателем можно оценить однородность совокупности?

Тема «Экономические индексы»

Пример вычисления индексов.

В таблице приведены данные о ценах и количестве проданных товаров за 2 месяца.

То- вар	Еди- ница изме- рения	Цена, руб.		Количество проданных товаров	
		март, p_0	апрель, p_1	март, q_0	апрель, q_1
Чай	Пачка	20	18	100	150
Кофе	Банка	60	70	200	200
Сыр	Кг	50	50	40	38

Определить индивидуальные и общие индексы цен, физического объема и стоимости товара.

Решение.

Индивидуальный индекс физического объема продукции i_q рассчитывается по формуле

$$i_q = \frac{q_1}{q_0}.$$

Определяем индивидуальные индексы для каждого продукта

$$i_{q \text{ чай}} = q_1 \text{ чай} / q_0 \text{ чай} = 150/100 = 1,5 \text{ или } 150\%,$$

то есть физический объем продаж чая увеличился в 1,5 раза или на 150% - 100% = 50%;

$$i_{q \text{ кофе}} = q_1 \text{ кофе} / q_0 \text{ кофе} = 200/200 = 1 \text{ или } 100\%,$$

то есть физический объем продаж кофе остался на прежнем уровне;

$$i_{q \text{ сыр}} = q_1 \text{ сыр} / q_0 \text{ сыр} = 38/40 = 0,95 \text{ или } 95\%,$$

то есть физический объем продаж сыра составил 95% от базисного уровня или уменьшился на 5%.

Индивидуальный индекс цены i_p рассчитывается по формуле

$$i_p = \frac{p_1}{p_0}.$$

Определяем индивидуальные индексы для каждого продукта

$$i_{p \text{ чай}} = p_1 \text{ чай} / p_0 \text{ чай} = 18/20 = 0,9 \text{ или } 90\%,$$

то есть цена 1 пачки чая составил 90% от базисного уровня или уменьшилась на 10%.

$$i_{p \text{ кофе}} = p_1 \text{ кофе} / p_0 \text{ кофе} = 70/60 = 1,167 \text{ или } 116,7\%,$$

то есть цена одной банки кофе возросла в 1,167 раза или увеличилась на 16,7% по сравнению с ценой марта.

$$i_{p \text{ сыр}} = p_1 \text{ сыр} / p_0 \text{ сыр} = 50/50 = 1 \text{ или } 100\%, \text{ то есть цена 1 кг сыра не изменилась.}$$

Индивидуальный индекс стоимости каждого вида продукции можно определить по формуле

$$i_{pq} = \frac{p_1 q_1}{p_0 q_0}$$

или можно применить соотношение между индексом цены и индексом физического объема

$$i_{pq} = i_p \times i_q$$

Используем последнее соотношение:

$$i_{pq \text{ чай}} = i_{p \text{ чай}} i_{q \text{ чай}} = 0,9 \cdot 1,5 = 1,35 \text{ или } 135\%,$$

то есть стоимость реализации чая увеличилась в 1,35 раза или на 35%;

$$i_{pq \text{ кофе}} = i_{p \text{ кофе}} i_{q \text{ кофе}} = 1,167 \cdot 1 = 1,167 \text{ или } 116,7\%,$$

то есть стоимость реализации кофе увеличилась на 16,7%;

$$i_{pq \text{ сыр}} = i_{p \text{ сыр}} i_{q \text{ сыр}} = 1 \cdot 0,95 = 0,95 \text{ или } 95\%, \text{ то есть товарооборот сыра составил 95\% от базисного уровня или уменьшился на 5\%.}$$

Индекс стоимости продукции, или товарооборота (I_{pq}), представляет собой отношение стоимости продукции текущего периода ($\sum p_1 q_1$) к стоимости продукции в базисном периоде ($\sum p_0 q_0$) и определяется по формуле

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{18 \times 150 + 70 \times 200 + 50 \times 38}{20 \times 100 + 60 \times 200 + 50 \times 40} = \frac{18600}{16000} = 1,163.$$

Следовательно, стоимость продукции (товарооборота) в апреле по сравнению с мартом возросла в 1,163 раза (товарооборот составил 116,3%). Стоимость продукции увеличилась на 16,3%, или 18600 – 16000 = 2600 руб.

Вычисляем общий индекс физического объема по формуле

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{20 \times 150 + 60 \times 200 + 50 \times 38}{20 \times 100 + 60 \times 200 + 50 \times 40} = \frac{16900}{16000} = 1,056, \text{ т.е. } 105,6\%,$$

Таким образом, за счет увеличения физического объема продукции на 5,6% (105,6% – 100%) ее стоимость в абсолютном выражении увеличилась на 900 руб. (16900 – 16000).

Для вычисления общего (агрегатного) индекса цен применим формулу

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{18 \times 150 + 70 \times 200 + 50 \times 38}{20 \times 150 + 60 \times 200 + 50 \times 38} = \frac{18600}{16900} = 1,101$$

Таким образом, в среднем по трем товарам цены выросли в 1,101 раза (или рост цен составил 110,1%). В результате увеличения цен на 10,1% (110,1% – 100%) покупатели заплатили на 1700 руб. больше в апреле, чем в марте (18600 – 16900 = 1700).

Задачи для самостоятельного выполнения

Задача 1.

Имеются данные о продаже товаров в торговых предприятиях района:

Вид товара	Товарооборот в действующих ценах, тыс.руб.		Изменение цен во 2-м квартале по сравнению с 1-м кварталом, %
	1-й квартал	2-й квартал	
Обувь	60	80	+ 10
Трикотаж	25	30	+ 5
Галантерея	40	35	- 5

Определить: общий индекс цен, общий индекс физического объема товарооборота, общий индекс товарооборота

Задача 2.

Имеются данные о ценах и количестве проданного товара

Вид товара	Единица измерения	Цена, руб.		Реализовано, тыс.ед.	
		1999 г.	2000 г.	1999 г.	2000 г.
Пиво	л	10	11	600	500
Рыба	кг	25	30	100	150

Определить: индивидуальные индексы цен, физического объема и товарооборота

Задача 3.

Цены на потребительские товары и услуги в регионе в январе по сравнению с предшествующим месяцем возросли на 5%, и в феврале по сравнению с январем – на 5%.

Как изменились цены в марте по сравнению с февралем, если

а) общий рост цен за 1 квартал данного года составил 110%

б) при расчете всех индексов использовались веса декабря предшествующего года

Задача 4.

Имеются данные о себестоимости и произведенной продукции на двух предприятиях

Вид продукции	Предприятие А		Предприятие Б	
	себестоимость, руб.	произведено, шт.	себестоимость, руб.	произведено, шт.
1	375	1018	384	624
2	120	965	120	980
3	415	383	418	1540

Рассчитать индекс себестоимости продукции предприятия А по сравнению с предприятием Б

Задача 5.

Определите изменение физического объема реализации товаров в текущем периоде по сравнению с предыдущим, если товарооборот возрос на 42,3%, а цены повысились на 13,7%

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОРГАНИЗАЦИИ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Цель самостоятельной работы по дисциплине – повышение эффективности изучения теоретической части дисциплины и полноценной работы на лабораторных занятиях, а также получение навыков самостоятельного проведения поиска, сбора информации, использования методов, в том числе математических, для анализа, моделирования и прогнозирования социально-экономических явлений

Самостоятельная работа основывается на программе дисциплины и методических указаниях по отдельным темам. 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. На этом этапе студент в основном должен работать в библиотеке и компьютерном зале. Для закрепления материала темы необходимо ответить на предлагаемые в пособиях вопросы и пререшать задачи по теме.

Тема «Прикладная статистика как наука. Информационная база статистики. Статистическое исследование»

Статистика как наука представляет собой целостную систему научных дисциплин: теория статистики, экономическая статистика и ее отрасли, социальная статистика и ее отрасли и т.д. . В отличие от классической математической статистики анализ данных не пытается вывести свойства окружающего мира исходя из специально собранных данных, а ориентирован на отыскание каких-либо образов, закономерностей, структуры в имеющихся данных.

Задача статистического исследования состоит в получении обобщающих характеристик и выявлении закономерностей на основе изучения объекта исследования. Важно уяснить, что статистическое исследование состоит из трех стадий: 1) статистическое наблюдение; 2) сводка и группировка результатов наблюдения; 3) анализ полученных обобщающих показателей. Все три стадии связаны между собой, на каждой из них используются специальные методы.

Первым этапом статистического исследования является статистическое наблюдение – сбор статистических данных. Важным моментом является рассмотрение статистического наблюдения как целенаправленного, научно-организованного процесса. Следует обратить внимание на программно-методологические вопросы его обеспечения. Для правильного выбора формы, вида и способа статистического наблюдения необходимо четко уяснить, что объектом исследования выступает определенная статистическая совокупность.

Следует обратить внимание, наиболее существенной информацией, которую должен иметь исследователь, прежде чем отобрать статистические методы, применяемые к данным, является *уровень измерения* каждой переменной во множестве данных. Традиционная классификация уровней измерения была разработана американским психологом С.С.Стивенсом (Stanley Smith Stevens) в 1946 году. Схема классификации основана на типах операций, допустимых для данной переменной. В порядке возрастания числа допустимых операций Стивенс идентифицировал четыре уровня: номинальный, порядковый, интервальный и отношений. Существуют и другие классификации уровней и измерения.

При изучении инструментальных средств большое внимание уделяется решению задачи прогнозирования и построению модели тренда. В этой связи рекомендуется повторить тему «Ряды динамики» из дисциплины «Статистика». Метод экстраполяции

тренда - один из пассивных методов прогнозирования, - и называется «наивным» прогнозом, так как предполагает строгую инерционность развития, которая представляется в виде проектирования прошлых тенденций в будущее, а главное — независимость показателей развития от тех или иных факторов. Как в теоретической части курса, так и на практических занятиях подробно рассматривается линейная модель тренда как простейший случай модели. Студенту следует на примере линейной модели уяснить методику построения модели, а для более сложных случаев необходимо использовать статистические пакеты прикладных программ.

Тема Анализ одномерных количественных данных

Все используемые в статистической практике показатели по форме выражения классифицируются на абсолютные, относительные и средние. Выбор конкретной формы зависит от имеющихся данных и поставленной задачи.

При изучении абсолютных величин следует обратить внимание, что они отражают физические размеры явлений или объем (численность) совокупности и всегда являются именованными числами.

Для аналитических целей в статистической практике широко применяются относительные величины, представляющие результат деления одного показателя на другой и выражающие соотношения между количественными характеристиками. При расчете относительных величин показатель, с которым проводится сравнение, является делителем и называется основанием или базой сравнения. Относительный показатель выражается в коэффициентах (разах), процентах, промилле и т.д.

В зависимости от изучаемого явления определяются относительные величины динамики, плана (планового задания), выполнения плана; относительные величины структуры, координации, интенсивности, сравнения. Необходимо уяснить назначение каждого вида относительных величин и метод их расчета. Следует также обратить внимание на связь между относительными величинами динамики, плана и выполнения плана.

Вариационный ряд – совокупность значений варьирующего признака и соответствующих им численностей единиц совокупности. Следует различать дискретные и интервальные вариационные ряды. Необходимо обратить внимание на способы определения числа интервалов, способы «закрытия» открытых интервалов. При изучении вариационных рядов надо обратить внимание на понятия «интервал» и «группа»: построение интервального вариационного ряда – это группировка единиц совокупности.

Для оценки достигнутого уровня изучаемого показателя, при расчете нормативов, при анализе и планировании применяются средние величины. Средняя величина — обобщающий показатель, характеризующий типический уровень признака в расчете на единицу совокупности в конкретных условиях места и времени. В средних величинах проявляется статистическая закономерность. Важно уяснить, что средняя величина должна вычисляться по однородной совокупности.

Степенные средние величины (средние значения) в зависимости от величины степени m называются арифметическая, квадратическая, гармоническая, геометрическая средняя. Кроме того, в зависимости от представления исходных данных бывают простые и взвешенные степенные средние величины, которые рассчитываются по соответствующим формулам. При выборе способа расчета (формулы) средней взвешенной величины необходимо прежде всего определить, какой признак является осредняемым, а какая величина выступает в качестве веса.

Структурные средние — мода и медиана — характеризуют величину значения признака, занимающего определенное значение в ранжированном вариационном ряду. В дискретном вариационном ряду мода и медиана отыскиваются непосредственно по определению. В интервальном вариационном ряду сначала определяется интервал, содержащий искомую величину, а затем внутри него вычисляется точечная мода (или медиана) по соответствующей формуле. Важно уяснить, что мода и медиана – это

определенные значения признака (не путать с номером единицы совокупности или частотой).

Для численной оценки вариации признака служат показатели: размах вариации, среднее линейное отклонение, дисперсия, среднее квадратическое (стандартное) отклонение. В зависимости от представления исходных данных расчет производится по соответствующим формулам (простые и взвешенные показатели). Относительные показатели вариации позволяют оценить однородность совокупности и сравнивать разные совокупности. Базой для сравнения служит средняя арифметическая. Чаще всего относительные показатели выражаются в процентах. Для расчетов показателей вариации рекомендуется применять табличный способ расчетов.

Тема «Выборочный метод в статистике»

Тема «Выборочный метод в статистике» взаимосвязана со всеми предыдущими темами, в особенности, со статистическим наблюдением, статистическими показателями и показателями вариации. Она также связана с курсами математики и теории вероятностей (закон больших чисел, теоремы Чебышева, Бернулли и др.). Важными вопросами являются определение способа отбора единиц совокупности, вычисление ошибок выборки и построение доверительных интервалов выборочных характеристик, расчет необходимого объема выборки.

Совокупность единиц, из которых производится отбор, принято называть генеральной совокупностью. Совокупность отобранных единиц из генеральной совокупности называется выборочной совокупностью.

При любом способе отбора должен соблюдаться принцип: каждой единице генеральной совокупности обеспечивается одинаковая вероятность (возможность) быть выбранной. Единица отбора совпадает с единицей наблюдения. Этим обеспечивается возможность распространения выводов, сделанных на основе выборочного наблюдения, на все генеральную совокупность.

Различают среднюю и предельную ошибки выборки. Расчет ошибок позволяет решить одну из главных проблем организации выборочного наблюдения – оценить представительность (репрезентативность) выборочной совокупности. Величина средней ошибки выборки рассчитывается дифференцировано в зависимости от способа отбора (повторный или бесповторный) и процедуры (вида) выборки по соответствующим формулам.

Тема «Анализ взаимосвязей двух количественных признаков»

На практике часто возникают задачи, когда нужно исследовать взаимосвязь между какими-либо непрерывными данными, например, между спросом на телевизоры и денежными доходами населения. Для анализа взаимосвязи используются корреляционный и регрессионный анализ. Корреляционно-регрессионный анализ как общее понятие включает в себя измерение тесноты, направления связи и установления аналитического выражения (формы) связи.

Чаще всего выделяют один результативный признак, на который влияют один или несколько факторов. Необходимо различать следующие варианты зависимостей.

1. Парная корреляция – связь между двумя признаками: результативным и факторным или двумя факторными.

2. Частная корреляция – зависимость между результативным и одним факторным признаком при фиксированном значении других факторных.

3. Множественная корреляция – зависимость результативного признака от двух или более факторных признаков, включенных в исследование.

Тема «Методы многомерного статистического анализа»

В зависимости от специализации и природы используемых методов в литературе класс методов, целью которых является группировка объектов, называется распознаванием образов, многомерной классификацией, таксономией, кластер - анализом или дискриминантным анализом. Суть метода распознавания образов состоит в следующем: объекты, обладающие типичными для всех чертами, образуют характерный класс, называемый образом. Требуется построить некоторый оптимальный алгоритм, который бы позволил по этим типичным чертам определить принадлежность данного объекта к тому или иному классу, то есть распознать образ. Следует проработать основные определения понятий функции расстояния, эталонных точек, иерархического дерева, критерия оптимальности многомерной классификации. При изучении темы рекомендуется рассмотреть примеры решения задач в двумерном признаковом пространстве, что позволяет визуализировать процессы построения классов.

При анализе многомерных данных обычно рассматривают задачи и методы снижения размерности признакового пространства, которые рассматриваются на примере методов главных компонент и факторного анализа. Студенту следует уяснить содержательную и математическую постановку задачи, последовательность этапов ее решения. Поскольку на практике задачу снижения размерности решается с помощью компьютерных программ, то необходимо уметь анализировать и интерпретировать полученные результаты обработки исходных данных.

Тема «Методы статистического анализа нечисловых данных»

При обработке экспериментальных данных, зафиксированных в номинальной шкале, непосредственно с самими данными можно выполнять *только* операцию проверки их совпадения или несовпадения. Это означает, что измеряя значение (категорию) признака, можно отнести этот объект к группе (классу) объектов с этой категорией, то есть можно считать частоты.

Для анализа взаимосвязи таких признаков применяются методы анализа таблиц сопряженностей и так называемый χ^2 - тест, основанный на использовании критерия согласия χ^2 Пирсона..

Под *ранговой корреляцией* понимается статистическая связь между порядковыми переменными. Эта связь анализируется на основе исходных статистических данных, представленных упорядочениями (ранжировками) n рассматриваемых объектов по разным свойствам. Следует рассмотреть часто встречающуюся в исследовательской работе задачу оценки согласованности мнений экспертов и освоить расчет коэффициентов Спирмена, Кендала, конкордации

При выполнении индивидуального творческого задания по теме «*Исследование рынка информационно-аналитических систем (по типам)*» следует использовать информационные ресурсы интернета, специализированных изданий. Кроме того, желательным моментом является проведение опроса пользователей и применение изученных методов для обработки данных исследования.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Боровиков В.П., Популярное введение в современный анализ данных в системе STATISTICA: учеб. пособие для вузов. – М.: Горячая линия-Телеком, 2013. – 288 с. ГРИФ УМО [Электронный ресурс]. — URL: http://e.lanbook.com/books/element.php?pl1_id=11828
2. Буре, В.М. Методы прикладной статистики в R и Excel. [Электронный ресурс] / В.М. Буре, Е.М. Парилина, А.А. Седаков. — Электрон. дан. — СПб. : Лань, 2016. — 152 с. — Режим доступа: <http://e.lanbook.com/book/81558> —. ГРИФ
3. Анализ данных на компьютере [Текст] : учебное пособие / Ю. Н. Тюрин, А. А. Макаров. - 4-е изд., перераб. - М. : ФОРУМ, 2012. - 368 с. В библиотеке ТУСУРа: 3 экз.
4. Информационные системы и технологии в экономике и управлении : Учебное пособие для вузов / Санкт-Петербургский государственный университет экономики и финансов ; ред. : В. В. Трофимов. - 2-е изд., перераб. и доп. - М. : Высшее образование, 2007. - 480 с. : ил. - (Университеты России). - Библиогр. в конце разд. - ISBN 5-9692-0114-6. Экземпляры всего: 20
5. 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. ГРИФ СибРУМЦ. В библиотеке ТУСУРа: 20 экз.
6. Статистические методы в управлении качеством продукции [Текст] : учебное пособие для вузов / В. В. Ефимов, Т. В. Барт. - 2-е изд., стереотип. - М. : КноРус, 2013. - 235 с. - В библиотеке ТУСУРа: 20 экз.
7. Информационные технологии управления: Учебник / Б.В. Черников. – М.: Форум, 2008; М.:инфра-М, 2008. -351 с. – (Высшее образование). - ISBN 978-5-8199-0347-6 (ИД ФОРУМ) – ГРИФ : В библиотеке ТУСУР: 10 экз
8. Руководство Осло. Рекомендации по сбору и анализу данных по инновациям [Текст] : вопросы измерения научно-технологической деятельности / Организация экономического сотрудничества и развития, Статистическое бюро европейских сообществ. - 3-е изд. - Томск : ТУСУР, 2011. - 206 с. В библиотеке ТУСУРа: 51 экз.
9. Воскобойников, Ю.Е. Регрессионный анализ данных в пакете MATHCAD + CD. [Электронный ресурс] — Электрон. дан. — СПб. : Лань, 2011. — 224 с. — Режим доступа: <http://e.lanbook.com/book/666>.
10. Статистика: Учебное пособие/ ред. : М. Р. Ефимова. - М.: Инфра-М, 2006. – 335 с. В библиотеке ТУСУРа 90 экз.

ПРИЛОЖЕНИЕ 1

Значения F-распределения Фишера

Уровень значимости $\alpha=0,05$

Число степеней свободы $\nu_1 = k-1=1$, $\nu_2 = n-k = n-2$,

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

$\nu_2 =$ $n-2$	F_{tab}	$\nu_2 =$ $n-2$	F_{tab}	$\nu_2 =$ $n-2$	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 2

ЧИСЛЕННОСТЬ ПЕРСОНАЛА, ЗАНЯТОГО НАУЧНЫМИ ИССЛЕДОВАНИЯМИ И РАЗРАБОТКАМИ

Вариант		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
	Российская Федерация	887729	885568	870878	858470	839338	813207	807066	801135	761252	742433	736540	735273
	Сибирский ФО	62477	63271	63052	63376	62494	60986	58647	56427	53956	53463	53024	52794
1	Республика Алтай	90	97	105	122	124	111	132	168	173	156	158	172
2	Республика Бурятия	1209	1200	1249	1299	1278	1231	1003	985	954	969	952	1144
3	Республика Тыва	285	323	343	334	330	327	332	416	414	425	416	415
4	Республика Хакасия	128	125	120	108	107	282	205	198	193	166	149	148
5	Алтайский край	3427	2941	2919	2980	2870	2732	2775	2731	2267	2054	1955	2182
6	Забайкальский край	680	775	840	823	779	509	608	520	512	335	322	313
7	Красноярский край	7196	7637	7486	7303	7087	7102	6846	6685	6287	6299	6475	6748
8	Иркутская область	5295	5408	5387	5184	4983	4829	4557	4910	4897	4919	4912	5075
9	Кемеровская область	2090	1916	1820	1594	1557	1476	1511	1496	1327	1336	1258	1231
10	Новосибирская область	25168	25494	25537	25888	25077	24791	23438	22561	21597	21622	21615	21569
11	Омская область	8872	9442	9478	9825	10127	9367	8983	7246	6961	6622	6125	5002
12	Томская область	8037	7913	7768	7916	8175	8229	8257	8511	8374	8560	8687	8795

ПРИЛОЖЕНИЕ 3

Таблица - Основные показатели развития регионов Сибирского федерального округа

	Валовой региональный продукт в 2010 г., млн. руб.	Площадь территории ¹ , тыс. км ²	Численность населения на 1 января 2012 г., тыс. человек	Средне-годовая численность занятых в экономике, тыс. человек	Средне-душевые денежные доходы (в месяц), руб.	Потребительские расходы в среднем на душу населения (в месяц), руб.	Средне-месячная 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	21635,8	92,9	208,4	91,7	13836,9	7179,0	15632,4	61628
Республика Бурятия	136374,0	351,3	971,4	417,4	15715,5	11340,0	19924,0	430210
Республика Тыва	30601,0	168,6	309,4	106,0	10962,8	4944,6	19163,1	47409
Республика Хакасия	93709,0	61,6	532,2	239,2	14222,8	9680,5	20689,5	292915
Алтайский край	299715,3	168,0	2407,2	1075,6	12499,9	9765,7	13822,6	757632
Забайкальский край	162100,2	431,9	1099,4	489,4	15968,8	10572,7	21099,6	650405
Красноярский край	1050158,5	2366,8	2838,4	1437,5	20145,5	14105,7	25658,6	1815754
Иркутская область	539245,6	774,8	2424,4	1121,7	16017,2	10580,2	22647,7	1975486
Кемеровская область	622513,0	95,7	2750,8	1302,0	16666,0	11237,2	20478,8	1406912
Новосибирская область	482026,5	177,8	2686,9	1305,1	18244,1	14898,1	20308,5	1229181
Омская область	371218,1	141,1	1974,8	945,5	17247,9	12663,1	19087,8	725451
Томская область	284292,0	314,4	1057,7	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.	в том числе		Ввод в действие общей площади жилых домов ² , 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		растениеводства	животноводства				
9	10	11	12	13	14	15	16	17	
862	1296	1654	8020	1205	6815	76,6	14312	11802	Республика Алтай
12808	51115	19826	13044	4605	8439	304,4	100938	41017	Республика Бурятия
3376	514	2868	4648	827	3821	52,4	13742	7033	Республика Тыва
26536	56595	21462	9371	3216	6155	156,2	46034	38064	Республика Хакасия
6041	189279	31991	93784	46670	47114	663,2	218077	70833	Алтайский край
40377	14365	18311	15154	3433	11721	276,9	106366	51557	Забайкальский край
266636	628113	95432	68598	32708	35890	1047,1	361607	303885	Красноярский край
129795	299406	81275	43610	18748	24862	755,2	225846	137995	Иркутская область
507993	385413	85949	38044	19621	18423	1082,6	287279	225131	Кемеровская область
19674	249816	61592	60425	26860	33565	1505,2	368292	142078	Новосибирская область
4411	529355	36013	66911	33535	33376	836,7	228595	83342	Омская область
137513	100598	25617	19420	7225	12195	457,6	93050	101927	Томская область

ПРИЛОЖЕНИЕ 4

Таблица Размер заработной платы работников (тыс.руб.)

Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
23	23	18	20	30
20	20	22	18	20
25	25	32	20	20
20	20	18	20	18
20	20	20	23	18
18	18	16	20	20
20	20	20	18	18
18	18	18	22	20
20	20	24	18	20
20	20	20	22	23
23	23	23	32	20
20	20	20	18	16
18	18	18	20	22
22	22	22	23	23
18	18	16	20	30
22	22	22	30	20
32	32	23	20	18
18	18	20	20	20
20	20	24	18	20
16	17	20	18	23
20	20	20	20	20
18	18	18	18	18
20	20	20	20	22
20	20	18	20	18
23	23	20	23	22
20	20	20	20	32
18	18	23	18	18
22	22	20	22	20
16	17	18	23	23
22	22	22	30	20
Число равных интервалов				
К=4	К=3	К=4	К=3	К=4

Вариант 6	Вариант 7	Вариант 8	Вариант 9	Вариант 10
23	23	18	20	25
20	20	22	18	20
25	25	28	20	20
20	20	18	20	18
20	20	20	23	18
18	18	16	20	20
20	20	20	18	18
18	18	18	22	20
20	20	24	18	20
20	20	20	22	23
23	23	23	27	20
20	20	20	18	18
18	18	18	20	22

22	16	22	23	23
18	18	16	20	23
22	22	22	27	20
29	32	23	20	17
18	18	20	20	20
20	20	24	18	20
17	17	20	18	23
20	20	20	20	20
18	18	18	18	18
20	20	20	20	22
20	20	18	20	18
23	23	20	23	22
20	20	20	20	25
18	18	23	18	18
22	22	20	22	20
25	17	18	23	23
22	22	22	25	20
Число равных интервалов				
K=3	K=4	K=4	K=3	K=4