ПРИДНЕСТРОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИСТЕТ
им. Т.Г. ШЕВЧЕНКО
Рыбницкий филиал

Кафедра прикладной информатики

ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

Лабораторный практикум
Часть I

Рыбница
2013

УДК 004.6
ББК 32.973
Б 39

Составители:
Е.А. Безвугляк, преп. каф. прикладной информатики
И.И. Сычева, преп. каф. прикладной информатики

Рецензенты:
Скодорова Л.К., к.соц.н., доц. каф. прикладной информатики
Лоскутова Е.В., доц. каф. прикладной информатики

Информационные системы и технологии: Лабораторный практикум. Часть I /
Б 39 Сост.: Е.А. Безвугляк, И.И. Сычева.– Рыбница, 2013. – 115 с.

Лабораторный практикум содержит методические рекомендации для студентов по изучению применения информационных систем и технологий с использованием возможностей статистического пакета для социальных наук SPSS 17.0. Методические рекомендации к каждой лабораторной работе включают в себя краткую характеристику изучаемого пакета, подробное описание хода работы, требования к содержанию отчета, контрольные вопросы и задания.

Предназначено для студентов, получающих экономическое образование по направлению подготовки «Прикладная информатика» профилю подготовки 351400 «Прикладная информатика в экономике».

УДК 004.6
ББК 32.973

Рекомендовано Научно-методическим советом ПГУ им. Т.Г. Шевченко

Содержание

Введение

Предлагаемый практикум содержит набор лабораторных работ по дисциплине «Информационные системы и технологии», которая изучается студентами I-го курса Рыбницкого филиала Приднестровского государственного университета им. Т. Г. Шевченко.

Лабораторный практикум рассчитан на освоение первой части учебной программы по дисциплине «Информационные системы и технологии» в соответствии с требованиями ГОСТа по по направлению подготовки «Прикладная информатика» профилю подготовки 351400 «Прикладная информатика в экономике».

Назначение данного лабораторного практикума – освоение студентами применения информационных систем и технологий на примере статистического пакета данных для социальных наук SPSS 17.0.

Практикум предполагает наличие у студентов знаний высшей математики, теории вероятностей, экономической теории, а также владении основами современных компьютерных технологий.

Содержание комплекса лабораторных работ направлено на формирование у студентов навыков работы с методами обработки данных, полученных в результате научных или производственных экспериментов, исследования различных процессов, выявления закономерностей в поведении объектов, процессов и систем.

Тематика лабораторных работ охватывает широкий диапазон заданий: от первичной обработки данных до анализа временных рядов.

Изложение материала ведется в логической последовательности, начиная с обоснования необходимости в знаниях современных информационных систем и технологий.

Лабораторный практикум, в целях более эффективного усвоения учебного материала, содержит краткое теоретическое введение по каждой теме, подробные методические указания с описанием этапов анализа, наглядные примеры анализа, реализация проведения анализа в статистическом пакете SPSS 17.0, варианты задач для самостоятельного решения.

Каждый студент индивидуально выполняет лабораторные работы практикума. Для этого он получает одно из заданий по указанному преподавателем варианту. Задание выполняется в соответствии с методическими указаниями под руководством преподавателя.

Лабораторная работа считается выполненной после её защиты. Для

защиты работы необходимо предоставить отчет, оформленный в соответствии со следующими требованиями.

Отчет по лабораторной работе оформляется в печатном или письменном виде индивидуально каждым студентом, выполнившим необходимые задания работы (независимо от того, выполнялись ли задачи индивидуально или в составе группы студентов). Текст печатается на бумаге формата А4 размером и с интервалом между строками, удобными для чтения, или пишется от руки так же на бумаге формата А4 или в специально отведенной для этого тетради. Страницы отчета следует пронумеровать (титульный лист не нумеруется, далее идет страница 2 и т.д.).

Титульный лист отчета должен содержать фразы: «Отчет по лабораторной работе (номер работы)» «Тема: (тема работы)», далее оформляется по стандарту. Внизу листа следует указать текущий год.

Отчет должен содержать следующие основные разделы:

1. Тема и цель работы.
2. Теоретическая часть.
3. Практическая часть.
4. Вывод.

Теоретическая часть должна содержать минимум необходимых теоретических сведений по теме работы. Объем должен быть таким, чтобы было ясно, что студент изучает. В случае надобности при объяснении результатов удобно сослаться на соответствующие формулы и заключения теоретической части.

В практической части студент должен отразить исходные данные и реализацию проведения анализа на ПК. Так же должен отразить полученные результаты и проанализировать их.

Объем отчета должен быть оптимальным для понимания того, что и как сделал студент, выполняя работу. Обязательные требования к отчету включают общую и специальную грамотность изложения, а также аккуратность оформления.

Лабораторная работа №1

Тема: Программа SPSS Statistics 17.0. Основы работы.

Цель: Ознакомиться с программой SPSS Statistics 17.0. Изучить ее функциональные возможности.

1. Программа SPSS Statistics 17.0

SPSS (аббревиатура англ. «Statistical Package for the Social Sciences», «статистический пакет для социальных наук») (рис.1.1) — компьютерная программа для статистической обработки данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в социальных науках.

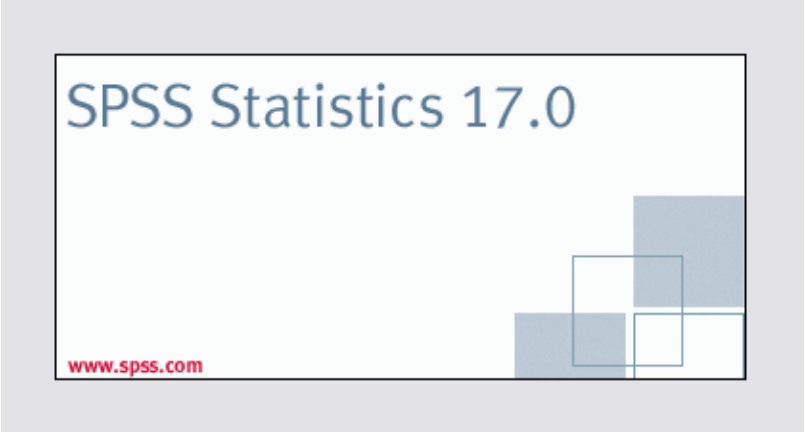


Рис.1.1. Загрузочное окно программы SPSS Statistics 17.0

По мнению некоторых авторов, SPSS «занимает ведущее положение среди программ предназначенных для статистической обработки информации».

Норман Най, Хедли Халл и Дейл Бент разработали первую версию системы в 1968 году, затем этот пакет развивался в рамках Чикагского университета.

Первое пользовательское руководство вышло в 1970 году в издательстве McGraw-Hill, а с 1975 года проект выделился в отдельную компанию SPSS Inc. Первая версия пакета под Microsoft Windows вышла в 1992 году. На данный момент существуют версии под MacOs X и Linux.

SPSS Statistics 17.0 предоставляет мощную систему анализа и управления данными с графическим интерфейсом, оснащенным удобными меню и простыми диалоговыми окнами, упрощающими работу пользователя. Большинство задач могут быть решены при помощи нескольких щелчков мышью.

В SPSS Statistics есть несколько типов окон:

- Редактор данных. В Редакторе данных отображается содержимое файла данных. С помощью Редактора данных Вы можете создавать новые файлы данных или изменять старые. Если у Вас открыто более одного файла данных, то для каждого файла существует отдельное окно Редактора данных.
- Viewer. Все статистические результаты, таблицы и диаграммы отображаются во Viewer. Вы можете редактировать вывод и сохранять его для дальнейшего использования. Окно Viewer открывается автоматически, когда выходные результаты создаются первый раз за сеанс.
- Редактор диаграмм. Вы можете изменять диаграммы в окнах диаграмм. Вы можете изменить цвета, шрифты, поворачивать оси, вращать трехмерные диаграммы и даже изменять типы диаграмм.
- Редактор синтаксиса. Вы можете вставить выбранные параметры статистических процедуры из диалоговых окон в окно синтаксиса, где они появятся в виде команд. После этого можно отредактировать синтаксис команды, чтобы использовать специальные возможности, которые недоступны через меню. Эти команды можно сохранить в файле для использования в последующих сеансах.

2. Редактор данных

Редактор данных - это окно, похожее по внешнему виду на окно электронной таблицы и предназначенное для создания и редактирования файлов данных. Окно Редактора данных открывается автоматически при запуске SPSS.

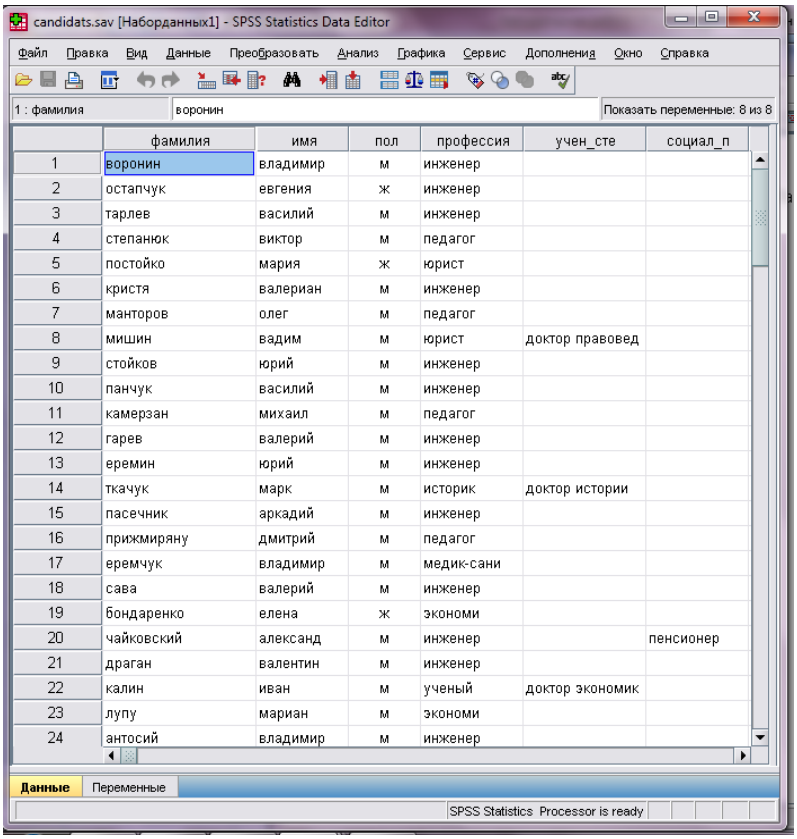
В Редакторе данных есть два режима работы с данными:

- Закладка Данные. В этом режиме можно просматривать и редактировать фактические значения данных.
- Просмотр переменных. В этом режиме можно просматривать и редактировать свойства переменных, включая метки переменных и значений, типы данных (например, текст, дата, или число), типы шкал измерения (номинальная, порядковая, или количественная) и определяемые пользователем пропущенные значения.

В обоих режимах просмотра данных можно добавлять, изменять и удалять информацию, содержащуюся в файле данных.

Закладка Данные

Закладка *Данные* (рис.1.2) очень похожа на электронную таблицу.



	фамилия	имя	пол	профессия	учен_сте	социал_п
1	воронин	владимир	м	инженер		
2	остапчук	евгения	ж	инженер		
3	тарлев	василий	м	инженер		
4	степанюк	виктор	м	педагог		
5	постойко	мария	ж	юрист		
6	кристя	валериан	м	инженер		
7	манторов	олег	м	педагог		
8	мишин	вадим	м	юрист	доктор правовед	
9	стойков	юрий	м	инженер		
10	панчук	василий	м	инженер		
11	камерзан	михаил	м	педагог		
12	гарев	валерий	м	инженер		
13	еремин	юрий	м	инженер		
14	ткачук	марк	м	историк	доктор истории	
15	пасечник	аркадий	м	инженер		
16	прижмиряну	дмитрий	м	педагог		
17	еремчук	владимир	м	медик-сани		
18	сава	валерий	м	инженер		
19	бондаренко	елена	ж	экономи		
20	чайковский	александ	м	инженер		пенсионер
21	драган	валентин	м	инженер		
22	калин	иван	м	ученый	доктор экономик	
23	лупу	мариан	м	экономи		
24	антосий	владимир	м	инженер		

Рис.1.2. Закладка «Данные» редактора данных

Однако существуют и важные отличия между вкладкой *Данные* и электронной таблицей:

- Строки являются наблюдениями. Каждая строка представляет наблюдение. Например, отдельный респондент в опросе является наблюдением, и данные о нем записываются в отдельную строку.
- Столбцы являются переменными. Каждый столбец представляет переменную или измеряемую характеристику. Например, каждый вопрос в анкеты - это одна переменная.
- Ячейки содержат значения. Каждая ячейка содержит одно значение переменной для одного наблюдения. Ячейка представляет собой пересечение строки наблюдения и столбца переменной. Ячейки содержат только значения данных. В отличие от электронных таблиц, ячейки в Редакторе данных не могут содержать формулы.
- Файл данных является прямоугольным. Размер файла данных определяется числом наблюдений и переменных. Вы можете вводить данные в любую ячейку. Если Вы вводите данные в

ячейку вне границ определенных наблюдений и переменных, прямоугольник данных будет расширен так, чтоб включить любые строки и/или столбцы между вводимой ячейкой и границами файла. В границах файла данных нет «пустых» ячеек. Для числовых переменных пустые ячейки преобразуются в системные пропущенные значения. Для текстовых переменных пустые ячейки преобразуются в пробелы.

Закладка Переменные

Закладка *Просмотр переменных* (*Переменные*) (рис.1.3) содержит описания свойств переменных, содержащихся в файле данных.

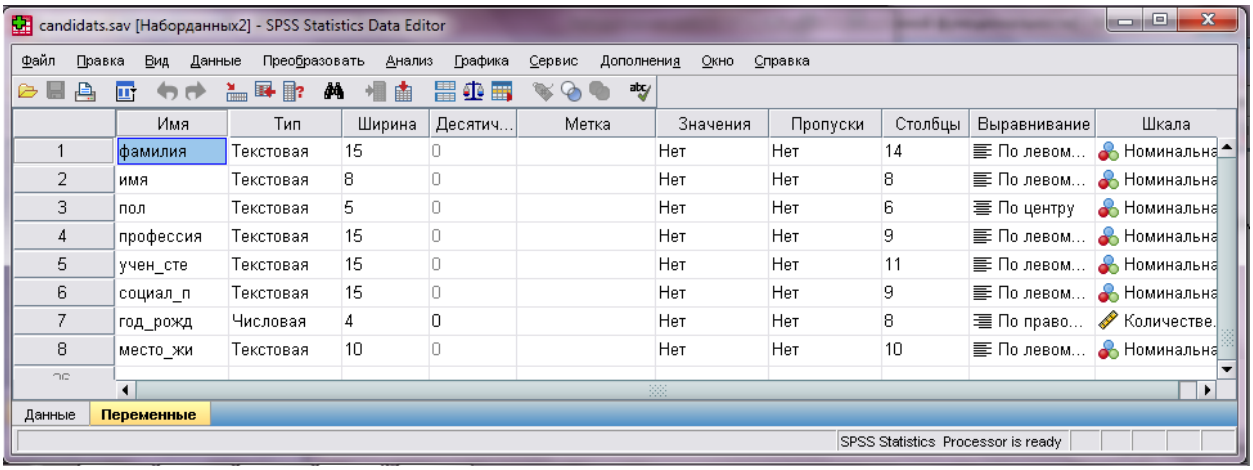


Рис.1.3. Закладка «Переменные» редактора данных

На вкладке *Просмотр переменных*:

- Строки - это переменные.
- Столбцы - это свойства переменных.

Вы можете не только добавлять или удалять переменные, но и изменять свойства переменных, включая следующие атрибуты:

- Имя переменной
- Тип переменной
- Число цифр или символов в переменной
- Число десятичных знаков
- Описательные метки переменных и значений
- Пользовательские пропущенные значения
- Ширину столбца
- Уровень измерения

Все эти свойства сохраняются вместе с сохраняемым файлом данных.

Кроме задания свойств переменных на вкладке *Просмотр переменных*,

существует еще два способа задания свойств переменных:

- Конструктор копирования свойств данных предоставляет возможность использования внешнего файла данных SPSS Statistics или другого набора данных в текущей сессии в качестве шаблона при определении свойств файла и переменных в активном наборе данных. Вы можете также использовать переменные активного набора данных в качестве шаблонов для других переменных активного набора данных. Процедура *Копировать свойства данных* доступна в меню *Данные* в окне *Редактора данных*.
- Процедура Задать свойства переменных (также доступная в меню *Данные* в окне *Редактора данных*) сканирует данные и выводит все уникальные значения данных для выбранных переменных, указывая на значения без меток и предоставляет возможность автоматического создания меток. Эта процедура особенно полезна для категориальных переменных, в которых категории представлены числовыми кодами, например, 0=Мужской,1=Женский.

3. Диалоговые окна

Большинство пунктов меню открывает диалоговые окна (рис.1.4). Диалоговые окна используются для выбора переменных и параметров анализа.

У диалоговых окон статистических процедур, как правило, есть два основных элемента:

- Список исходных переменных. Список переменных активного набора данных В списке исходных переменных находятся имена переменных только тех типов, которые могут быть использованы в данной процедуре. Использование коротких и длинных текстовых переменных ограничено во многих процедурах.
- Список (списки) выбранных переменных. Один или несколько списков переменных, выбранных для анализа, например, списки зависимых и независимых переменных.

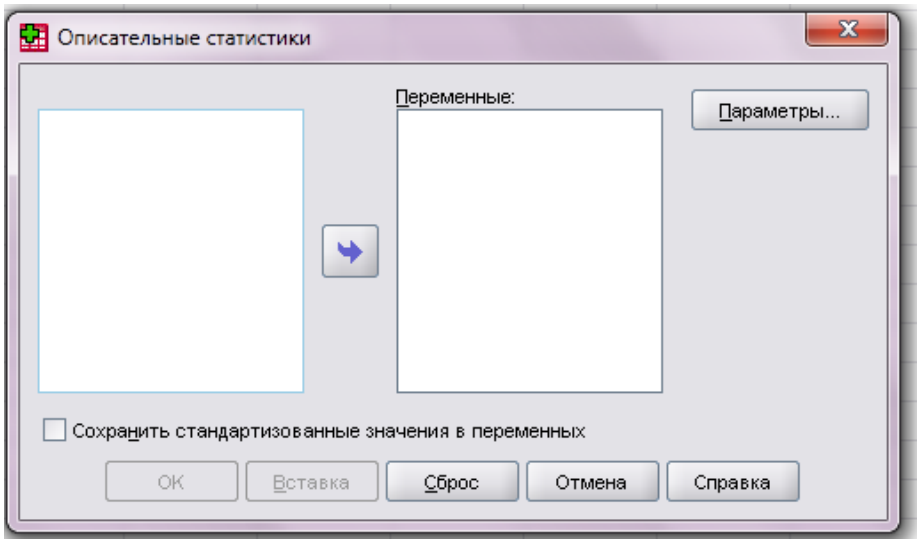


Рис.1.4. Диалоговое окно «Описательные статистики»

Выбор переменных

Чтобы выделить одну переменную, просто щелкните по ней в списке исходных переменных, а затем перетащите ее на целевой список переменных. Для перемещения переменных между списками можно также использовать кнопку со стрелкой. Если список выбранных переменных только один, переменную можно выбрать, дважды щелкнув по ней.

Можно также выбрать сразу несколько переменных:

- Чтобы выбрать группу переменных, которые расположены в списке последовательно друг за другом, щелкните мышью по первой, а затем, удерживая клавишу «*Shift*», щелкните по последней переменной.
- Чтобы выбрать группу переменных, которые расположены в списке не последовательно, щелкните мышью по первой переменной, а затем, удерживая клавишу «*Ctrl*», щелкните по следующей переменной и т.д.

.

Управляющие элементы диалоговых окон

В большинстве диалоговых окон SPSS есть пять стандартных кнопок:

OK. Запускает выполнение процедуры. После того как Вы выберете все переменные и параметры для процедуры, щелкните по OK, чтобы запустить процедуру и закрыть диалоговое окно.

Вставка. Генерирует команду синтаксиса на основе параметров, выбранных в диалоговом окне, и помещает ее текст в окно Редактора синтаксиса. После этого команду можно изменить и добавить в нее параметры, недоступные в диалоговых окнах.

Сброс. Все заданные в диалоговом окне параметры процедуры

изменяются на параметры по умолчанию, списки выбранных переменных очищаются.

Отмена. Все изменения, сделанные в диалоговом окне при последнем обращении к нему, отменяются, а окно закрывается. Параметры, выбранные в диалоговых окнах, сохраняются до окончания сеанса работы в SPSS, если Вы сами не решите их изменить.

Справка. Контекстно-зависимая справка. Эта кнопка открывает стандартное окно Справки, содержащее информацию о диалоговом окне.

4. Статистические данные для анализа

Имена переменных

Для имен переменных должны выполняться следующие правила:

- Имя каждой переменной должно быть уникальным; дублирующиеся имена не допускаются.
- Имена переменных могут иметь длиной до 64 байт (символов), первый символ в имени переменной должен быть буковой. Последующие символы могут быть любой комбинацией букв, чисел, точек (.) и не пунктуационных символов. (Шестьдесят четыре байта обычно означают 64 символа в однобайтовых языках (например, английский, французский, немецкий, русский, и др.) и 32 символа в двухбайтовых языках (например, японский, китайский, корейский). Много текстовых символов которые занимают только один байт в кодовой странице занимают два или больше байт в кодировке Unicode. Например, «й» это один байт в формате кодовой страницы , но занимает два байта в формате Unicode; так «гйsumй» это шесть байт в формате кодовой страницы и восемь байт в формате Unicode.)
- Имена переменных не могут содержать пробелов.
- Если имя переменной начинается с символа # - эта переменная является служебной.
- Создавать служебные переменные можно только при помощи синтаксиса.
- Если имя переменной начинается с символа $-эта переменная является системной. В задаваемых пользователем переменных нельзя задать символ $ в качестве первого символа в имени переменной.
- В именах переменных можно использовать точку, знак подчеркивания, а также символы $, # и @. Например, A._$@#1 - это

допустимое имя переменной.

- Следует избегать задания имен переменных, заканчивающихся точкой, поскольку точка в таких случаях может быть воспринята как символ окончания команды синтаксиса.
- Следует избегать имен переменных, заканчивающихся символом подчеркивания, поскольку возможен конфликт с именами, создаваемыми автоматически командами и процедурами.
- В именах переменных не могут использоваться зарезервированные ключевые слова. Зарезервированными словами являются: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO и WITH.
- Имена переменных могут состоять из любого сочетания символов в верхнем и в нижнем регистрах. Регистр сохраняется при отображении имен переменных.

Если длинное имя переменной нужно перенести в выводе на несколько строк, строка разрывается на символах подчеркивания, на точках и там, где регистр символов изменяется с нижнего на верхний.

Тип переменной

В диалоговом окне *Тип переменной* (рис.1.5) задается тип данных для каждой переменной. По умолчанию все новые переменные - числовые. Вы можете использовать диалоговое окно Тип переменной для изменения типа данных.

Содержимое диалогового окна *Тип переменной* зависит от выбранного типа данных. Для некоторых типов данных появляются поля для ввода ширины переменной и числа знаков после запятой, для других вы можете просто выбрать формат из списка с форматами.

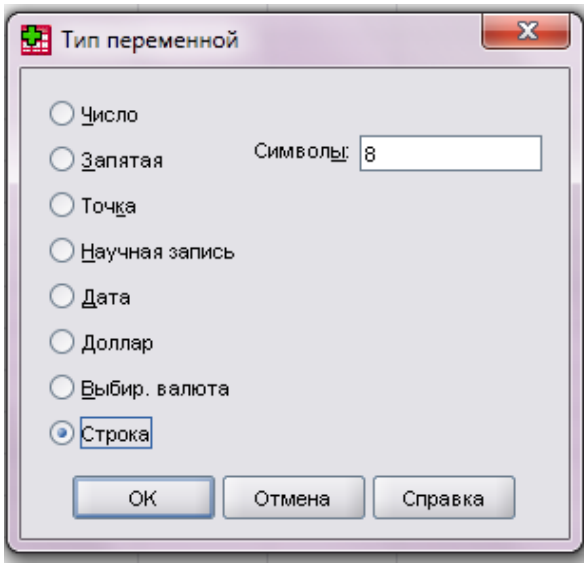


Рис.1.5. Диалоговое окно «Тип переменных»

Доступны следующие типы данных:

Числовая. Переменная, значения которой являются числами. Значения отображаются в стандартном числовом формате. При вводе данных Редактор данных принимает числовые значения в стандартном формате или в научной записи.

Запятая. Числовая переменная, значения которой отображаются с запятыми, разделяющими каждые три разряда, а для отделения дробной части используется точка. При вводе данных типа Запятая, Редактор данных принимает числа с запятыми или без запятых, или в форме научной записи. В значениях не могут содержаться запятые справа от десятичного разделителя.

Точка. Числовая переменная, значения которой отображаются с точками, разделяющими каждые три разряда, а для отделения дробной части используется запятая. При вводе данных типа с разделением групп цифр точками, Редактор данных принимает числа с точками или без точек, или в форме научной записи. В значениях не могут содержаться точки справа от десятичного разделителя.

Научная запись. Позволяет задать числовую переменную, значения которой выводятся с показателем степени, представленным буквой E, за которой идет знак и величина степени десятки. Редактор данных принимает в качестве таких переменных числовые значения, как со степенью, так и без. Показателю степени может предшествовать E или D, а также необязательный знак или только знак, например, 123, 1.23E2, 1.23D2, 1.23E+2, 1.23+2.

Дата. Числовая переменная, значения которой отображаются в одном из нескольких форматов календарной даты или времени. Формат выбирается из списка. Разделителями могут быть слэши, дефисы, точки, запятые или пробелы. Диапазон столетия при двузначном отображении года определяется установками в диалоговом окне Параметры (меню Правка, Параметры, вкладка Данные).

Доллар. Числовая переменная отображается со значком доллара вначале ($), точками, отделяющими группы по три разряда, и точкой в качестве десятичного разделителя. Значения данных можно вводить как со знаком доллара вначале, так и без него.

Выбираемая валюта. Числовая переменная, значения которой выводятся в одном из денежных форматов, заданного пользователем на вкладке Валюта диалогового окна Параметры. Заданные символы валюты нельзя использовать при вводе данных, однако они выводятся в Редакторе Данных.

Текстовая. Переменная, значения которой не являются числовыми, не может использоваться в вычислениях. Текстовая переменная может

содержать любые символы, однако их число не должно превышать заданную величину. Заглавные и строчные буквы считаются разными символами. Такие значения называют также буквенно-цифровыми.

Как задать типы переменных:

1. Щелкните по кнопке с тремя точками в ячейке, находящейся на пересечении столбца *Тип* и строки переменной, тип которой нужно изменить.
2. Выберите тип переменной в диалоговом окне *Тип переменной*.
3. Щелкните по *ОК*.

Выравнивание переменной

Выравнивание управляет выводом значений данных и/или меток значений в Редакторе данных. По умолчанию числовые переменные выровнены по правому краю, а текстовые переменные - по левому. Выравнивание влияет только на представление (вывод на экран) данных в Редакторе данных.

Шкала измерения переменной

Вы можете задать шкалу измерения переменной: количественную (числовые данные с интервальным или абсолютным уровнем измерения), порядковую, или номинальную. Номинальные и порядковые данные могут быть текстовыми (алфавитно-цифровыми) или числовыми.

- Номинальные. Переменную можно рассматривать как номинальную, когда ее значения представляют категории без естественного упорядочения, например, подразделение компании, где работает наемный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религиозную концессию.
- Порядковые. Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности обслуживанием от крайней неудовлетворенности до крайней удовлетворенности. Примеры порядковых переменных включают баллы, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение.
- Количественная. Переменную можно рассматривать как количественную, когда ее значения представляют упорядоченные категории с осмысленной метрикой, так что уместно сравнивать

расстояния между значениями. Примеры количественной переменной включают возраст в годах и доход в тысячах долларов.

Примечание: В текстовых порядковых переменных для отражения истинного порядка категорий принимается алфавитный порядок. Например, для строковой переменной со значениями низкий, средний, высокий интерпретируемый порядок категорий следующий: высокий, низкий, средний, что не соответствует правильному порядку. Вообще говоря, для представления порядковых данных более надежно использовать числовые коды.

Новым числовым значениям, созданным за время сессии, назначается количественная шкала измерения.

Значки, которые отображаются рядом с переменными в списках диалоговых окон, дают информацию о типе переменной и уровне измерения (табл.1.1).

Таблица 1.1

Значки типа данных, уровня измерения и списка переменных

Уровень измерения	**Тип данных**			
	Числовой	**Текстовый**	**Дата**	**Время**
Количественный		(не задается)		
Порядковый				
Номинальный				

Как просмотреть или изменить свойства переменных:

1. Активизируйте окно *Редактора данных*.
2. Дважды щелкните по имени переменной в заголовке столбца на вкладке *Данные*, или щелкните по вкладке *Просмотр переменных*.
3. Чтобы добавить новую переменную, введите имя переменной в любую пустую строку.
4. Выберите свойства, которые необходимо задать или изменить (рис.1.6).

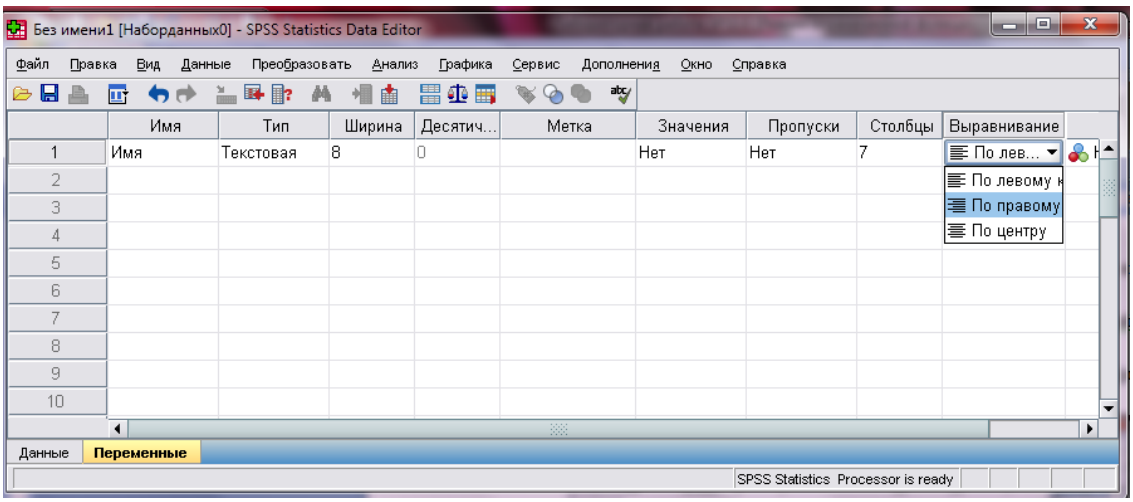


Рис.1.6. Изменение свойства «Выравнивание» переменной в редакторе данных на закладке «Переменные»

5. Ввод данных

Данные, используемые для анализа, хранятся в различных форматах, и в SPSS для Windows имеются средства, обеспечивающие доступ к большинству из них.

- Таблицы Excel и Lotus
- Таблицы баз данных из многих источников, включая Oracle, SQL Server, Access, dBASE и другие.
- Текстовые файлы данных.
- Файлы данных SPSS Statistics, созданные в других операционных системах.
- Файлы данных SYSTAT.
- Файлы данных SAS
- Файлы данных Stata

Вы можете непосредственно вводить данные в Редакторе данных на вкладке *Данные*. Вы можете вводить данные в любую ячейку. Можно вводить данные последовательно по каждому наблюдению или по переменным в выбранной области или в отдельные ячейки.

- Активная ячейка выделена жирной рамкой.
- Имя переменной и номер строки активной ячейки отображаются в левом верхнем углу Редактора Данных.
- Когда Вы выбираете ячейку и вводите значение, это значение отображается в редакторе значения ячеек в верхней части Редактора данных.
- Значения данных не записываются, пока Вы не нажмете на *Enter* или не перейдете в другую ячейку.
- Для переменных всех типов, кроме простых числовых, прежде

чем вводить данные, необходимо сначала задать тип переменной.

- Если вы вводите значение в пустой столбец, Редактор данных автоматически создает новую переменную и присваивает ей имя по умолчанию.

Формат ввода и формат представления

В зависимости от формата, представление значений в Редакторе данных может отличаться от того представления, в котором эти значения были введены в компьютер и хранятся в нем.

Ниже приведены несколько общих правил:

- Для числовых переменных, переменных с запятой и с точкой Вы можете вводить значения с любым числом десятичных позиций (до 16-ти), и все значения будут сохраняться в компьютере. На вкладке Данные отображается только заданное количество знаков и отображаемые числа, имеющие большее число знаков после запятой, округляются. Однако, во всех расчетах используются полные значения со всеми десятичными знаками.
- Для строковых переменных все переменные дополняются справа пробелами до заданной ширины. В текстовых переменных с максимальной шириной три знака значение Да хранится как 'Да ' и не оно не эквивалентно значению ' Да'.
- Для формата дата можно использовать слеш, тире, пробелы, подчеркивания, запятые или точки в качестве разделителей между значениями дня, месяца и года. Можно вводить числа, трехбуквенные аббревиатуры или полные названия месяцев. Даты формата dd-mmm-yy отображаются с тире в качестве разделителей и трехбуквенными аббревиатурами для представления месяца. Данные формата dd/mm/yy и mm/dd/yy отображаются с слешем в качестве разделителей и числами для представления месяца. Внутреннее представление даты - это время в секундах, прошедшее с 14 октября 1582 года. Диапазон веков для дат, состоящих из двух знаков, представляющих год, задается в настройках SPSS (в меню Правка, Параметры, вкладка Данные).
- Для ввода данных в форматах времени Вы можете использовать двоеточия, разделители периодов (или запятые, или точки) или пробелы для разделения часов, минут и секунд. Время отображается с двоеточиями в качестве разделителя. Внутренне значения времени хранятся как число секунд, представляющее

интервал времени. Например, 10:00:00 внутренне храниться как 36000, что представляет собой 60 (число секунд в минуте) * 60 (число минут в часе) * 10 (часов).

5.1. Сохранение файлов данных в формате Excel

Вы можете сохранить данные в одном из трех форматов Microsoft Excel. Excel 2.1, Excel 97 и Excel 2007.

В Excel 2.1 и Excel 97 ограничивается количество столбцов — 256; так что сохраняются только первые 256 переменные.

В Excel 2007 ограничения на столбцы составляет 16000, так что сохраняются только первые 16000 переменные.

В Excel 2.1 ограничения на строки составляет 16 384, так что сохраняются только первые 16 384 наблюдения.

В Excel 97 и Excel 2007 также имеются ограничения на количество строк на листе, однако книга может иметь несколько страниц, и при превышении предела для одной страницы создается несколько страниц.

6. Работа с выводом результатов. Окно Viewer

Результаты выполнения процедур SPSS выводятся в окно, называемое *Viewer* (*Вывод*) (рис.1.7). В этом окне Вы сможете с легкостью найти именно ту часть вывода результатов, которая вам нужна. Вы можете управлять выводом и создавать документы, содержащие в точности такой вывод, который Вам требуется, структурированный и отформатированный должным образом.

Вы можете использовать *Вывод* для:

- Обзора результатов.
- Показать или скрывать выбранные таблицы и диаграммы.
- Изменить порядок следования результатов, перемещая элементы.
- Переместить объекты между средством просмотра и другими приложениями.

Окно *Viewer* делится на две панели:

- В левой панели находится структура содержимого.
- Правая панель содержит сами результаты – статистические таблицы, диаграммы и текстовой вывод.

Если щелкнуть по объекту в структуре, можно перейти непосредственно к соответствующей таблице или диаграмме. Вы можете передвигать правую границу схемы вывода, чтобы изменить ширину схемы вывода.

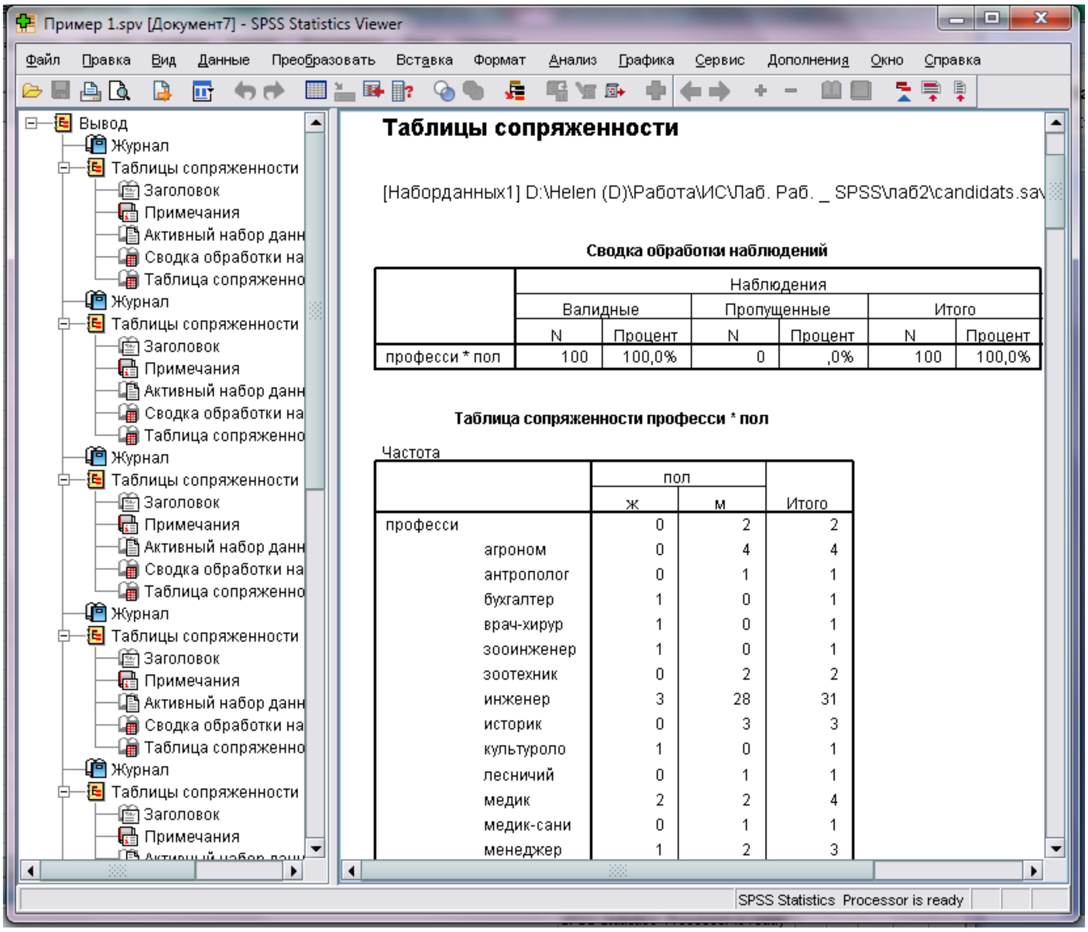


Рис.1.7. Окно «Viewer» («Вывод»)

Сохранение результатов

Содержимое *Viewer* можно сохранить как документ *Вывод*. Сохраненный документ будет включать обе панели окна *Viewer* (схему и результаты).

Сохранение документа *Вывод*:

1. В меню окна *Viewer* выберите: *Файл/Сохранить*.
2. Выберите папку, в которой будете сохранять документ.
3. Введите имя документа и нажмите на *Сохранить*.

7. Получение справки

В SPSS есть несколько видов Справки:

Меню *Справка*

При помощи меню *Справка* можно получить доступ к основной Справке, Учебнику и техническим справочным материалам.

- *Темы.* Этот пункт меню Справка позволяет получить доступ к закладкам Содержание, Указатель, Поиск, которые позволяют находить конкретные темы Справки.

- *Учебник.* Иллюстрированные пошаговые инструкции по использованию многих базовых возможностей. Совсем не обязательно изучать весь Учебник целиком, от начала до конца. Можно выбирать темы для изучения, просматривать темы в любом порядке и использовать указатель и содержание для поиска нужных тем.
- *Примеры анализа.* Практические примеры выполнения различных типов анализа и интерпретации полученных результатов. Поскольку используемые для примеров файлы данных находятся на жестком диске, Вы можете выполнять все шаги, описанные в примерах, и наглядно видеть, как были получены те или иные результаты. Процедуру для изучения можно выбрать в содержании или найти в указателе.
- *Репетитор по статистике.* Конструктор, позволяющий находить процедуру анализа, подходящую для использования. После того, как Вы за несколько шагов выберете подходящие варианты, Репетитор по статистике откроет диалоговое окно процедуры, соответствующей выбранным Вами критериям. Репетитор по статистике позволяет получить доступ к большинству процедур анализа и создания отчетов в SPSS, а также ко многим графическим процедурам.
- *Руководство по синтаксису.* Подробная информация о синтаксисе содержится в справочной системе, а также в отдельном документе - Руководстве по синтаксису.
- *Алгоритмы.* Алгоритмы большинства статистических процедур доступны в двух формах: в справочной системе, а так же в отдельном документе PDF.
- *Контекстно-зависимая справка.* Во многих местах в интерфейсе можно вызвать контекстно-зависимую справку.

Кнопки Справка в диалоговых окнах

В большинстве диалоговых окон есть кнопки *Справка*, прямо выводящие к теме Справки, которая относится именно к этому диалоговому окну. Темы справки предоставляют общую информацию и ссылки на аналогичные темы.

Синтаксис

Находясь в редакторе синтаксиса, поместите курсор в любое место в команде и нажмите клавишу *F1*. На экране появится синтаксическая

диаграмма команды. Полная документация по синтаксису вызывается щелчком ссылки в списке связанных тем или в закладке *Содержание* окна *Справка*.

8. Основные шаги в анализе данных

Анализировать данные при помощи SPSS Statistics совсем несложно. Все, что необходимо сделать, это:

1. Загрузите свои данные в SPSS Statistics. Можно открыть файл, сохраненный ранее в SPSS Statistics, считать файл электронной таблицы, базу данных или текстовый файл или ввести данные непосредственно в редакторе данных.
2. Выбрать процедуру. Выбрать в меню процедуру для расчета статистик или создания диаграммы.
3. Выбрать переменные для анализа. Переменные файла данных отображаются в диалоговом окне процедуры.
4. Запустить процедуру и посмотреть результат. Результаты выводятся в окне *Viewer*.

9. Контрольные вопросы

1. Для чего предназначен пакет SPSS Statistic?
2. Какие типы окон есть в программе?
3. Что собой представляет Редактор данных и какие вкладки он содержит?
4. Какие существуют правила присвоения имен переменным?
5. Какие существуют типы и шкалы измерений переменных?
6. Каким правилам необходимо придерживаться при вводе данных?
7. Где можно увидеть результаты выполнения процедур?
8. Какие существуют виды Справки в программе SPSS Statistic?
9. Каковы основные шаги в анализе данных?

Лабораторная работа №2

Тема: Первичная обработка статистических данных в пакете SPSS Statistics 17.0. Одномерный анализ.

Цель: Формирование умений первичной обработки статистических данных. Выработка навыков работы в статистическом пакете SPSS Statistics 17.0.

1. Введение

Для того чтобы выполнить статистическое исследование, необходима научно-обоснованная информационная база. Она формируется в результате статистического наблюдения, которое является начальной стадией экономико-статистического исследования.

Статистическим наблюдением называется планомерный научно-обоснованный сбор данных или сведений о социально-экономических явлениях и процессах. В процессе обработки, анализа статистические данные становятся информацией. Не все данные, факты, собранные в процессе наблюдения, могут быть использованы для дальнейшего исследования. Они должны отвечать определенным требованиям. Важнейшими требованиями являются достоверность данных, сопоставимость данных или единообразие.

В данной работе проводится описательный (дескриптивный) анализ отдельных переменных. К нему относятся создание частотной таблицы, вычисление статистических характеристик и графическое представление.

2. Построение частотных таблиц

Полученные статистические данные необходимо ввести в первый столбец редактора данных.

Чтобы указать наименование столбца и тип вводимых данных необходимо в редакторе данных (окно SPSS Statistics Data Editor) (рис.2.1) дважды щелкнуть на ячейке с надписью *пер* или щелкнуть на вкладке *Переменные* на нижнем краю таблицы.

В обоих случаях вы перейдете в режим просмотра переменных.

Чтобы задать имя переменной, введите в текстовом поле *Имя* выбранное имя переменной. Затем нажмите клавишу <*Tab*>, чтобы подтвердить ввод и перейти к установке типа переменной.

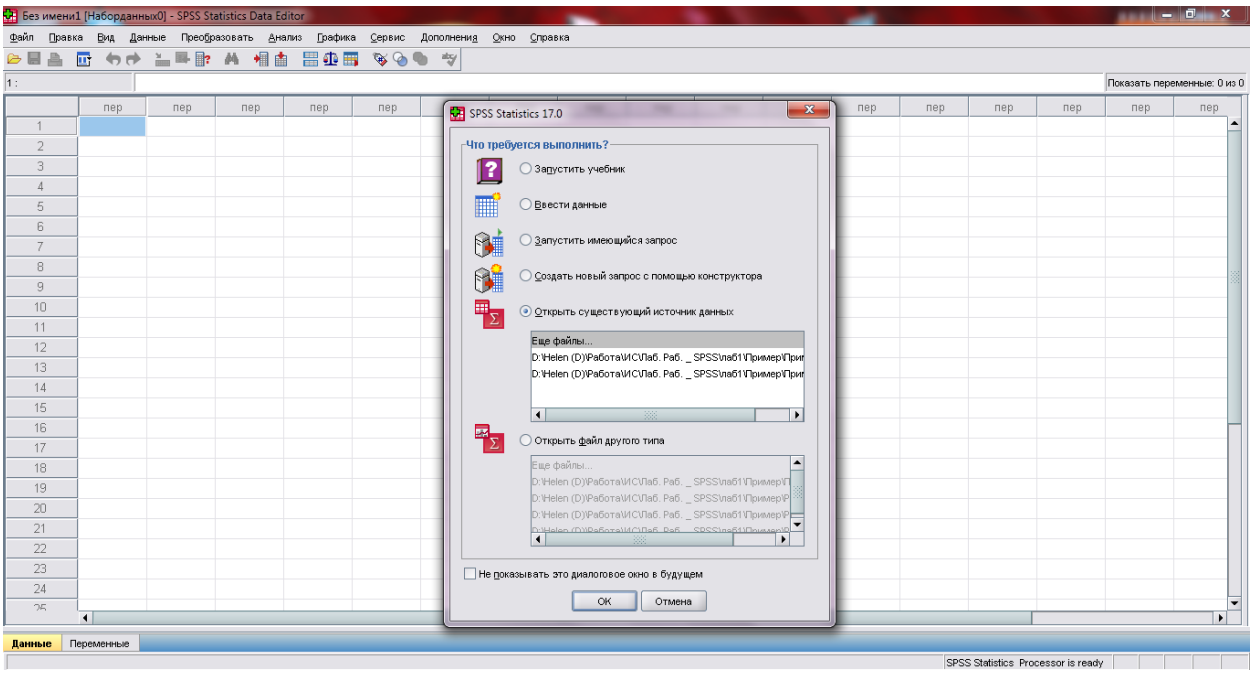


Рис.2.1. Окно SPSS Statistics Data Editor

Для построения вариационного ряда в меню *Данные* следует выбрать пункт *Сортировать наблюдения*. Эта функция упорядочивает значения случайной величины в порядке возрастания или убывания (рис.2.2).

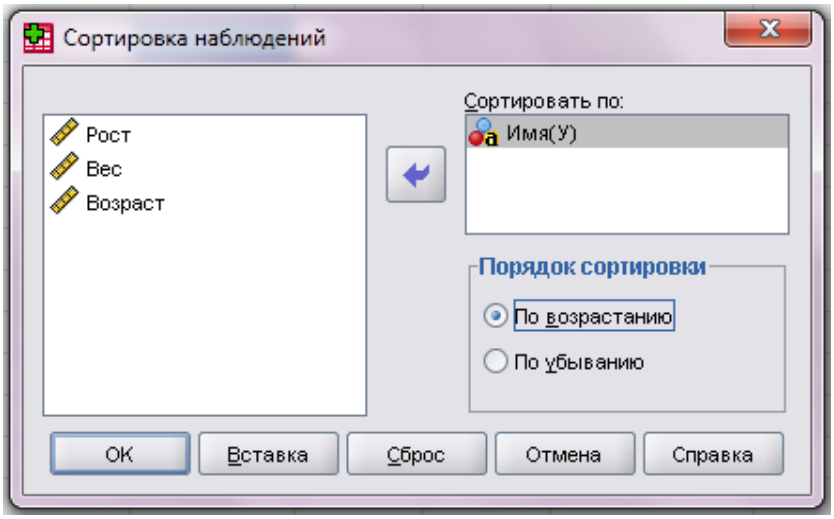


Рис.2.2. Диалоговое окно «Сортировка наблюдений»

Первым этапом статистического анализа данных, как правило, является частотный анализ. Для построения частотной таблицы (в виде ряда распределения частот) применительно к вашим данным выберите в меню команды *Анализ/Описательные статистики/Частоты.*

В появившемся диалоговом окне *Частоты* (рис.2.3) выберите опцию *Вывести частотные таблицы*, кнопкой со стрелочкой перенесите изучаемую переменную в список выходных параметров и подтвердите

операцию кнопкой *OK*.

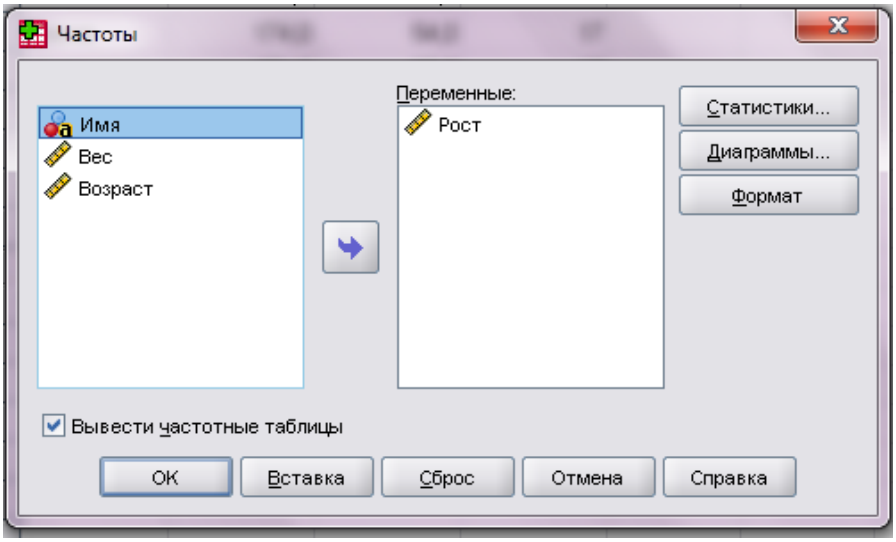


Рис.2.3. Диалоговое окно «Частоты»

В окне просмотра результатов *Вывод* отобразится таблица частот. Перед самой частотной таблицей выводится небольшая таблица с обзором допустимых и отсутствующих значений.

Чтобы вывести частотную таблицу, отсортированную по убыванию частоты, поступите следующим образом:

- выберите в меню команды *Анализ/Описательные статистики/Частоты*;
- перенесите рассматриваемую переменную в список выходных переменных;
- при активной опции *Вывести частотные таблицы* щелкните на кнопке *Формат*. Откроется диалоговое окно *Частоты:Формат* (рис.2.4).

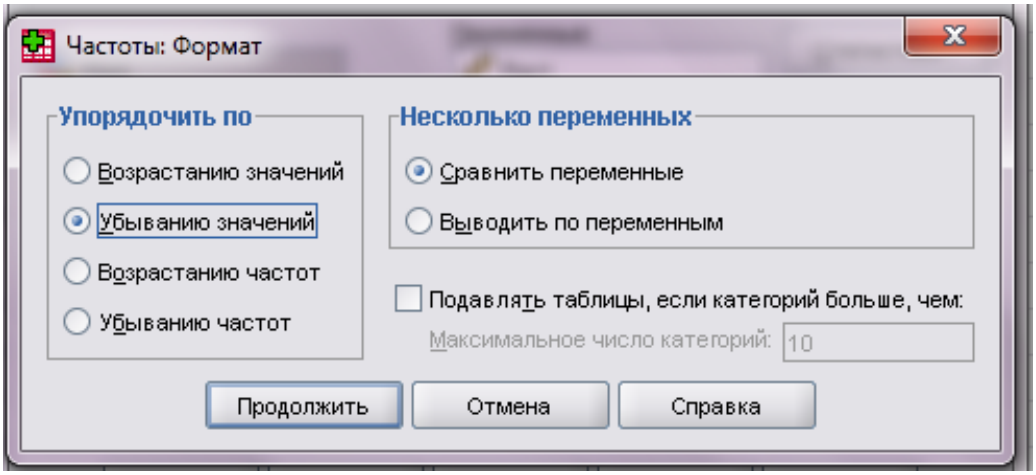


Рис.2.4. Диалоговое окно Frequencies:Format

В группе *Упорядочить по* выберите порядок, в котором будут отображены значения в частотной таблице. Возможны следующие варианты:

- По возрастанию значений. Это настройка по умолчанию.
- По убыванию значений.
- По возрастанию частот.
- По убыванию частот.

Кроме того, флажок *Подавлять таблицы, если категорий больше, чем:* позволяет избежать вывода длинных частотных таблиц.

- Выберите *по Убыванию значений*.
- Подтвердите выбор кнопкой *Продолжить*.
- Щелкните *OK*, чтобы начать вычисление. В результате частоты в таблице будут расположены в порядке убывания.

3. Вывод статистических характеристик

Следующим этапом частотного анализа данных является получение описательной статистики. Для вычисления статистических характеристик случайной величины необходимо выполнить следующее: в диалоге *Частоты* щелкнуть на кнопке *Статистики*. Откроется диалоговое окно *Частоты:Статистики* (рис.2.5).

В группе *Значения процентилей* можно выбрать следующие варианты:

- *Квартили.* Будут показаны первый, второй и третий квартили.
- *Процентили для* (Точки раздела): Будут вычислены значения процентилей, разделяющие выборку на группы наблюдений, которые имеют одинаковую ширину, то есть включают одно и то же количество измеренных значений. По умолчанию предлагается количество групп 10. Если задать, например, 4, то будут показаны квартили, то есть квартили соответствуют процентилям 25,50 и 75. видно, что число показываемых процентилей на единицу меньше заданного числа групп.
- *Процентили.* Здесь имеются в виду значения процентилей, определяемые пользователем.

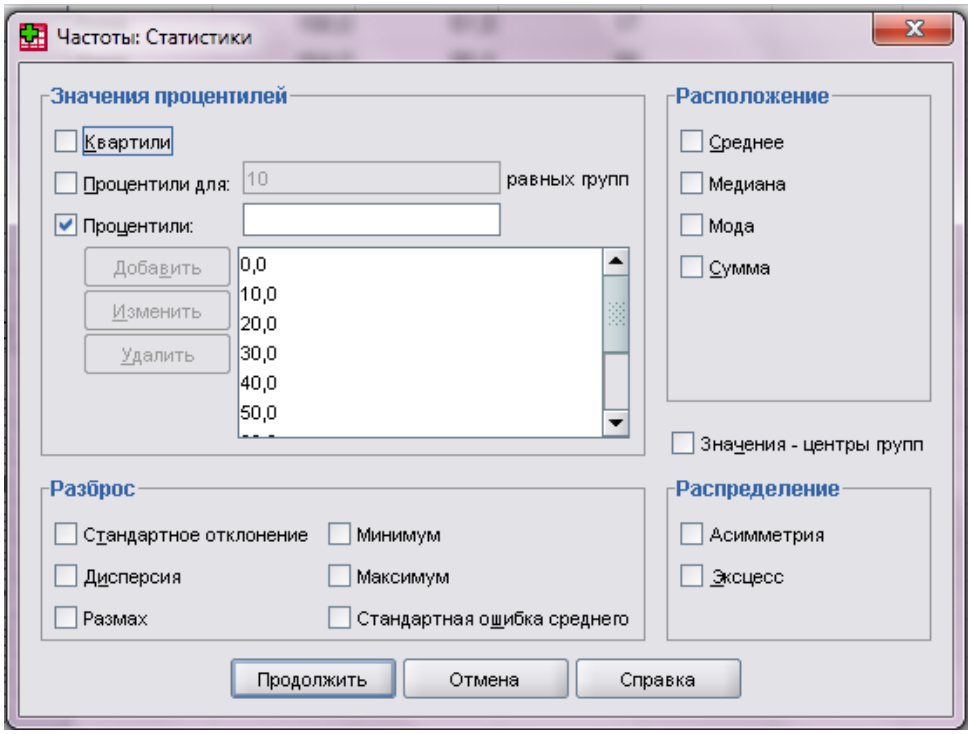


Рис.2.5. Диалоговое окно «Частоты: Статистики»

Введите значение процентиля в пределах от 0 до 100 и щелкните на кнопке *Добавить*. Повторите эти действия для всех желаемых значений процентилей.

В группе *Разброс* можно выбрать меры разброса:

- *Стандартное отклонение.* Оно равно квадратному корню из дисперсии. В интервале шириной, равной удвоенному стандартному отклонению, который отложен по обе стороны от среднего значения, располагается примерно 67% всех значений выборки, подчиняющейся нормальному распределению.
- *Дисперсия.* Она определяется как сумма квадратов отклонений всех измеренных значений от их среднеарифметического значения, деленная на количество измерений минус 1. (В SPSS встроена функция исправленной дисперсии, поэтому в знаменателе присутствует «минус 1»).
- Размах – разница между наибольшим значением (максимумом) и наименьшим значением (минимумом).
- *Минимум.*
- *Максимум.*
- *Стандартная ошибка* среднего значения. В интервале шириной, равной удвоенной стандартной ошибке, отложенному вокруг среднего значения, располагается среднее значение генеральной совокупности с вероятностью примерно 67%. Стандартная

ошибка определяется как стандартное отклонение, деленное на квадратный корень из объема выборки.

В группе *Расположение* можно выбрать следующие характеристики:

- *Среднее значение.* Это арифметическое среднее измеренных значений; оно определяется как сумма значений, деленная на их количество.
- *Медиана* – это точка на шкале измеренных значений, выше и ниже которой лежит по половине всех измеренных значений.
- *Мода* – это значение, которое наиболее часто встречается в выборке. Если распределение имеет несколько мод, то говорят, оно мультимодально или многомодально (имеет два и более «пика»).
- *Сумма* всех значений.

В группе *Распределение* можно выбрать следующие меры несимметричности распределения:

- *Ассиметрия* (коэффициент ассиметрии) – это мера отклонения распределения частоты от симметричного распределения, то есть такого, у которого на одинаковом удалении от среднего значения по обе стороны выборки данных располагается одинаковое количество значений. Если наблюдения подчиняются нормальному распределению, то асимметрия равна нулю. Для проверки на нормальное распределение можно применять следующее правило: если асимметрия значительно отличается от нуля, то гипотезу о том, что данные взяты из нормально распределенной генеральной совокупности, следует отвергнуть. Если вершина асимметричного распределения сдвинута к меньшим значениям, то говорят о положительной асимметрии, в противоположном случае – об отрицательной.
- *Эксцесс* (Коэффициент вариации) – указывает, является ли распределение пологим (при большом значении коэффициента) или крутым. Эксцесс равен нулю, если наблюдения подчиняются нормальному распределению. Поэтому для проверки на нормальное распределение можно применять ещё одно правило: если коэффициент вариации значительно отличается от нуля, то гипотезу о том, что данные взяты из нормально распределенной генеральной совокупности, следует отвергнуть.

4. Графическое представление данных

В п.1.2 была описана методика построения частотных таблиц.

Однако результаты частотного распределения можно представить графически.

Чтобы создать столбчатую диаграмму для частотного распределения:

- Выберите в меню команды *Анализ/Описательные статистики/Частоты*
- Перенесите нужную переменную в список выходных переменных.
- Щелкните на кнопке *Диаграммы*. Откроется диалоговое окно *Частоты: Диаграммы* (рис.2.6).
- Выберите в группе *Тип диаграммы* пункт *Столбиковые*, а в группе *Значения на диаграмме* - пункт *Проценты*. Подтвердите выбор кнопкой *Продолжить*. Вы вернетесь в диалог *Частоты*.

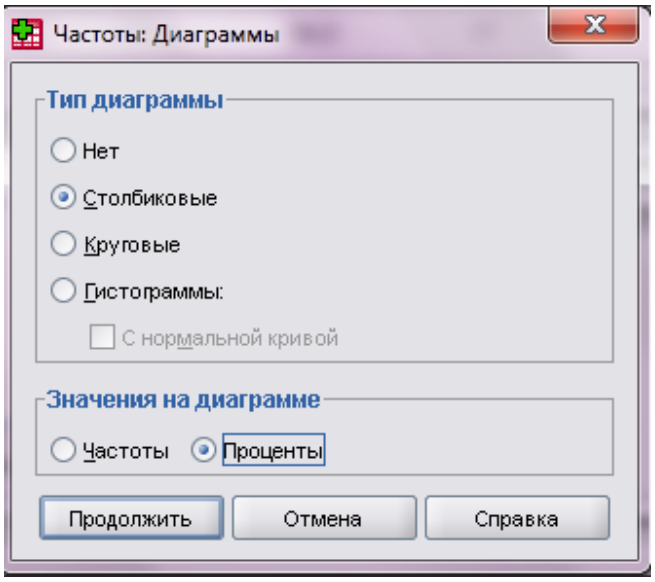


Рис.2.6. Диалоговое окно Частоты: Диаграммы

- В диалоговом окне *Частоты* снимите флажок *Вывести частотные таблицы*.
- Щелкните *OK*. Диаграмма будет показана в окне просмотра.

Чтобы усовершенствовать вид диаграммы, выполните следующие действия.

- Чтобы начать редактирование, дважды щелкните в области диаграммы. Диаграмма будет показана в *Редакторе диаграмм*.
- На панели инструментов *Редактора диаграмм* щелкните на кнопке меток столбцов  .

Откроется диалоговое окно *Свойства*. (Изучите все вкладки данного окна самостоятельно.) После чего появятся надписи на всех столбцах диаграммы.

После щелчка мышью на любом столбце об будет обведен голубой линией. Это значит, что области столбцов готовы для редактирования, для чего необходимо открыть окно *Свойства* двойным кликом мыши. (Изучите все вкладки данного окна самостоятельно.) Таким же образом вызывается окно редактирования всей диаграммы в целом, для этого необходимо дважды щелкнуть мышью в свободной области диаграммы и откроется окно *Свойства*. (Изучите все вкладки данного окна самостоятельно.)

Изменить заголовок диаграммы можно так еж, если дважды на нем щелкнуть. (Изучите все вкладки данного окна самостоятельно.)

5. Пример

В группе пациентов проводились измерения показателей систолического (верхнего) и диастолического (нижнего) кровяного давления, представленные в таблице 2.1.

Таблица 2.1

Показатели систолического и диастолического кровяного давления

Систолическое	154	136	91	125	133	125	93	80	132	107	142	115	114	120	141
Диастолическое	108	90	54	89	93	77	43	50	125	76	96	74	79	71	90

Для представленной информации нет смысла выводить таблицы частот, т.к. невооруженным глазом видно, что более двух раз ни одно из значений не встречается, иначе получим длинную частотную таблицу.

Определим статистические характеристики: первый, второй и третий квартили, стандартное отклонение, минимум, максимум, среднее значение – для каждой выборки отдельно. Получим следующие результаты (рис.2.7):

Статистики

		Систолическое	Диастолическое
N	Валидные	15	15
	Пропущенные	0	0
Среднее		120,53	81,00
Стд. отклонение		20,866	21,693
Минимум		80	43
Максимум		154	125
Процентили	25	107,00	71,00
	50	125,00	79,00
	75	136,00	93,00

Рис.2.7. Статистические характеристики систолического и диастолического кровяного давления

Таким образом, в наблюдаемой группе пациентов среднее значение давления близко к показателю 120/80, что по принятым стандартам является нормальным. Разброс (стандартное отклонение) показателей верхнего давления относительно среднего значения (120,53) составил 20,87 единиц, а нижнего относительно своего среднего значения (81) – 21,69 единиц. Это значит, что нижнее давление менее стабильно.

Минимальное и максимальное систолическое давление соответственно равны 80 и 154, в то время как диастолическое колеблется в пределах от 43 до 125. У 50 % респондентов исследуемые показатели близки к оптимальным, о чем свидетельствуют значения 25-й и 75-й процентилей, т.е. у половины пациентов верхнее давление находится в диапазоне 107-136, а нижнее – 71-93.

Заметим, что медиана и 50-я процентиль (вторая квартиль) – это одно и тоже, поэтому выбирая опцию *Квартили*, нет смысла выбирать ещё и *Медиана*, т.к. произойдет повторное вычисление одной и той же статистической характеристики.

6. Общие задания

1. Получите у преподавателя номер индивидуального задания.
2. Определите, к какой статистической шкале можно отнести ваши данные. Обоснуйте свое решение.
3. Введите данные в столбцы редактора данных, присвойте имя, тип и другие необходимые для ваших переменных характеристики.
4. Постройте таблицу частот и по полученной таблице сделайте первичные выводы относительно ваших данных.
5. Вычислите статистические характеристики и дайте анализ полученным результатам.
6. Для исходных данных постройте столбчатую диаграмму, а также гистограмму с выводом графика нормального распределения.
7. Усовершенствуйте их с помощью редактора диаграмм.

7. Требования к отчету

Отчет по лабораторной работе №2 предоставляется в печатном виде и должен содержать:

1. Теоретический материал.
2. Исходные данные индивидуального задания.
3. Результаты выполнения практических заданий с предоставлением необходимых расчетов, графиков и т.д.
4. Выводы по полученной информации.

После предоставления отчета, студент подтверждает знание изучаемого материала и защищает лабораторную работу за компьютером.

8. Индивидуальные задания

1. Проведите сравнительный анализ цен двух акций, измеряемых в течение одного дня.

Акция 1:

10.628	14.425	14.164	19.049	19.181
13.446	16.322	10.103	9.984	17.958
11.444	2.778	18.964	18.039	21.932
13.690	13.731	17.262	13.387	20.942

Акция 2:

20.076	16.805	18.242	16.048	15.297
18.524	21.308	11.826	18.065	16.467
14.697	26.415	16.810	21.020	21.320
15.369	11.022	21.181	15.252	16.628

2. Исследовать показатели дохода и расхода на некотором предприятии.

№	Доход	Расход	№	Доход	Расход	№	Доход	Расход
1	30000	20000	8	51700	50000	15	44000	20000
2	20000	15000	9	48200	45000	16	50000	55000
3	28100	28000	10	50000	25000	17	48250	45000
4	44500	50000	11	28250	20000	18	60000	60000
5	30000	31000	12	26100	20000	19	61000	62000
6	31100	30000	13	31450	31400	20	60500	62000
7	68500	30000	14	32100	30000			

3. Проанализировать урожайность овощей (ц/га):

Год	Урожайность	Год	Урожайность	Год	Урожайность	Год	Урожайность
1964	102,3	1973	137	1982	203	1991	104,5
1965	93,9	1974	166	1983	163	1992	103
1966	113	1975	167,4	1984	191,7	1993	78
1967	104,3	1976	175,8	1985	174	1994	79,7
1968	126	1977	171	1986	189	1995	50,8
1969	104	1978	202,3	1987	188	1996	41
1970	131	1979	207,8	1988	166,4	1997	37
1971	102,2	1980	206	1989	144	1998	61
1972	137,9	1981	142	1990	173		

4. Провести первичную обработку статистических данных. Величина денежного вознаграждения.

10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

5. Провести первичную обработку статистических данных по двум выборкам. Индексы загрязнения атмосферы:

№ поста\месяц	1	2	3	4	5	6	7	8	9	10	11	12
1	2,60	2,04	1,60	2,96	2,51	2,97	2,55	2,32	1,77	1,93	1,70	2,86
2	2,51	2,27	2,33	2,34	2,95	2,41	2,91	2,94	2,14	2,75	1,07	2,05

6. Провести первичную обработку следующих статистических данных.

Показатели	Март	Апрель	Май	Июнь	Июль	Август
Приход (руб.)	32550	33038.25	33533.82	34035.83	34547.38	35065.59
Затраты на товары (руб.)	19316	19490	19665	19842	20021	20201

7. Имеются сведения о заготовке сена и надоях молока по хозяйствам района. Провести первичный анализ.

№ хозяйства	Заготовлено сена (т)	Надои молока в среднем от коровы (кг)	№ хозяйства	Заготовлено сена (т)	Надои молока в среднем от коровы (кг)
1	135	6,2	10	150	6,5
2	120	5,5	11	201	4,2
3	315	9,0	12	50	6,3
4	171	4,1	13	45	3,8
5	169	3,5	14	117	5,2
6	135	6,3	15	35	2,0
7	61	5,3	16	17	5,8
8	131	3,5	17	21	6,1
9	72	4,7			

8. Проанализировать данный временной ряд. Реализация масла сливочного (тыс.т.)

Месяцы	Годы					
	1	2	3	4	5	6
Январь	44	42	43	44	43	46
Февраль	45	40	41	43	40	44
Март	49	42	42	43	44	47
Апрель	47	41	42	44	45	49
Май	38	36	35	39	41	43
Июнь	37	35	36	38	40	39
Июль	39	37	37	41	43	44
Август	41	40	39	42	42	45
Сентябрь	47	48	46	47	49	51
Октябрь	52	51	50	55	52	56
Ноябрь	49	50	47	50	61	60
декабрь	50	45	46	52	54	63

9. Статистически исследуйте последние три столбца в штатном расписании административно- управленческого персонала.

Должность	***Кол-во***	***оклад, грн.***	***годовой ФОТ, грн.***	***Отчисления, грн.***
Генеральный директор	1	900	10800	4050
Технический директор	1	800	9600	3600
Экономист – маркетолог	1	800	9600	3600
Главный бухгалтер	1	750	9000	3375
консультант	3	500	6000	2250
фотограф	2	550	6600	2475
механик	1	173	2076	778,5
уборщица	2	150	1800	675
ИТОГО	12		55476	20803,5

10. Сравните разброс коррозии труб с антикоррозийным покрытием и труб без покрытия.

С покрытием	39	43	43	52	52	59	40	45	47	62	40	27
Без покрытия	42	37	61	74	55	57	44	55	37	70	52	55

11. Проведите первичную обработку данных по заработной плате.

Должности		Количество (чел.)	Средняя ЗП (в месяц)	Фонд ЗП
1.	Директор	1	3 000	3000
2.	Главный инженер	1	2 600	2600
3.	Главный технолог	1	2 400	2400
4.	Зам директора по коммерции	1	2 000	2000
5.	Главный бухгалтер	1	1 900	1900
5.	Экономист (маркетолог)	1	1 400	1400
7.	Конструктор	1	1 500	1500
8.	Инженер по снабжению	1	1 400	1400
9.	Механик	1	1 400	1400
10.	Электрик	1	1 400	1400
11.	Бухгалтер	1	800	800
12.	Секретарь	1	600	600
13.	Кладовщик	1	900	900
14.	Участок лесопиления	6	1 100	6600
15.	Участок сушки	2	1 100	2200
16.	Участок профилирования	2	1 100	2200
17.	Участок упаковки	1	850	850
18.	Участок сортировки	1	900	900
19.	Участок столярных изделий	2	900	1800
Итого		27		44850

12. Посредством описательных статистик исследуйте эффективность рекламной кампании, рассмотрев следующие данные о ежедневном обороте фирмы за 15 дней до и после публикации рекламы.

№	До	После	№	До	Досле	№	До	После
1	100,744	115,842	6	87,659	122,346	11	120,777	110,822
2	102,497	99,655	7	109,510	117,175	12	113,219	95,506
3	81,305	98,710	8	102,047	113,978	13	77,832	121,507
4	105,532	120,969	9	97,978	101,172	14	97,831	90,734
5	96,715	101,850	10	89,960	115,694	15	97,105	114,223

13. Проведите первичную обработку данных по оценкам тестов чтения и арифметики.

Чтение	43	58	45	53	37	58	55	61	46	64	46	62	60	56
Арифметика	32	25	28	30	22	25	22	20	20	30	21	28	34	28

14. Провести первичную обработку статистических данных.

Рост девушек, см.

165	170	166	160	154	163	161
164	158	159	163	157	164	170
163	166	162	164	161	175	167
166	178	164	167	163	173	174
164	167	165	164	165	165	161
164	168	158	168	152	172	164
162	173	166	161	164	165	163

15. Провести первичную обработку статистических данных.

Рост юношей, см: 182, 183, 168, 174, 165, 174, 163, 168, 179, 185, 171, 174, 180, 175, 179, 181, 169, 184, 172, 174.

16. Исследуйте динамический ряд среднегодовых удоев молока от одной коровы.

Год	Удой молока, кг	Год	Удой молока, кг	Год	Удой молока, кг
1961	2532	1969	3177	1977	3648
1962	2317	1970	3181	1978	3475
1963	2341	1971	3201	1979	3475
1964	2513	1972	3192	1980	3579
1965	2968	1973	3156	1981	3473
1966	2956	1974	3364	1982	3385
1967	3041	1975	3489	1983	3701
1968	3182	1976	3587	1984	3854

17. Обработайте данные. Время реакции на свет и на звук, в миллисекундах.

Номер испытуемого	Время реакции на звук	Время реакции на свет	Номер испытуемого	Время реакции на звук	Время реакции на свет
1	223	181	10	191	156
2	104	194	11	197	178
3	209	173	12	183	160
4	183	153	13	174	164
5	180	168	14	176	169
6	168	176	15	155	155
7	215	163	16	115	122
8	172	152	17	163	144
9	200	155			

18. Исследуйте динамический ряд выплавки стали с 1960 по 1979гг.

Год	Выплавка стали, млн. т	Год	Выплавка стали, млн. т	Год	Выплавка стали, млн. т
1960	65,3	1967	102,2	1973	131,5
1961	70,8	1968	106,5	1974	136,2
1962	76,3	1969	110,3	1975	141,3
1963	80,2	1970	115,9	1976	144,8
1964	85,0	1971	120,7	1977	146,7
1965	91,0	1972	125,6	1978	151,5
1966	96,9				

19. Проанализируйте данный временной ряд. Реализация масла сливочного (тыс. т.).

Месяцы	**Годы**					
	1	**2**	**3**	**4**	**5**	**6**
1	44	42	43	44	43	46
2	45	40	41	43	40	44
3	49	42	42	43	44	47
4	47	41	42	44	45	49
5	38	36	35	39	41	43
6	37	35	36	38	40	39
7	39	37	37	41	43	44
8	41	40	39	42	42	45
9	47	48	46	47	49	51
10	52	51	50	55	52	56
11	49	50	47	50	61	60
12	50	45	46	50	54	63

20. Известны средние издержки (СИЗ) и товарооборот (ТР) фирмы за некоторый промежуток времени. Проведите первичную обработку этой информации.

СИЗ	ТР	СИЗ	ТР	СИЗ	ТР	СИЗ	ТР
4,009	13	4,052	12	4,089	13	4,114	13
4,022	13	4,056	12	4,101	13	4,114	12
4,033	12	4,060	12	4,102	14	4,126	13
4,048	12	4,066	14	4,109	12	4,133	14
4,051	13	4,082	14	4,112	14	4,135	13

9. Контрольные вопросы

1. Что понимают под статистическим наблюдением?
2. Как построить частотную таблицу?
3. Какие характеристики данных относятся к описательным статистикам?
4. Что понимается под квартилью?
5. Что такое первый, второй и третий квартили? Что они показывают?
6. Как задаются процентили?
7. Что понимается под модой и медианой?
8. Объясните процесс графического представления данных.

Лабораторная работа №3

Тема: Таблицы сопряженности.

Цель: Сформировать практические навыки определения взаимосвязи между переменными посредством создания таблиц сопряженности.

1. Введение

В лабораторной работе №2 рассматривались только отдельные переменные, т.е. проводился одномерный анализ данных. Перейдем к двумерному анализу, т.е. будем определять, существует ли взаимосвязь между двумя и более переменными.

Исследуем зависимость между двумя переменными. Связь между неметрическими переменными (т.е. переменными, относящимися к номинальной шкале или к порядковой шкале с не очень большим количеством категорий) лучше всего представить в форме таблиц сопряженности.

2. Создание таблиц сопряженности

Рассмотрим файл *candidats.sav* (данные по списку кандидатов в депутаты Парламента Республики Молдова от ПКРМ на выборах 2005г.).

Загрузите файл с вашими данными. Для создания таблиц сопряженности и вычисления на их основе меры связанности, выберите меню команды *Анализ*/*Описательные статистики/Таблицы сопряженности.*

Откроется диалоговое окно *Таблицы сопряженности*. Здесь в списке исходных переменных можно выбрать переменные для строк и столбцов таблицы сопряженности. Для каждого сочетания двух переменных будет создана таблица сопряженности. Например, если в списке *Строки* находится три переменных, а в списке *Столбцы* – две, то мы получим 3x2=6 таблиц сопряженности.

Построим таблицу сопряженности из переменных «*пол*» и «*профессия*». Чтобы таблица приняла более удобный для анализа вид, перенесем переменную «*пол*» в список *Столбцы,* а «*профессия*» – в список *Строки* (рис.3.1). (Сравните с таблицей, когда в строках – пол, а в столбцах - профессия).

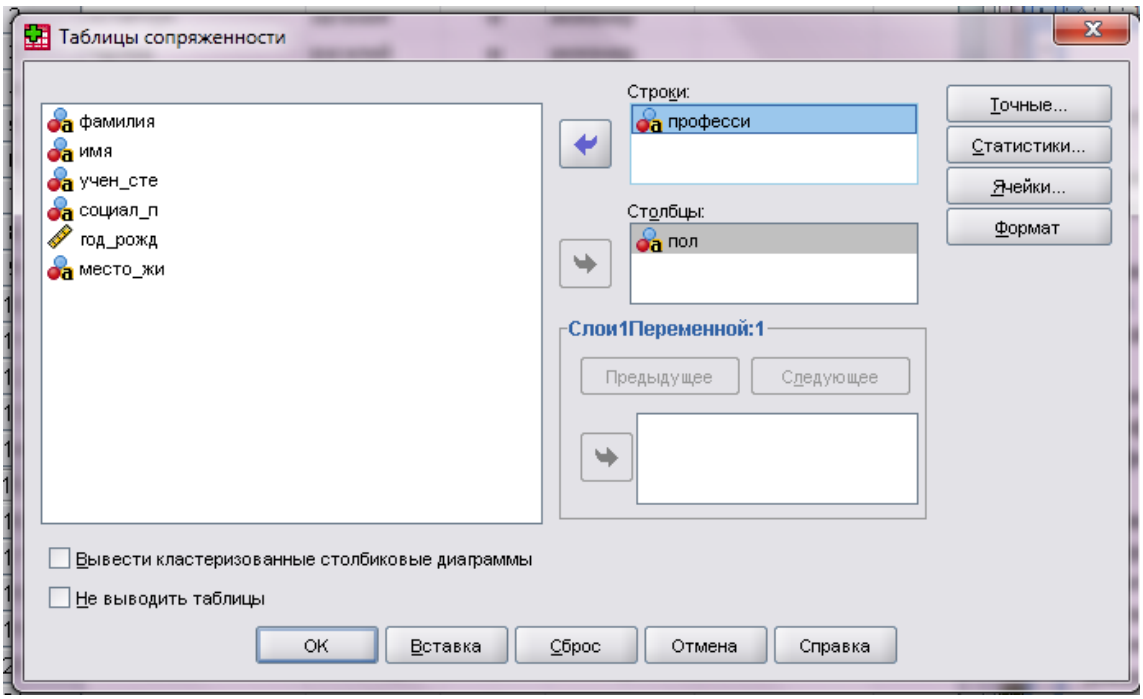


Рис.3.1. Диалоговое окно «Таблица сопряженности»

Первая таблица содержит информацию о числе самих наблюдений (рис.3.2).

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
профессии * пол	100	100,0%	0	,0%	100	100,0%

Рис.3.2. Таблица «Сводка обработки наблюдений»

Из второй таблицы (рис.3.3) (собственно таблицы сопряженности) видно, что два наблюдения содержат пропущенные (или утерянные) значения в переменной «*профессия*».

Переменная «*пол*» является столбцовой переменной, т.к. каждое её значение отображается в отдельном столбце. Соответственно, «*профессия*» - это переменная строк. Значение в каждой ячейке таблицы – количество наблюдений (частота).

Числа в последней строке и в последнем столбце *Итого* показывают суммы значений соответственно по строкам и столбцам.

Таблица сопряженности профессия * пол

Частота

		пол		
		ж	м	Итого
профессия		0	2	2
	агроном	0	4	4
	антрополог	0	1	1
	бухгалтер	1	0	1
	врач-хирур	1	0	1
	зооинженер	1	0	1
	зоотехник	0	2	2
	инженер	3	28	31
	историк	0	3	3
	культуроло	1	0	1
	лесничий	0	1	1
	медик	2	2	4
	медик-сани	0	1	1
	менеджер	1	2	3
	педагог	4	6	10
	преподават	2	4	6
	психолог	1	0	1
	радиофизик	0	1	1
	строитель	0	1	1
	техник	1	2	3
	товаровед	0	1	1
	ученый	0	3	3
	физик	0	1	1
	филолог	1	0	1
	экономи	1	5	6
	юрист	3	7	10
Итого		23	77	100

Рис.3.3. Таблица сопряженности «Профессия*пол»

Так, из столбца *Итого* видно, что из общего числа кандидатов в депутаты по профессиональной принадлежности превалируют инженерные (31 из 100), педагогические (10 из 100) и юридические (10 из 100) специальности. Причем большинство кандидатов, как в данных специальностях, так и во всей совокупности – это мужчины (см. строку *Итого* и строки наиболее массовых профессий).

- В данном примере из 100 человек – 77 мужчин и 23 женщины.
- Среди женщин-кандидатов наибольшей популярностью пользуется, в первую очередь, педагогическое направление (4 педагога и 2 преподавателя), затем – инженерное и юридическое (по 3 для каждого направления) и медицинское (2).
- Мужчины наиболее перспективными считают:
  1. инженерные профессии;
  2. педагогику и преподавание;
  3. юриспруденцию;

4. экономику.

Для данной таблицы сопряженности параметры приняты по умолчанию, поэтому в каждой ячейке отображается только абсолютная частота.

Более тщательно исследовать существование зависимости позволяет вычисление значений ожидаемых частот. Чтобы определить эти значения, выполните следующие действия:

- выберите меню команды *Анализ*/*Описательные статистики*/*Таблицы сопряженности;*
- соответствующие переменные перенесите в список строк и список столбцов;
- щелкните на кнопке *Ячейки*. Откроется диалоговое окно *Таблицы сопряженности: Вывод в ячейках* (рис.3.4).

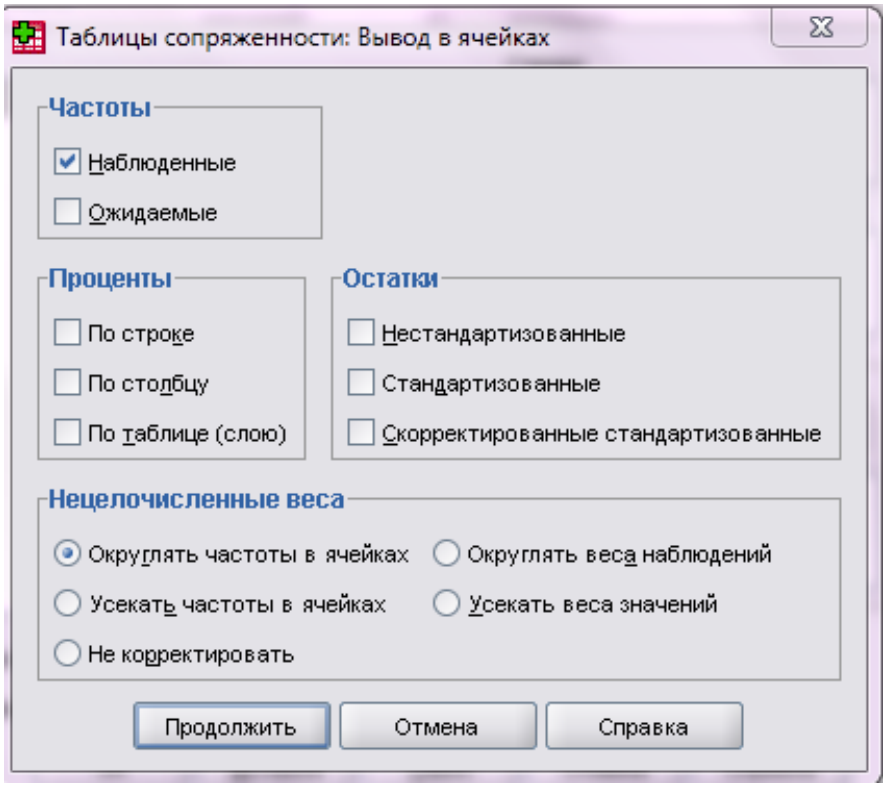


Рис.3.4. Диалоговое окно «Таблицы сопряженности: Вывод в ячейках»

По умолчанию в ячейках таблицы сопряженности отображаются только наблюдаемые значения частот. В группе *Частоты* можно выбрать один или более следующих вариантов:

- Наблюдаемые. Будут отображаться наблюдаемые частоты. Это настройка по умолчанию.
- Ожидаемые. Если установить этот флажок, будут отображаться ожидаемые частоты. Они вычисляются как

произведение сумм соответствующей строки и столбца, деленное на общую сумму частот.

- установите флажок *Ожидаемые*;
- щелкните *Продолжить* и затем *ОК*. Вы получите таблицу сопряженности, где под наблюдаемыми частотами расположены ожидаемые значения (рис.3.5).

Таблица сопряженности профессия * пол

			пол		
			ж	м	Итого
профессия		Частота	0	2	2
		Ожидаемая частота	,5	1,5	2,0
	агроном	Частота	0	4	4
		Ожидаемая частота	,9	3,1	4,0
	антрополог	Частота	0	1	1
		Ожидаемая частота	,2	,8	1,0
	бухгалтер	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
	врач-хирур	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
	зооинженер	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
	зоотехник	Частота	0	2	2
		Ожидаемая частота	,5	1,5	2,0
	инженер	Частота	3	28	31
		Ожидаемая частота	7,1	23,9	31,0
	историк	Частота	0	3	3
		Ожидаемая частота	,7	2,3	3,0
	культуроло	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
	лесничий	Частота	0	1	1
		Ожидаемая частота	,2	,8	1,0
	медик	Частота	2	2	4
		Ожидаемая частота	,9	3,1	4,0
	медик-сани	Частота	0	1	1
		Ожидаемая частота	,2	,8	1,0
	менеджер	Частота	1	2	3
		Ожидаемая частота	,7	2,3	3,0

Рис.3.5. Таблица сопряженности «Профессия*пол», содержащая ожидаемые частоты

Ещё одну возможность выявления существования зависимости между переменными дает вычисление остатков. Эти остатки являются показателем того, насколько сильно наблюдаемые и ожидаемые частоты отклоняются друг от друга.

Чтобы получить остатки частот, выберите меню команды *Анализ*/*Описательные статистики/Таблицы сопряженности,* перенесите переменные соответственно в список строк и список столбцов, затем щелкните на кнопке *Ячейки*. Флажки *Наблюдаемые* и *Ожидаемые* следует

оставить помеченными.

В группе *Остатки* можно выбрать один или более следующих вариантов отображения:

- Нестандартизированные (ненормированные). Отображаются ненормированные остатки, т.е. разность наблюдаемых (f_o) и ожидаемых (f_e) частот.
- Стандартизированные (Нормированные): Отображаются нормированные остатки. Для этого ненормированные остатки делятся на квадратный корень из ожидаемой частоты:

$$\frac{f_0 - f_e}{\sqrt{f_e}} \qquad (2.1)$$

  Нормированные остатки полезны при последующем проведении анализа тестов по критерию χ^2.
- Скорректированные стандартизированные (Уточненные и нормированные). Нормированные остатки вычисляются с учетом сумм по строкам и столбцам:

$$\frac{f_0 - f_e}{\sqrt{f_e * \left(\left(1 - {}^{z}/_{N}\right) * \left(1 - {}^{s}/_{N}\right)\right.}} \qquad (2.2)$$

где z – сумма по текущей строке,

s – сумма по текущему столбцу,

N – общая сумма частот.

Установите флажок *Нестандартизированные* и щелкните *Продолжить*, а в главном диалоговом окне – *OK*.

Перед вами появится таблица сопряженности, содержащая абсолютные частоты, ожидаемые частоты и остаток (рис.3.6).

Таблица сопряженности профессия * пол

			пол		
			ж	м	Итого
профессия		Частота	0	2	2
		Ожидаемая частота	,5	1,5	2,0
		Остаток	-,5	,5	
	агроном	Частота	0	4	4
		Ожидаемая частота	,9	3,1	4,0
		Остаток	-,9	,9	
	антрополог	Частота	0	1	1
		Ожидаемая частота	,2	,8	1,0
		Остаток	-,2	,2	
	бухгалтер	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
		Остаток	,8	-,8	
	врач-хирур	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
		Остаток	,8	-,8	
	зооинженер	Частота	1	0	1
		Ожидаемая частота	,2	,8	1,0
		Остаток	,8	-,8	
	зоотехник	Частота	0	2	2
		Ожидаемая частота	,5	1,5	2,0
		Остаток	-,5	,5	
	инженер	Частота	3	28	31
		Ожидаемая частота	7,1	23,9	31,0
		Остаток	-4,1	4,1	
	историк	Частота	0	3	3
		Ожидаемая частота	,7	2,3	3,0
		Остаток	-,7	,7	

Рис.3.6. Таблица сопряженности «Профессия*пол», содержащая абсолютные частоты, ожидаемые частоты и остаток

3. Графическое представление таблиц сопряженности

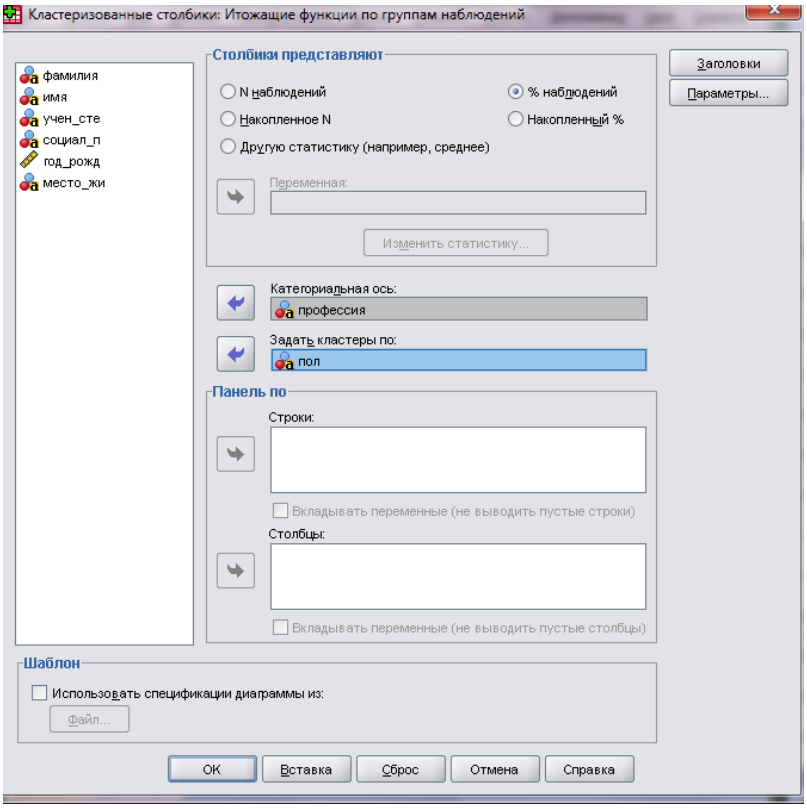


Рис.3.7. Диалоговое окно «Кластеризованные столбики: Итожащие функции по группам переменных»

Чтобы сделать более наглядными данные, содержащиеся в таблицах сопряженности, их можно представить визуально. Для этого выберите в меню команды *Графика/Устаревшие диалоговые окна/Столбики*.

- Выберите пункт *Кластер*, оставьте предлагаемую по умолчанию опцию *Итоги по группам наблюдений* и щелкните на кнопке *Задать*. Откроется окно *Кластеризованные столбики: Итожащие функции по группам переменных* (рис.3.7).
- Выберите пункт *% наблюдений*.
- Применительно к рассмотренному выше примеру, перенесите переменную «профессия» в поле *Категориальная ось*, «пол» - в *Задать кластеры по*.
- Щелкните *Заголовки*. Откроется окно*Ззаголовки* (рис.3.8).
- В поле *Строка 1* введите заголовок диаграммы, в поле *Подзаголовок* – подзаголовок, а в поле *Сноска, строка 1* – дополнительные сведения, если необходимо. Подтвердите ввод кнопкой *Продолжить*.

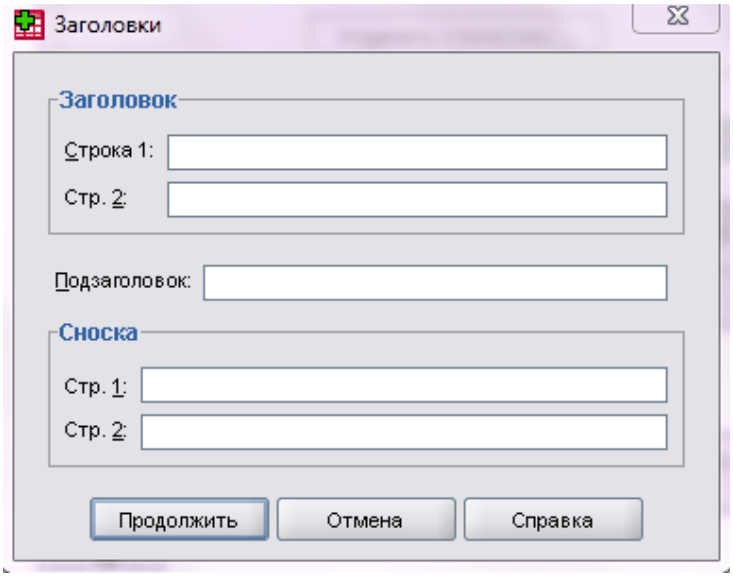


Рис.3.8: Диалоговое окно «Заголовки»

- После щелчка на кнопке *Параметры* откроется одноименное диалоговое окно.
- Снимите в нем флажок *Выводить группы, заданные пропущенными значениями*, если необходимо.
- Щелкните *Продолжить*, затем – *OK*. в окне просмотра появится график (рис.3.9).
- Откройте редактор диаграмм и измените диаграмму.

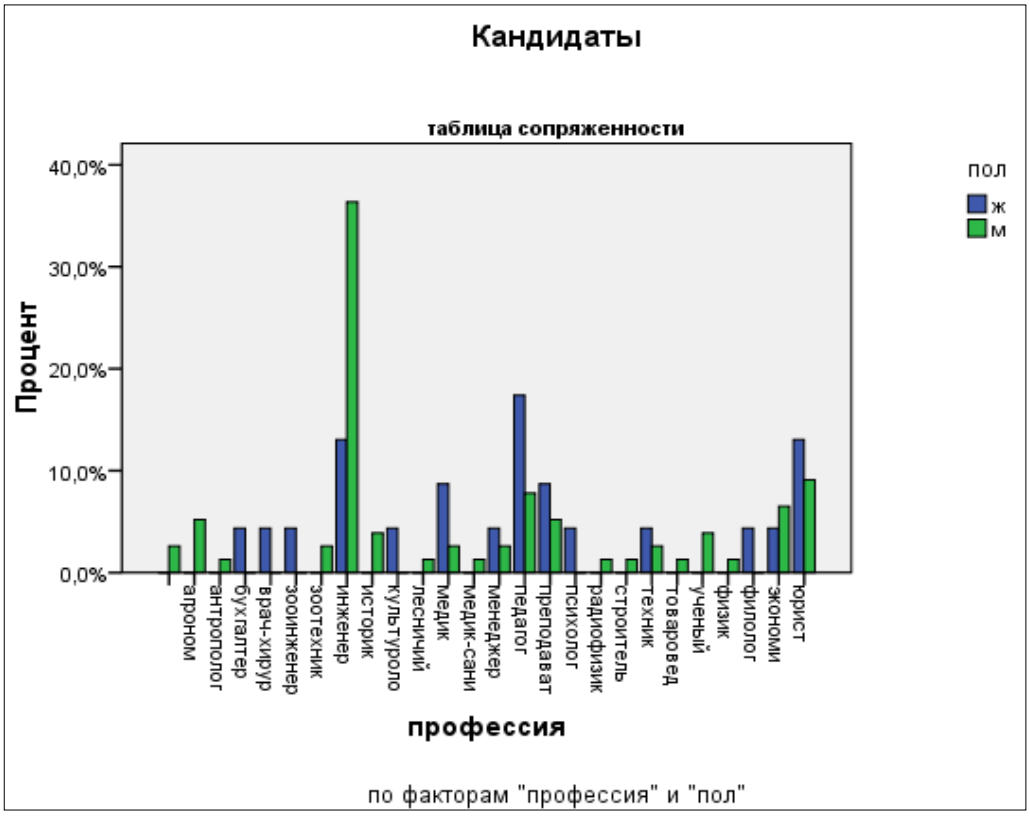


Рис.3.9. Диаграмма таблицы сопряженности «профессия*пол»

Можно не вызывать меню *Графика*, а просто установить в диалоге *Таблицы сопряженности* флажок *Вывести кластеризованные столбиковые диаграммы*.

4. Статистические критерии для таблиц сопряженности

Чтобы получить статистические критерии для таблиц сопряженности, щелкните на кнопке *Статистики* в диалоговом окне *Таблицы сопряженности*. Откроется окно *Таблицы сопряженности: Статистики*, в котором можно выбрать один или несколько критериев (рис.3.10).

- Тест χ^2 (хи-квадрат)
- Корреляции
- Меры связанности для переменных, относящихся к номинальной шкале
- Меры связанности для переменных, относящихся к порядковой шкале
- Меры связанности для переменных, относящихся к интервальной шкале
- Коэффициент Каппа (к)
- Мера риска
- Тест Мак-Немара

- Статистики Кокрена и Мантеля-Хенцеля

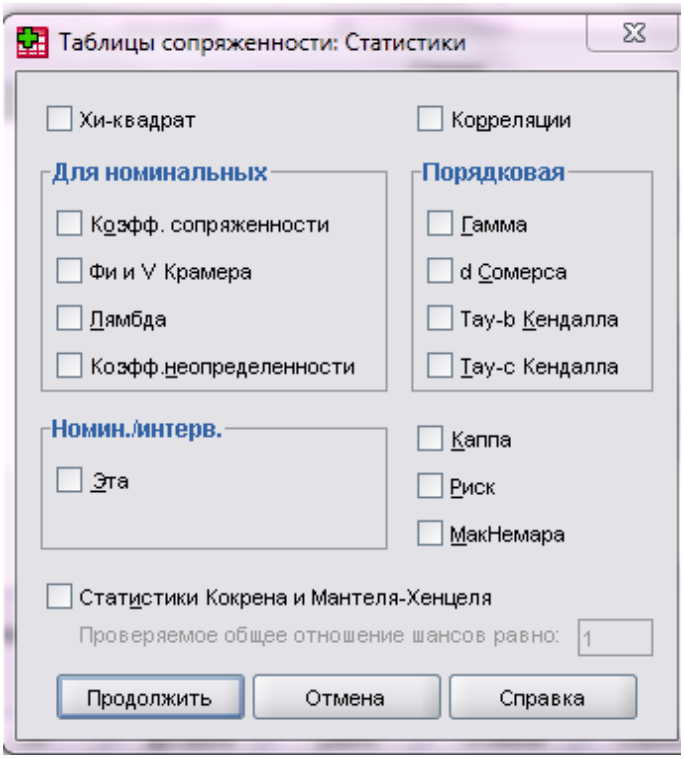


Рис.3.10. Диалоговое окно «Таблицы сопряженности: Статистики»

4.1. Тест хи-квадрат (χ^2)

При проведении теста хи-квадрат проверяется взаимная независимость двух переменных таблицы сопряженности. **Две переменные считаются взаимно независимыми, если наблюдаемые частоты (f_o) в ячейках совпадают с ожидаемыми частотами (f_e).**

Чтобы провести тест хи-квадрат с помощью SPSS Statistics, в диалоговом окне *Таблицы сопряженоости* кнопкой *Сброс* удалите все возможные настройки.

Перенесите нужные переменные в списки строк и столбцов (для нашего примера это – профессия и пол). Щелкните на кнопке *Ячейки*. В диалоговом окне установите, кроме предлагаемого по умолчанию флажка *Наблюдаемые*, ещё флажки *Ожидаемые* и *Стандартизированные*. Подтвердите ввод кнопкой *Продолжить*.

Щелкните на кнопке *Статистики*. В открывшемся диалоговом окне установите флажок *Хи-квадрат*. Щелкните *Продолжить*, а в главном окне – *ОК*.

Для файла о кандидатах, выполнив тест χ^2, в появившемся окне *Вывод* получим:

1. Таблицу с информацией о числе самих наблюдений.

2. Таблицу сопряженности.
3. Результаты теста хи-квадрат (рис.3.11).

Критерии хи-квадрат

	Значение	ст.св.	Асимпт. значимость (2-стор.)
Хи-квадрат Пирсона	33,881[a]	25	,110
Отношение правдоподобия	36,237	25	,068
Кол-во валидных наблюдений	100		

а. В 48 (92,3%) ячейках ожидаемая частота меньше 5. Минимальная ожидаемая частота равна ,23.

Рис.3.11. Результаты теста Хи-квадрат

Здесь указаны значения критерия хи-квадрат по формуле Пирсона и отношение правдоподобия, т.е. критерий хи-квадрат с поправкой на правдоподобие.

Критерий хи-квадрат по формуле Пирсона

В формуле Пирсона $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ вычисляется сумма квадратов стандартизованных (нормированных) остатков по всем полям таблицы сопряженности. Поэтому поля с более высоким стандартизованным остатком вносят более весомый вклад в численное значение критерия χ^2 и, следовательно, в значимый результат.

Правило: Считается, что существует значимое различие между наблюдаемой и ожидаемой частотой, если нормированный остаток больше или равен 2. Другие предельные значения принимаются в соответствии со следующей таблицей.

Таблица 3.1

Предельные значения

Нормированный остаток	Уровень значимости
>=2,0	p<0,05 (*)
>=2,6	p<0,01 (**)
>=3,3	p<0,001 (***)

Однако эти правила применимы только в случае, если ожидаемая частота не меньше 5.

Корректность проведения теста хи-квадрат определятся двумя условиями:

1. во-первых, ожидаемые частоты <5 должны встречаться не более чем

в 20 % полей таблицы;

2. во-вторых, суммы по строкам и столбцам всегда должны быть больше нуля.

В рассматриваемом нами примере формула Пирсона не дает даже минимально значимую величину критерия хи-квадрат (p>0,05). Кроме того, как указывает примечание после таблицы теста χ^2, 92,3 % полей имеют ожидаемую частоту менее 5. Так как допустимый предел в 20 % намного превышен, то расхождение между наблюдаемыми и ожидаемыми частотами считаем значительным.

Критерий хи-квадрат с поправкой на правдоподобие

Альтернативой формуле Пирсона для вычисления χ^2 является поправка на правдоподобие: $\chi^2 = -2 \cdot \sum f_o \cdot \ln \frac{f_e}{f_o}$

При большом объеме выборки формула Пирсона и подправленная формула дают очень близкие результаты. В нашем примере критерий хи-квадрат с поправкой на правдоподобие составляет 36,237.

4.2. Коэффициенты корреляции

Тест χ^2 позволяет лишь выяснить сам факт существования статистической зависимости между двумя признаками. Далее будем определять силу этой зависимости, её вид и направленность. **Критерии количественной оценки зависимости между переменными называются коэффициентами корреляции или мерами связанности.**

В качестве коэффициента корреляции между переменными, принадлежащими порядковой шкале, применяется коэффициент Спирмена, а для переменных с интервальной шкалой – коэффициент корреляции Пирсона, называемый также линейной корреляцией. При этом следует учесть, что каждую дихотомическую переменную, т.е. переменную, принадлежащую к номинальной шкале и имеющую две категории, можно рассматривать как порядковую.

Проверим, существует ли корреляция между переменными «ширина» и «вес» из файла данных «изучение крабов».

Для определения коэффициентов корреляции в диалоговом окне *Таблицы сопряженности* сбросьте все возможные настройки и перенесите соответствующие переменные в список строк и список столбцов. Щелкните на кнопке *Статистики*. Установите флажок *Корреляции*. Щелкните *Продолжить*. В диалоге *Таблицы сопряженности* откажитесь от вывода таблиц. Нажмите *ОК*.

Будут вычислены коэффициенты корреляции Спирмена и Пирсона, а также проведена проверка их значимости.

В данном случае обе переменные «ширина» и «вес» принадлежат к интервальной шкале. Проделав вышеуказанные действия, получим следующую таблицу (рис.3.12).

Симметричные меры

		Значение	Асимптотическая стандартная ошибка[a]	Прибл. T[b]	Прибл. значимость
Интервальная по интервальной	R Пирсона	,838	,051	12,665	,000[c]
Порядковая по порядковой	Корреляция Спирмена.	,884	,034	15,621	,000[c]
Кол-во валидных наблюдений		70			

a. Не подразумевая истинность нулевой гипотезы.

b. Используется асимптотическая стандартная ошибка в предположении истинности нулевой гипотезы.

c. На основании нормальной аппроксимации.

Рис.3.12. Симметричные меры переменных «ширина*вес»

Для интервальных переменных следует рассматривать коэффициент Пирсона (для порядковых - Спирмена). Он является максимально значимым ($p<0,001$) и составляет 0,838. Значит, между шириной и весом крабов изучаемой совокупности существует сильная корреляция (взаимосвязь). Переменные коррелируют положительно. Следовательно, большие по весу особи имеют больший размер тела, и наоборот.

4.3. Меры связанности для переменных с номинальной шкалой

Корреляции неприменимы для номинальных переменных, т.к. между их кодировками невозможно установить порядкового отношения и, следовательно, они не могут быть расположены в определенном, рационально объяснимом порядке.

Наилучшее средство для анализа таких зависимостей – тест хи-квадрат, а также анализ наблюдаемых и ожидаемых частот и нормированных остатков.

Кроме того, существуют критерии, позволяющие количественно оценить степень взаимной зависимости или независимости двух переменных, принадлежащих к номинальной шкале. Причем значение 0 соответствует полной независимости переменных, а 1 – их максимальной зависимости. Меры связанности не могут иметь отрицательных значений, т.к. при отсутствии порядкового отношения нельзя дать ответ на вопрос о направлении зависимости.

Чтобы задать все меры связанности для номинальных переменных,

необходимо в диалоговом окне *Таблицы сопряженности* после нажатия кнопки *Статистики* установить все флажки в группе *Для номинальных*.

5. Общие задания

1. Для рассмотренных в данной работе примеров проведите все возможные вычисления, постройте график, если это позволяет тип данных и проанализируйте получившиеся результаты.
2. Получите у преподавателя номер индивидуального задания.
3. Определите, к какой статистической шкале относятся статистические данные.
4. Введите данные в столбцы редактора данных, присвойте имя, тип и другие, необходимые для ваших переменных характеристики.
5. Определите, какие из рассмотренных выше статистических методик применимы к вашим данным. Ответ обоснуйте.
6. Проведите необходимые расчеты по вашим данным и дайте анализ полученным результатам.

6. Требования к отчету.

Отчет по лабораторной работе №3 предоставляется в печатном виде и должен содержать:

1. Теоретический материал.
2. Анализ результатов, полученных после выполнения задания 1.
3. Исходные данные и обоснование типа шкалы.
4. Результаты выполнения практических заданий с предоставлением необходимых расчетов, графиков и т.д.
5. Выводы по полученной информации.

После предоставления отчета, студент подтверждает знание изучаемого материала и защищает лабораторную работу за компьютером.

7. Индивидуальные задания

1. Выполнение контрольной работы

№ п.п.	Балл	Пол	Группа
1	2,000	1,000	1,000
2	3,000	1,000	1,000
3	1,000	1,000	1,000
4	4,000	2,000	1,000
5	5,000	2,000	1,000
6	3,000	2,000	1,000
7	6,000	1,000	2,000
8	7,000	1,000	2,000
9	5,000	1,000	2,000

10	8,000	2,000	2,000
11	9,000	2,000	2,000
12	7,000	2,000	2,000

2÷6. Размеры поршневых колец

№ п.п.	SAMPLE (модель)	SIZE
1	1	74,030
2	1	74,002
3	1	74,019
4	1	73,992
5	1	74,008
6	2	73,995
7	2	73,992
8	2	74,001
9	2	74,011
10	2	74,004
11	3	73,988
12	3	74,024
13	3	74,021
14	3	74,005
15	3	74,002
16	4	74,002
17	4	73,996
18	4	73,993
19	4	74,015
20	4	74,009
21	5	73,992
22	5	74,007
23	5	74,015
24	5	73,989
25	5	74,014
26	6	74,009
27	6	73,994
28	6	73,997
29	6	73,985
30	6	73,993
31	7	73,995
32	7	74,006
33	7	73,994
34	7	74,000
35	7	74,005
36	8	73,985
37	8	74,003
38	8	73,993
39	8	74,015
40	8	73,988
41	9	74,008
42	9	73,995
43	9	74,009
44	9	74,005

45	9	74,004
46	10	73,998
47	10	74,000
48	10	73,990
49	10	74,007
50	10	73,995
51	11	73,994
52	11	73,998
53	11	73,994
54	11	73,995
55	11	73,990
56	12	74,004
57	12	74,000
58	12	74,007
59	12	74,000
60	12	73,996
61	13	73,983
62	13	74,002
63	13	73,998
64	13	73,997
65	13	74,012
66	14	74,006
67	14	73,967
68	14	73,994
69	14	74,000
70	14	73,984
71	15	74,012
72	15	74,014
73	15	73,998
74	15	73,999
75	15	74,007
76	16	74,000
77	16	73,984
78	16	74,005
79	16	73,998
80	16	73,996
81	17	73,994
82	17	74,012
83	17	73,986
84	17	74,005
85	17	74,007
86	18	74,006
87	18	74,010
88	18	74,018
89	18	74,003
90	18	74,000
91	19	73,984
92	19	74,002
93	19	74,003
94	19	74,005

95	19	73,997
96	20	74,000
97	20	74,010
98	20	74,013
99	20	74,020
100	20	74,003
101	21	73,988
102	21	74,001
103	21	74,009
104	21	74,005
105	21	73,996
106	22	74,004
107	22	73,999
108	22	73,990
109	22	74,006
110	22	74,009
111	23	74,010
112	23	73,989
113	23	73,990
114	23	74,009
115	23	74,014
116	24	74,015
117	24	74,008
118	24	73,993
119	24	74,000
120	24	74,010
121	25	73,982
122	25	73,984
123	25	73,995
124	25	74,017
125	25	74,013

7÷11. Syrup Loss data (Изготовление сиропа)

	NOZZLE	OPERATOR	PRESSURE	LOSS
1	1	A	Low	-35
2	1	A	Low	-25
3	1	A	Medium	100
4	1	A	Medium	85
5	1	A	High	4
6	1	A	High	5
7	1	B	Low	-45
8	1	B	Low	-60
9	1	B	Medium	-10
10	1	B	Medium	30
11	1	B	High	-40
12	1	B	High	-30
13	1	C	Low	-40
14	1	C	Low	15
15	1	C	Medium	80

16	1	C	Medium	54
17	1	C	High	31
18	1	C	High	36
19	2	A	Low	17
20	2	A	Low	24
21	2	A	Medium	55
22	2	A	Medium	120
23	2	A	High	-23
24	2	A	High	-5
25	2	B	Low	-65
26	2	B	Low	-58
27	2	B	Medium	-55
28	2	B	Medium	-44
29	2	B	High	-64
30	2	B	High	-62
31	2	C	Low	20
32	2	C	Low	4
33	2	C	Medium	110
34	2	C	Medium	44
35	2	C	High	-20
36	2	C	High	-31
37	3	A	Low	-39
38	3	A	Low	-35
39	3	A	Medium	90
40	3	A	Medium	113
41	3	A	High	-30
42	3	A	High	-55
43	3	B	Low	-55
44	3	B	Low	-67
45	3	B	Medium	-28
46	3	B	Medium	-26
47	3	B	High	-61
48	3	B	High	-52
49	3	C	Low	15
50	3	C	Low	-30
51	3	C	Medium	110
52	3	C	Medium	135
53	3	C	High	54
54	3	C	High	4

12÷14. Ирисы Фишера: размеры чашелистиков и лепестков для трех типов ирисов

(Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris)

	SEPALLEN	SEPALWID	PETALLEN	PETALWID	IRISTYPE
1	5,0	3,3	1,4	,2	SETOSA
2	6,4	2,8	5,6	2,2	VIRGINIC
3	6,5	2,8	4,6	1,5	VERSICOL
4	6,7	3,1	5,6	2,4	VIRGINIC
5	6,3	2,8	5,1	1,5	VIRGINIC

6	4,6	3,4	1,4	,3	SETOSA
7	6,9	3,1	5,1	2,3	VIRGINIC
8	6,2	2,2	4,5	1,5	VERSICOL
9	5,9	3,2	4,8	1,8	VERSICOL
10	4,6	3,6	1,0	,2	SETOSA
11	6,1	3,0	4,6	1,4	VERSICOL
12	6,0	2,7	5,1	1,6	VERSICOL
13	6,5	3,0	5,2	2,0	VIRGINIC
14	5,6	2,5	3,9	1,1	VERSICOL
15	6,5	3,0	5,5	1,8	VIRGINIC
16	5,8	2,7	5,1	1,9	VIRGINIC
17	6,8	3,2	5,9	2,3	VIRGINIC
18	5,1	3,3	1,7	,5	SETOSA
19	5,7	2,8	4,5	1,3	VERSICOL
20	6,2	3,4	5,4	2,3	VIRGINIC
21	7,7	3,8	6,7	2,2	VIRGINIC
22	6,3	3,3	4,7	1,6	VERSICOL
23	6,7	3,3	5,7	2,5	VIRGINIC
24	7,6	3,0	6,6	2,1	VIRGINIC
25	4,9	2,5	4,5	1,7	VIRGINIC
26	5,5	3,5	1,3	,2	SETOSA
27	6,7	3,0	5,2	2,3	VIRGINIC
28	7,0	3,2	4,7	1,4	VERSICOL
29	6,4	3,2	4,5	1,5	VERSICOL
30	6,1	2,8	4,0	1,3	VERSICOL
31	4,8	3,1	1,6	,2	SETOSA
32	5,9	3,0	5,1	1,8	VIRGINIC
33	5,5	2,4	3,8	1,1	VERSICOL
34	6,3	2,5	5,0	1,9	VIRGINIC
35	6,4	3,2	5,3	2,3	VIRGINIC
36	5,2	3,4	1,4	,2	SETOSA
37	4,9	3,6	1,4	,1	SETOSA
38	5,4	3,0	4,5	1,5	VERSICOL
39	7,9	3,8	6,4	2,0	VIRGINIC
40	4,4	3,2	1,3	,2	SETOSA
41	6,7	3,3	5,7	2,1	VIRGINIC
42	5,0	3,5	1,6	,6	SETOSA
43	5,8	2,6	4,0	1,2	VERSICOL
44	4,4	3,0	1,3	,2	SETOSA
45	7,7	2,8	6,7	2,0	VIRGINIC
46	6,3	2,7	4,9	1,8	VIRGINIC
47	4,7	3,2	1,6	,2	SETOSA
48	5,5	2,6	4,4	1,2	VERSICOL
49	7,2	3,2	6,0	1,8	VIRGINIC
50	4,8	3,0	1,4	,3	SETOSA
51	5,1	3,8	1,6	,2	SETOSA
52	6,1	3,0	4,9	1,8	VIRGINIC
53	4,8	3,4	1,9	,2	SETOSA
54	5,0	3,0	1,6	,2	SETOSA
55	5,0	3,2	1,2	,2	SETOSA

56	6,1	2,6	5,6	1,4	VIRGINIC
57	6,4	2,8	5,6	2,1	VIRGINIC
58	4,3	3,0	1,1	,1	SETOSA
59	5,8	4,0	1,2	,2	SETOSA
60	5,1	3,8	1,9	,4	SETOSA
61	6,7	3,1	4,4	1,4	VERSICOL
62	6,2	2,8	4,8	1,8	VIRGINIC
63	4,9	3,0	1,4	,2	SETOSA
64	5,1	3,5	1,4	,2	SETOSA
65	5,6	3,0	4,5	1,5	VERSICOL
66	5,8	2,7	4,1	1,0	VERSICOL
67	5,0	3,4	1,6	,4	SETOSA
68	4,6	3,2	1,4	,2	SETOSA
69	6,0	2,9	4,5	1,5	VERSICOL
70	5,7	2,6	3,5	1,0	VERSICOL
71	5,7	4,4	1,5	,4	SETOSA
72	5,0	3,6	1,4	,2	SETOSA
73	7,7	3,0	6,1	2,3	VIRGINIC
74	6,3	3,4	5,6	2,4	VIRGINIC
75	5,8	2,7	5,1	1,9	VIRGINIC
76	5,7	2,9	4,2	1,3	VERSICOL
77	7,2	3,0	5,8	1,6	VIRGINIC
78	5,4	3,4	1,5	,4	SETOSA
79	5,2	4,1	1,5	,1	SETOSA
80	7,1	3,0	5,9	2,1	VIRGINIC
81	6,4	3,1	5,5	1,8	VIRGINIC
82	6,0	3,0	4,8	1,8	VIRGINIC
83	6,3	2,9	5,6	1,8	VIRGINIC
84	4,9	2,4	3,3	1,0	VERSICOL
85	5,6	2,7	4,2	1,3	VERSICOL
86	5,7	3,0	4,2	1,2	VERSICOL
87	5,5	4,2	1,4	,2	SETOSA
88	4,9	3,1	1,5	,2	SETOSA
89	7,7	2,6	6,9	2,3	VIRGINIC
90	6,0	2,2	5,0	1,5	VIRGINIC
91	5,4	3,9	1,7	,4	SETOSA
92	6,6	2,9	4,6	1,3	VERSICOL
93	5,2	2,7	3,9	1,4	VERSICOL
94	6,0	3,4	4,5	1,6	VERSICOL
95	5,0	3,4	1,5	,2	SETOSA
96	4,4	2,9	1,4	,2	SETOSA
97	5,0	2,0	3,5	1,0	VERSICOL
98	5,5	2,4	3,7	1,0	VERSICOL
99	5,8	2,7	3,9	1,2	VERSICOL
100	4,7	3,2	1,3	,2	SETOSA
101	4,6	3,1	1,5	,2	SETOSA
102	6,9	3,2	5,7	2,3	VIRGINIC
103	6,2	2,9	4,3	1,3	VERSICOL
104	7,4	2,8	6,1	1,9	VIRGINIC
105	5,9	3,0	4,2	1,5	VERSICOL

106	5,1	3,4	1,5	,2	SETOSA
107	5,0	3,5	1,3	,3	SETOSA
108	5,6	2,8	4,9	2,0	VIRGINIC
109	6,0	2,2	4,0	1,0	VERSICOL
110	7,3	2,9	6,3	1,8	VIRGINIC
111	6,7	2,5	5,8	1,8	VIRGINIC
112	4,9	3,1	1,5	,1	SETOSA
113	6,7	3,1	4,7	1,5	VERSICOL
114	6,3	2,3	4,4	1,3	VERSICOL
115	5,4	3,7	1,5	,2	SETOSA
116	5,6	3,0	4,1	1,3	VERSICOL
117	6,3	2,5	4,9	1,5	VERSICOL
118	6,1	2,8	4,7	1,2	VERSICOL
119	6,4	2,9	4,3	1,3	VERSICOL
120	5,1	2,5	3,0	1,1	VERSICOL
121	5,7	2,8	4,1	1,3	VERSICOL
122	6,5	3,0	5,8	2,2	VIRGINIC
123	6,9	3,1	5,4	2,1	VIRGINIC
124	5,4	3,9	1,3	,4	SETOSA
125	5,1	3,5	1,4	,3	SETOSA
126	7,2	3,6	6,1	2,5	VIRGINIC
127	6,5	3,2	5,1	2,0	VIRGINIC
128	6,1	2,9	4,7	1,4	VERSICOL
129	5,6	2,9	3,6	1,3	VERSICOL
130	6,9	3,1	4,9	1,5	VERSICOL
131	6,4	2,7	5,3	1,9	VIRGINIC
132	6,8	3,0	5,5	2,1	VIRGINIC
133	5,5	2,5	4,0	1,3	VERSICOL
134	4,8	3,4	1,6	,2	SETOSA
135	4,8	3,0	1,4	,1	SETOSA
136	4,5	2,3	1,3	,3	SETOSA
137	5,7	2,5	5,0	2,0	VIRGINIC
138	5,7	3,8	1,7	,3	SETOSA
139	5,1	3,8	1,5	,3	SETOSA
140	5,5	2,3	4,0	1,3	VERSICOL
141	6,6	3,0	4,4	1,4	VERSICOL
142	6,8	2,8	4,8	1,4	VERSICOL
143	5,4	3,4	1,7	,2	SETOSA
144	5,1	3,7	1,5	,4	SETOSA
145	5,2	3,5	1,5	,2	SETOSA
146	5,8	2,8	5,1	2,4	VIRGINIC
147	6,7	3,0	5,0	1,7	VERSICOL
148	6,3	3,3	6,0	2,5	VIRGINIC
149	5,3	3,7	1,5	,2	SETOSA
150	5,0	2,3	3,3	1,0	VERSICOL

15÷17. Планирование эксперимента: 5 операторов, 3 испытания, 8 деталей (5 operators, 3 trials, 8 parts)

	OPERATOR	PART	TRIAL (испытание)	MEASURE (размер)
1	SMITH	1	1	108,65

2	SMITH	2	1	111,91
3	SMITH	3	1	96,23
4	SMITH	4	1	95,68
5	SMITH	5	1	111,59
6	SMITH	6	1	94,20
7	SMITH	7	1	109,42
8	SMITH	8	1	108,95
9	SMITH	1	2	107,20
10	SMITH	2	2	113,50
11	SMITH	3	2	94,78
12	SMITH	4	2	100,10
13	SMITH	5	2	110,74
14	SMITH	6	2	96,56
15	SMITH	7	2	107,93
16	SMITH	8	2	112,38
17	SMITH	1	3	103,62
18	SMITH	2	3	112,78
19	SMITH	3	3	90,07
20	SMITH	4	3	93,01
21	SMITH	5	3	112,33
22	SMITH	6	3	98,09
23	SMITH	7	3	107,43
24	SMITH	8	3	107,22
25	HILL	1	1	97,12
26	HILL	2	1	106,46
27	HILL	3	1	85,62
28	HILL	4	1	93,07
29	HILL	5	1	107,51
30	HILL	6	1	88,57
31	HILL	7	1	98,95
32	HILL	8	1	105,71
33	HILL	1	2	102,04
34	HILL	2	2	101,83
35	HILL	3	2	87,16
36	HILL	4	2	90,84
37	HILL	5	2	106,28
38	HILL	6	2	89,94
39	HILL	7	2	102,35
40	HILL	8	2	103,31
41	HILL	1	3	103,67
42	HILL	2	3	104,59
43	HILL	3	3	83,82
44	HILL	4	3	92,12
45	HILL	5	3	108,22
46	HILL	6	3	90,52
47	HILL	7	3	103,34
48	HILL	8	3	106,51
49	JONES	1	1	102,10
50	JONES	2	1	115,21
51	JONES	3	1	94,20

52	JONES	4	1	97,20
53	JONES	5	1	113,68
54	JONES	6	1	93,35
55	JONES	7	1	110,55
56	JONES	8	1	106,26
57	JONES	1	2	105,72
58	JONES	2	2	112,10
59	JONES	3	2	93,12
60	JONES	4	2	97,09
61	JONES	5	2	112,96
62	JONES	6	2	98,69
63	JONES	7	2	109,68
64	JONES	8	2	110,77
65	JONES	1	3	101,59
66	JONES	2	3	108,54
67	JONES	3	3	91,37
68	JONES	4	3	95,48
69	JONES	5	3	112,68
70	JONES	6	3	91,61
71	JONES	7	3	104,79
72	JONES	8	3	111,54
73	HANKS	1	1	104,03
74	HANKS	2	1	110,46
75	HANKS	3	1	90,59
76	HANKS	4	1	92,87
77	HANKS	5	1	111,09
78	HANKS	6	1	94,49
79	HANKS	7	1	106,48
80	HANKS	8	1	110,72
81	HANKS	1	2	107,50
82	HANKS	2	2	113,66
83	HANKS	3	2	93,94
84	HANKS	4	2	94,11
85	HANKS	5	2	111,03
86	HANKS	6	2	94,00
87	HANKS	7	2	109,05
88	HANKS	8	2	110,47
89	HANKS	1	3	109,38
90	HANKS	2	3	107,96
91	HANKS	3	3	88,49
92	HANKS	4	3	101,06
93	HANKS	5	3	109,37
94	HANKS	6	3	90,05
95	HANKS	7	3	104,45
96	HANKS	8	3	106,58
97	MILLER	1	1	103,11
98	MILLER	2	1	108,11
99	MILLER	3	1	90,18
100	MILLER	4	1	94,81
101	MILLER	5	1	112,81

102	MILLER	6	1	91,63
103	MILLER	7	1	106,90
104	MILLER	8	1	104,93
105	MILLER	1	2	105,12
106	MILLER	2	2	106,59
107	MILLER	3	2	87,57
108	MILLER	4	2	96,49
109	MILLER	5	2	110,76
110	MILLER	6	2	88,98
111	MILLER	7	2	105,38
112	MILLER	8	2	105,32
113	MILLER	1	3	103,63
114	MILLER	2	3	105,94
115	MILLER	3	3	88,20
116	MILLER	4	3	94,47
117	MILLER	5	3	108,50
118	MILLER	6	3	91,28
119	MILLER	7	3	101,73
120	MILLER	8	3	104,56

18÷20. Исследование пульса (Pulsation data, Milliken and Johnsn (1992, p. 6))

	PULSE	TASK
1	27	1
2	31	1
3	26	1
4	32	1
5	39	1
6	37	1
7	38	1
8	39	1
9	30	1
10	28	1
11	27	1
12	27	1
13	34	1
14	29	2
15	28	2
16	37	2
17	24	2
18	35	2
19	40	2
20	40	2
21	31	2
22	30	2
23	25	2
24	29	2
25	25	2
26	34	3
27	36	3

28	34	3
29	41	3
30	30	3
31	44	3
32	44	3
33	32	3
34	32	3
35	31	3
36	34	4
37	34	4
38	43	4
39	44	4
40	40	4
41	47	4
42	34	4
43	31	4
44	45	4
45	28	4
46	28	5
47	28	5
48	26	5
49	35	5
50	31	5
51	30	5
52	34	5
53	34	5
54	26	5
55	20	5
56	41	5
57	21	5
58	28	6
59	26	6
60	29	6
61	25	6
62	35	6
63	34	6
64	37	6
65	28	6
66	21	6
67	28	6
68	26	6

8. Контрольные вопросы.

1. Что показывает таблица сопряженности и как ее вывести?
2. Для чего вычисляются остатки частот?
3. Какие существуют возможности визуализации таблиц сопряженности?
4. Объясните методику расчета критерия хи-квадрат по формуле

Пирсона.

5. Что утверждается правилом Пирсона?
6. Что такое однонаправленное соотношение между коррелированными переменными?
7. Укажите свойства линейного коэффициента корреляции (Пирсона). Как его определить?
8. Что представляет собой коэффициент Спирмена?
9. Как найти коэффициент сопряженности признаков?

Лабораторная работа №4

Тема: Сравнение средних в группах данных.

Цель: Научиться объяснять статистическое различие выборок с помощью анализа средних значений.

1. Введение

Сравнение средних значений различных выборок относится к наиболее часто применяемым методам статистического анализа. При этом всегда должен быть выяснен вопрос, можно ли объяснить имеющееся различие средних значений статистическими колебаниями или нет. В последнем случае говорят о значимом различии.

Задачи сравнения средних часто возникают на практике, например, вы можете сравнить средний доход двух групп людей: имеющих высшее образование и не имеющих высшего образования.

Сравнение средних значений применяется для переменных, измеренных в непрерывной шкале, такими переменными являются, например, доход или артериальное давление. Переменные, измеренные в бедных шкалах, исследуются с помощью специальных методов. В частности, категориальные переменные исследуются с помощью таблиц сопряженности. Переменные, измеренные в порядковых шкалах, исследуются методами непараметрической статистики.

При сравнении средних значений выборок предполагается, что выборки подчиняются нормальному распределению. Если это не так, то вычисляются медианы и для сравнения выборок используется непараметрический тест.

Заметим, что возможны два варианта организации данных: вы можете иметь дело с независимыми группами наблюдений или с зависимыми группами наблюдений. Если вы случайным образом разбили выборку на две части и сравниваете показатели в первой и второй группе, то, скорее всего, вы имеете дело с независимыми группами.

2. Сравнение двух независимых выборок

Т-критерий является наиболее часто используемым методом, позволяющим выявить различия между средними двух выборок. Напомним: переменные должны быть измерены в достаточно богатой шкале, например количественной.

Т-критерий может применяться, даже если размер выборки очень небольшой (например, 10) и если переменные нормально распределены

(внутри групп), а дисперсии наблюдений в группах не слишком различны.

Предположение о нормальности можно проверить, исследуя распределение (например, визуально с помощью гистограмм) или применяя критерий нормальности (хи-квадрат).

Равенство дисперсий в двух группах можно проверить с помощью **F-критерия Ливиня**, который включен в таблицу вывода t-критерия в SPSS Statistics.

P-уровень значимости t-критерия равен вероятности ошибочно отвергнуть гипотезу об отсутствии различия между средними выборок, когда она верна (т.е. когда средние в действительности равны).

Чтобы применить t-критерий для независимых выборок, требуется, по крайней мере, одна независимая (группирующая) переменная и одна зависимая переменная.

Вначале с помощью группирующей переменной (например, «м» и «ж», если группирующей переменной является Пол) данные разбиваются на две группы. Далее в каждой группе вычисляется среднее значение зависимой переменной, например артериальное давление или доход. Эти выборочные средние сравниваются между собой.

Будем работать с файлом «adstudy», в котором содержатся результаты социологического опроса: пол, предпочтение (выбор одного из двух напитков), объем ежедневно потребляемой жидкости. Цель нашего исследования: определить, зависит ли среднесуточный объём потребляемой жидкости от пола человека.

Выберите с помощью меню *Анализ/Сравнение средних/Т-критерий для независимых выборок* (рис.4.1).

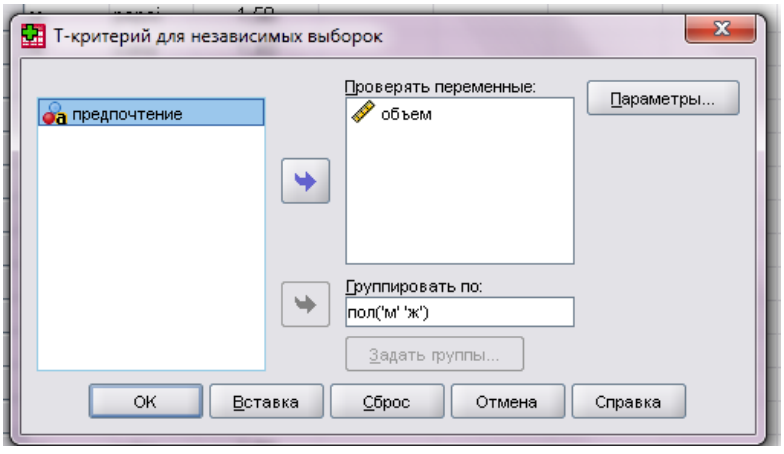


Рис.4.1. Диалоговое окно «Т-критерий для независимых выборок»

В открывшемся диалоговом окне *Т-критерий для независимых выборок* выполните следующие действия:

- в списке исходных переменных выберите зависимую

переменную, и перенести её в список тестируемых переменных *Проверять переменные*. Для нашего примера – это объем потребляемой жидкости;

- таким же способом перенесите группирующую переменную (пол) в поле *Группировать по*;
- после щелчка на кнопке *Задать группы* в новом окне можно ввести значения двух категорий для группирующей переменной. Для группы 1 введем значение «*м*» (мужчина), для группы 2 – «*ж*» (женщина) и нажмем *Продолжить*;
- для кнопки *Параметры* оставьте те настройки, которые установлены по умолчанию;
- запустите Т-тест, щелкнув *ОК*. В окне просмотра *Вывод* появятся результаты (рис.4.2).

Групповые статистики

	пол	N	Среднее	Стд. отклонение	Стд. ошибка среднего
объем	м	17	1,6059	,27493	,06668
	ж	13	1,5023	,37359	,10362

Критерий для независимых выборок

		Критерий равенства дисперсий Ливиня		t-критерий равенства средних						
									95% доверительный интервал разности средних	
		F	Знч.	t	ст.св.	Значимость (2-сторонняя)	Разность средних	Стд. ошибка разности	Нижняя граница	Верхняя граница
объем	Предполагается равенство дисперсий	,812	,375	,876	28	,389	,10357	,11825	-,13865	,34580
	Равенство дисперсий не предполагается			,841	21,263	,410	,10357	,12322	-,15248	,35963

Рис.4.2. Т-критерий для независимых выборок

Выведенные таблицы содержат:

- количество наблюдений, средние значения, стандартные отклонения и стандартные ошибки средних в обеих группах;
- результаты теста Ливня на равенство дисперсий.

Как правило, гипотеза о равенстве (гомогенности) дисперсий не принимается, если тест Ливня дает значение $p<0,05$ (гетерогенность дисперсий). Для случаев как гомогенности (равенства), так и гетерогенности (неравенства) выводятся следующие характеристики:

- результаты t-теста: значение распределения t, количество степеней свободы df, вероятность ошибки p (2-сторонняя);
- а также: разница средних значений, её стандартная ошибка и доверительный интервал.

В данном примере $p=0,389$, т.е. объем потребляемой жидкости не зависит от пола. Тест Ливня также показывает, что нет связи между полом человека и объемом выпиваемой за день жидкости ($p=0,375$).

3. Сравнение двух зависимых выборок

Определим, значимо ли изменяется частота пульса пациентов в зависимости от вида выполняемого упражнения (файл данных «Пульс»). Для этого сравним переменные «*пульс*» и «*task*» при помощи t-теста для зависимых выборок.

- Выберите в меню *Анализ/Сравнение средних/Т-критерий для парных выборок*;
- В диалоговом окне *Т-критерий для парных выборо* перенесите переменные «*пульс*» и «*task*» в поле *Парные переменные* (рис.4.3);

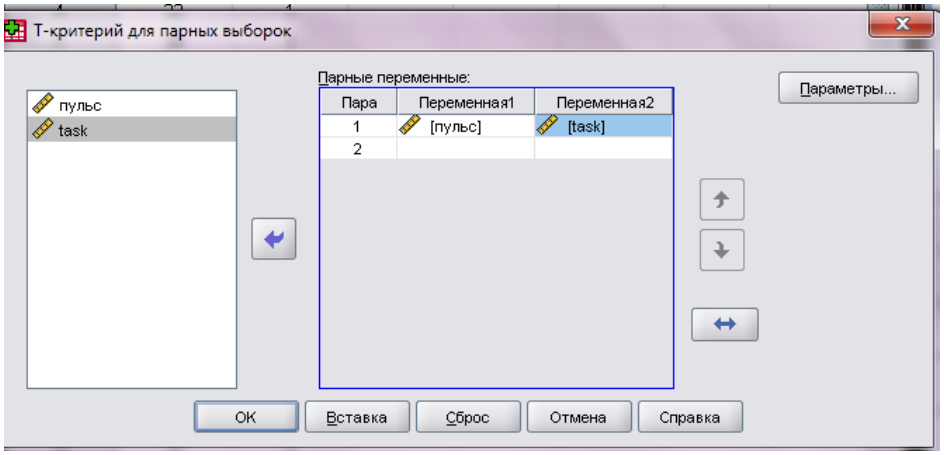


Рис.4.3. Диалоговое окно Т-критерий для парных выборок

- Щелкните *OK*. В окне просмотра *Вывод* появятся результаты (рис.4.4).

Статистики парных выборок

		Среднее	N	Стд. отклонение	Стд. ошибка среднего
Пара 1	пульс	32,31	68	6,242	,757
	task	3,43	68	1,765	,214

Корреляции парных выборок

		N	Корреляция	Знч.
Пара 1	пульс & task	68	-,137	,266

Критерий парных выборок

		Парные разности					t	ст.св.	Значимость (2-сторонняя)
		Среднее	Стд. отклонение	Стд. ошибка среднего	95% доверительный интервал разности средних: Нижняя граница	95% доверительный интервал разности средних: Верхняя граница			
Пара 1	пульс - task	28,882	6,715	,814	27,257	30,508	35,469	67	,000

Рис.4.4. Т-критерий для парных выборок

Выведенные таблицы содержат:

- средние значения, количество наблюдений, стандартные

отклонения и стандартные ошибки средних для обеих переменных;

- коэффициент корреляции (момент произведений Пирсона) между переменными и значимость его отклонения от нуля;
- среднее значение, количество наблюдений, стандартное отклонение и стандартную ошибку разницы;
- результаты t-теста: значение распределения t, количество степеней свободы df, вероятность ошибки p (2-сторонняя).

В данном примере значимость t-теста составила p=0,000 (максимальный уровень значимости, т.к. p<0,001), т.е. сложность физических упражнений является определяющей величиной при измерении пульса человека.

Чтобы выполнить t-тест только для заданных условий:

- Выберите в меню *Данные/Отобрать наблюдения* (рис.4.5);
- Выберите опцию *Если выполняется условие*. Щелчком на кнопке *Если* откройте диалоговое окно, в котором можно сформулировать необходимое условие (например: пульс<27) (рис.4.6).
- Щелкните *Продолжить*, а в основном диалоге – *OK*.
- Снова запустите t-тест. Теперь он будет выполняться только для заданного условия. Сравните полученные результаты.

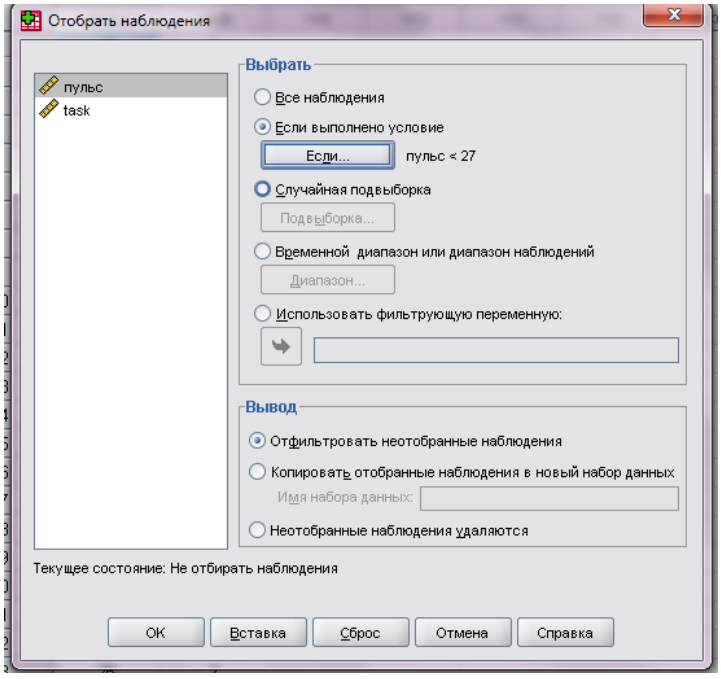


Рис.4.5. Диалоговое окно «Отобрать наблюдения»

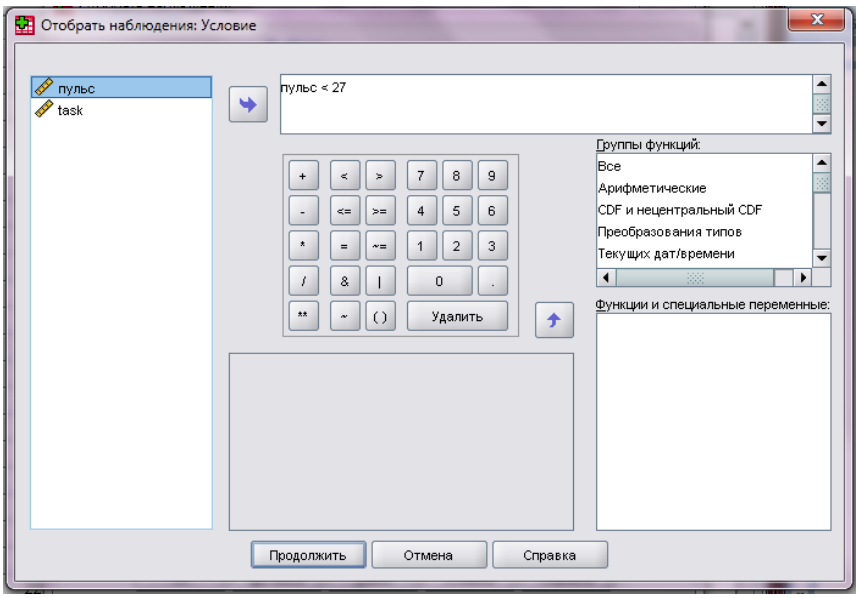


Рис.4.6. Диалоговое окно «Отобрать наблюдения: Условие»

Чтобы последующий анализ снова можно было проводить с использованием всех наблюдений, откройте диалоговое окно *Отобрать наблюдения* и выберите в нем опцию *Все наблюдения*.

4. Сравнение более двух независимых выборок

В файле «*лазер*» приведены значения угловой расходимости излучения лазера Y в минутах в зависимости от длины резонатора L и диаметра образца D (см.). Исследуем, зависит ли расходимость от диаметра образца. Для этого будем сравнивать 4 выборки величины Y, определяемые величиной D.

№ \ D	0,5	1,0	1,5	2,0
1	8.261	5.626	6.315	5.741
2	6.285	8.722	5.945	6.250
3	3.132	5.627	6.893	6.913
4	7.433	8.998	6.352	8.783

- Выберите в меню *Анализ/Сравнение средних/Однофакторный дисперсионный анализ* (рис.4.7);
- Переменную Y перенесите в список *Зависимые переменные*, а переменную D – в поле *Фактор*, т.е. переменных, определяющих значение зависимой переменной;
- кнопкой *Параметры* задайте следующее: вывод описательной статистики (флажок *Описательные*) и проверку на гомогенность дисперсий (флажок *Проверка однородности дисперсий*) (рис.4.8). Щелкните *Продолжить*;
- Чтобы выполнить апостериорный тест, вернувшись в основное диалоговое окно, щелкните на кнопке *Апостериорные*.

Откроется диалоговое окно *Однофакторный дисперсионный анализ: апостериорные множественные сравнения* (рис.4.9);

- выберите *тест Дункана* (*Дункан*). При значимом результате дисперсионного анализа этот тест показывает, какие именно возрастные группы значимо отличаются друг от друга. По умолчанию установлен уровень значимости 0,05. Щелкните *Продолжить;*
- Запустите тест, щелкнув *ОК*.

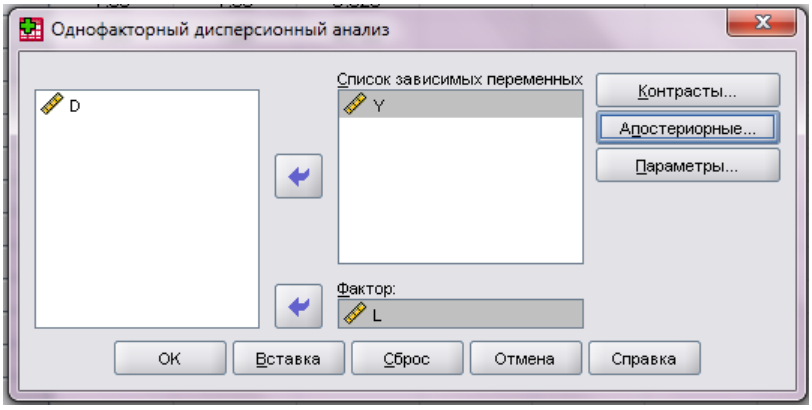


Рис.4.7. Диалоговое окно «Однофакторный дисперсионный анализ»

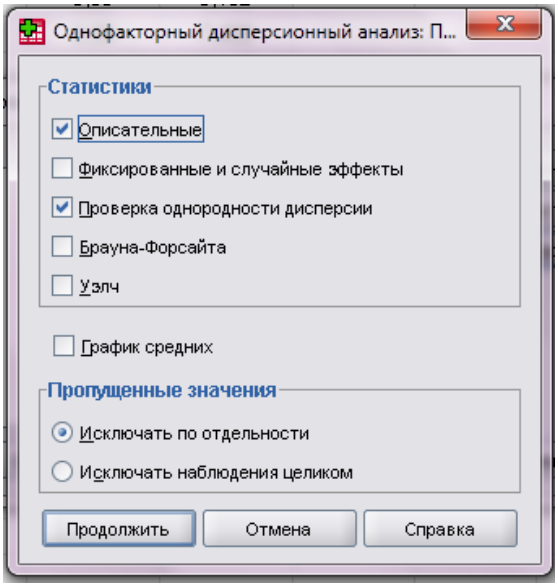


Рис.4.8. Диалоговое окно «Однофакторный дисперсионный анализ: Параметры»

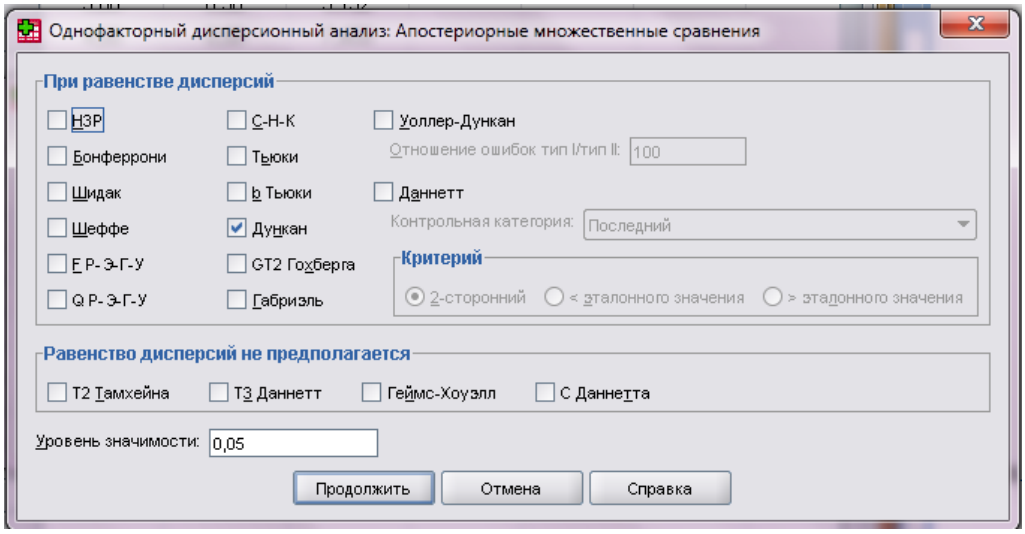


Рис.4.9. Диалоговое окно «Однофакторный дисперсионный анализ: апостериорные множественные сравнения»

В окне просмотра *Вывод* появятся 4 таблицы:

- Описательные статистики, в которой содержатся: число наблюдений, средние значения, стандартные отклонения и стандартные ошибки средних, 95 % доверительные интервалы, минимумы и максимумы для всех слоев фактора.
- Критерий однородности дисперсий (Статистика Ливня).
- Типовая схема дисперсионного анализа, включая вероятность ошибки p (значимость) для оценки общей значимости.
- Результаты многорангового теста Дункана с выделением однородных подмножеств.

Для нашего примера дисперсионный анализ дает незначимый результат (p=0,969), т.е. в рассматриваемых группах дисперсия угловой расходимости излучения лазера слишком различна.

Тест Дункана выделил всего одну гомогенную подгруппу (см. столбец *Подмножество для альфа = 0,05*). Это означает, что не наблюдается большого отличия средних показателей расходимости во всех четырех группах. Кроме того, результаты теста Дункана показывают, что с увеличением диаметра образца D увеличивается (хотя и незначительно) и расходимость Y.

Повторим этот анализ для переменных Y и L. Окажется, что изменение расходимости в большей степени зависит от длины резонатора L, чем от диаметра образца.

5. t-тест одной выборки

Этот тест позволяет выяснить, отличается ли среднее значение,

полученное на основе данной выборки, от предварительно заданного контрольного значения.

- Выберите с помощью меню *Анализ/Сравнение средних/Одновыборочный Т-критерий*;
- В открывшемся диалоговом окне *Одновыборочный Т-критерий* перенесите изучаемую переменную в поле *Проверять переменные* и введите контрольное значение в поле *Проверяемое значение*. Запустите вычисления, щелкнув *ОК* (рис.4.10).
- Кнопкой *Параметры* можно задать вместо 95% любой другой доверительный интервал. Значение доверительного интервала может принимать значения в промежутке от 1 до 99%.

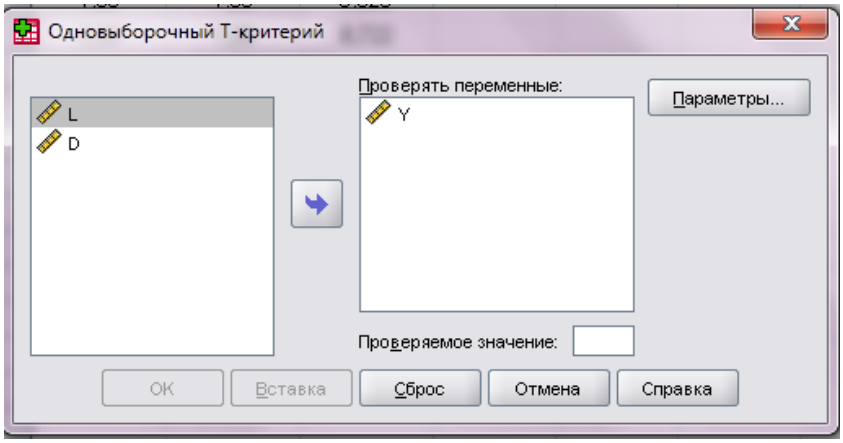


Рис.4.10. Диалоговое окно «Одновыборочный Т-критерий»

- Введите несколько проверяемых значений. Проанализируйте полученные результаты.

6. Общие задания

1. Определите, есть ли зависимость между предпочтениями людей при выборе напитка и объемом потребления жидкости (файл «*adstudy*»). Если есть, то какой напиток утоляет жажду в меньшей степени?
2. Исследуйте эффективность рекламной кампании, изучив ежедневный оборот фирмы за 15 дней до и после публикации рекламы (файл «*реклам_компания*»).
3. Получите индивидуальное задание и определите, какие из рассмотренных выше статистических методик применимы к вашим данным. Ответ обоснуйте.
4. Проведите необходимые расчеты по вашим данным и дайте анализ полученным результатам.

7. Требования к отчету

Отчет по лабораторной работе №4 предоставляется в печатном виде и должен содержать:

1. Ответы на контрольные вопросы.
2. Анализ результатов, полученных после выполнения практических заданий 1 и 2.
3. Исходные данные, полученные вами у преподавателя.
4. Результаты выполнения индивидуального практического задания.
5. Выводы по полученной информации.

8. Индивидуальные задания

1÷2. По территориям Центрального района известны данные за 1995 г.

Район	Среднемесячная начисленная заработная плата, тыс. руб., x	Доля денежных доходов, направленных на прирост сбережений во вкладах, займах, сертификатах и на покупку валюты, в общей сумме среднедушевого денежного дохода, %, y
Брянская обл.	289	6,9
Владимировская обл.	334	8,7
Ивановская обл.	300	6,4
Калужская обл.	343	8.4
Костромская обл.	356	6,1
Орловская обл.	289	9,4
Рязанская обл.	341	11,0
Смоленская обл.	327	6,4
Тверская обл.	357	9,3
Тульская обл.	352	8,2
Ярославская обл.	381	8,6

3÷4. По территориям Центрального района известны данные за 1995 г.

Район	Прожиточный минимум в среднем на одного пенсионера в месяц, тыс. руб., x	Средний размер назначенных ежемесячных пенсий, тыс. руб., y
Брянская обл.	178	240
Владимирская обл.	202	226

Ивановская обл.	197	221
Калужская обл.	201	226
Костромская обл.	189	220
г.Москва	302	250
Московская обл.	215	237
Орловская обл.	166	232
Рязанская обл.	199	215
Смоленская обл.	180	220
Тверская обл.	181	222
Тульская обл.	186	231
Ярославская обл.	250	229

5÷6. По территориям Центрального и Волго-Вятского районов известны данные за ноябрь 1997г.

Район	Прожиточный минимум в среднем на душу населения, тыс. руб., x	Средняя заработная плата и выплаты социальнога характера, тыс. руб., y
Брянская обл.	289	615
Владимирская обл.	338	727
Ивановская обл.	287	584
Калужская обл.	324	753
Костромская обл.	307	707
Орловская обл.	304	657
Рязанская обл.	307	654
Смоленская обл.	290	693
Тверская обл.	314	704
Тульская обл.	304	780
Ярославская обл.	341	830
Респ. Марий Эл	364	554
Респ. Мордовия	342	560
Чувашская Респ.	310	545
Кировская обл.	411	672
Нижегородская обл.	304	796

7÷8. По территориям Волго-Вятского, Центрально-Черноземного и Поволжского районов известны данные за ноябрь 1997 г.

Район	Средняя заработная плата и выплаты социального характера, тыс. руб.,x	Потребительские расходы в расчете на душу населения, тыс. руб.,y

Респ. Марий Эл	554	302
Респ. Мордовия	560	360
Чувашская Респ.	545	310
Кировская обл.	672	415
Нижегородская обл.	796	452
Белгородская обл.	777	502
Воронежская обл.	632	355
Курская обл.	688	416
Липецкая обл.	833	501
Тамбовская обл.	577	403
Респ. Калмыкия	584	208
Респ. Татарстан	949	462
Астраханская обл.	888	368
Волгоградская обл.	831	399
Пензенская обл.	562	342
Саратовская обл.	665	354
Ульяновская обл.	705	558

9÷10. По территории Северного, Северо-Западного и Центрального районов известны данные за ноябрь 1997 г.

Район	Денежные доходы на душу населения, тыс. руб., x	*Потребительские расходы на душу населения, тыс. руб.,y*
Респ. Карелия	913	596
Респ. Коми	1095	417
Архангельская обл	606	354
Вологодская обл.	876	526
Мурманская обл.	1314	934
Ленинградская обл	593	412
Новгородская обл.	754	525
Псковская обл.	528	367
Брянская обл.	520	364
Владимирская обл	539	336
Ивановская обл.	540	409
Калужская обл.	682	452
Костромская обл.	537	367
Московская обл.	589	328
Орловская обл.	626	460
Рязанская обл.	521	380
Смоленская обл.	626	439

Тверская обл.	521	344
Тульская обл.	658	401
Ярославская обл.	746	514

11÷12. По территориям Восточно-Сибирского и Дальневосточного районов известны данные за ноябрь 1997 г.

Район	Денежные доходы на душу населения, тыс. руб., x	*Потребительские расходы на душу населения, тыс. руб.,y*
Респ. Бурятия	524	408
Респ. Тыва	371	249
Респ. Хакасия	453	253
Красноярский край	1006	580
Иркутская обл.	997	651
Усть-Ордынский Бурятский авт. округ	217	139
Читинская обл.	486	322
Респ. Саха (Якутия)	1989	899
Еврейская авт. обл.	595	330
Чукотский авт. округ	1550	446
Приморский край	937	642
Хабаровский край	761	542
Амурская обл.	767	504
Камчатская обл.	1720	861
Магаданская обл.	1735	707
Сахалинская обл.	1052	557

13÷ 14. По территориям Уральского и Западно-Сибирского районов известны данные за ноябрь 1997 г.

Район	Денежные доходы на душу населения, тыс. руб., x	Потребительские расходы на душу населения, тыс. руб., y
Респ. Башкортостан	632	461
Удмуртская Респ.	738	524
Курганская обл.	515	298
Оренбургская обл.	640	351
Пермская обл.	942	624
Свердловская обл.	888	584
Челябинская обл.	704	425
Респ. Алтай	603	277
Алтайский край	439	321
Кемеровская обл.	985	573
Новосибирская обл.	735	576

Омская обл.	760	588
Томская обл.	830	497
Тюменская обл.	2093	863

15÷17. По территориям Уральского и Западно-Сибирского районов известны данные за ноябрь 1997г.

Район	Средняя заработная плата и выплаты социального характера тыс. руб.,x	Потребительские расходы в расчете на душу населения, тыс. руб.,y
Репс. Башкортостан	912	461
Удмуртская Респ.	809	524
Курганская обл.	748	298
Оренбургская обл.	847	351
Пермская обл.	1087	624
Свердловская обл.	1074	584
Челябинская обл.	1008	425
Респ. Алтай	682	277
Алтайский край	697	321
Кемеровская обл.	1251	573
Новосибирская обл.	967	576
Омская обл.	898	588
Томская обл.	1263	497
Тюменская обл.	3027	863

18÷20. По территориям Уральского и Западно-Сибирского районов известны данные за ноябрь 1997 г.

Район	Денежные доходы на душу населения, тыс. руб., x	Потребительские расходы на душу населения, тыс. руб., y
Респ. Башкортостан	632	461
Удмуртская Респ.	738	524
Курганская обл.	515	298
Оренбургская обл.	640	351
Пермская обл.	942	624
Свердловская обл.	888	584
Челябинская обл.	704	425
Респ. Алтай	603	277
Алтайский край	439	321
Кемеровская обл.	985	573
Новосибирская обл.	735	576
Омская обл.	760	588
Томская обл.	830	497
Тюменская обл.	2093	863

9. Контрольные вопросы

1. Для чего применяется сравнение средних?
2. Для какого типа переменных применяется сравнение средних значений?
3. Как формально определяется t-критерий (Стьюдента) для двух выборок?
4. Как провести различные t-тесты в программе SPSS Statistics?
5. Как проверяется гипотеза о равенстве дисперсий двух нормальных выборок?
6. Что показывает уровень значимости?
7. Какую гипотезу проверяет тест Ливня?
8. Что позволяет исследовать t-тест для одной выборки?

Лабораторная работа № 5

Тема: Корреляционный анализ.

Цель: Научиться определять меры тесноты связи между статистическими данными.

1. Введение

Степень зависимости (связь) между двумя переменными называется **корреляцией**.

В корреляционных связях между изменением факторного и результативного признака нет полного соответствия, воздействие отдельных факторов проявляется лишь в среднем при массовом наблюдении фактических данных. В простейшем случае применения корреляционной зависимости величина результативного признака рассматривается как следствие изменения только одного фактора (например, энерговооруженность труда рассматривается как причина роста производительности труда).

Простейшим примером обнаружения связи является сопоставление двух параллельных рядов - ряда значений факторного признака и соответствующих ему значений результативного признака. Значения факторного признака располагают в возрастающем порядке и затем прослеживают направление изменения величины результативного признака. Результативный признак (функцию) чаще всего обозначают через Y, а факторный признак - через X.

Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные.

- Переменные с интервальной и с номинальной шкалой: коэффициент корреляции Пирсона.
- Хотя бы одна из двух переменных имеет порядковую шкалу либо не является нормально распределенной: ранговая корреляция Спирмана или τ (тау) Кендалла.
- Одна из двух переменных является дихотомической: точечная двухрядная корреляция. Эта возможность в SPSS Statistics отсутствует. Вместо этого может быть применен расчет ранговой корреляции.

Дихотомической (бинарной) называется переменная с номинальной шкалой, имеющая две категории. Например, дихотомической будет переменная «пол», которой можно присвоить категории: 1=мужской, 2=женский.

- Обе переменные являются дихотомическими: четырёхполевая корреляция. Данный вид корреляции рассчитывается в SPSS Statistics на основании определения мер расстояния и мер сходства.

Расчет коэффициента корреляции между двумя недихотомическими переменными не лишен смысла только тогда, когда связь между ними линейна. Если, например, U-образная (неоднозначная), то коэффициент корреляции непригоден для использования в качестве меры силы связи: его значение стремится к нулю.

2. Графическое представление корреляционной зависимости

Для графического представления корреляционной связи можно использовать прямоугольную систему координат с осями, которые соответствуют обеим переменным. Каждая пара значений маркируется при помощи определенного символа. Такой график называется **диаграммой рассеяния**. Она строится для переменных, которые как минимум относятся к интервальной шкале.

Для построения диаграммы рассеяния, после открытия необходимого файла SPSS Statistics, выберите в меню *Графика*/*Устаревшие диалоговые окна/Рассеяния/Точки*. Откроется диалоговое окно *Рассеяния/Точки* (рис.5.1), предоставляющее различные возможности построения диаграмм рассеяния.

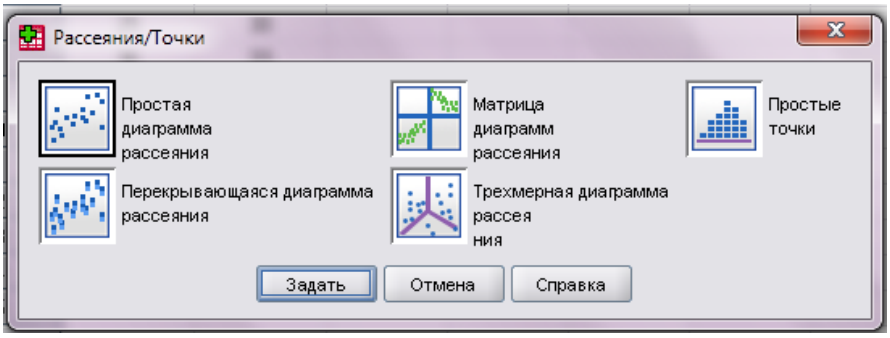


Рис.5.1. Диалоговое окно «Рассеяния/Точки»

2.1. Простая диаграмма рассеяния

В диалоговом окне *Рассеяния/Точки* щелкните на области *Простая диаграмма рассеяния* и кнопке *Задать*. Открывшееся диалоговое окно позволяет создать график зависимости между двумя переменными. Из списка исходных переменных перенесите те, которые подлежат исследованию соответственно в поле *Ось Y* и в поле *Ось X* и нажмите *OK*.

На диаграмме рассеяния можно наносить регрессионные линии. Для этого в окне просмотра результатов двойным щелком перенесите диаграмму рассеяния в редактор диаграмм.

- В списке меню редактора диаграмм выберите *Элементы/Линия аппроксимации для итога*. Откроется диалоговое окно *Свойства* (рис.5.2).
- На вкладке *Линия аппроксимации* выберите метод аппроксимации *Линейная регрессия* и в группе *Доверительные интервалы* отметьте опцию *Среднее*. Так для регрессионной прямой вы получите 95% доверительный интервал.
- Покиньте диалоговое окно нажатием *Применить*.

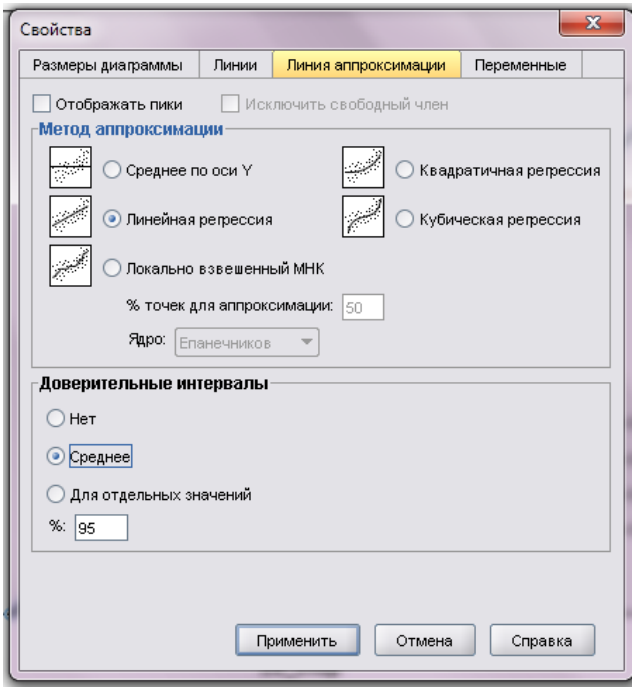


Рис.5.2. Диалоговое окно редактора диаграмм «Свойства»

Теперь на рассматриваемой диаграмме рассеяния присутствуют регрессионные прямые и соответствующие им доверительные интервалы (рис.5.3).

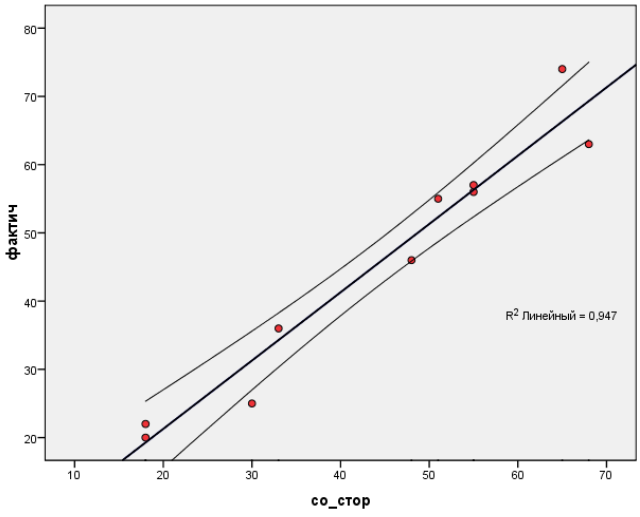


Рис.5.3. Диаграмма рассеяния

2.2. Матричные диаграммы рассеяния

Этот метод применяется для отображения нескольких диаграмм рассеяния на одном графике.

В диалоговом окне *Рассеяния/Точки* (см. рис.5.1) щелкните на области *Матрица диаграмм рассеяния* и кнопке *Задать*. Теперь поочередно перенесите все необходимые переменные в поле *Матричные переменные*, и нажмите *OK*.

Число строк и столбцов в матричной диаграмме соответствует количеству переменных. Каждая ячейка является диаграммой рассеяния для одной пары переменных. Диагональные ячейки содержат метки переменных, находящихся в соответствующих ячейках матрицы.

Для матричных диаграмм также можно организовать построение различных линий регрессии (рис.5.4).

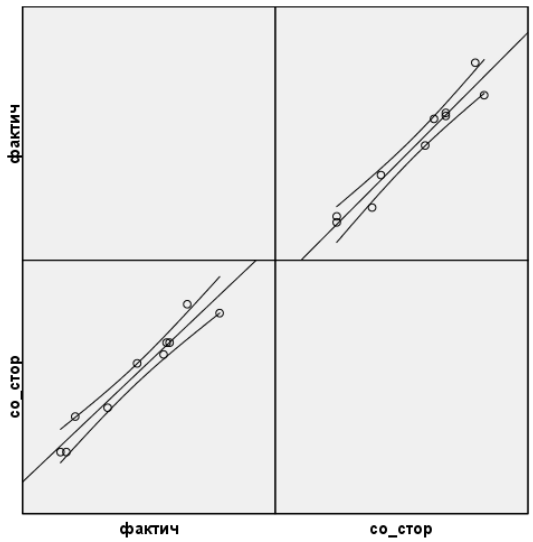


Рис.4. Матричная диаграмма рассеяния

2.3. Наложенные диаграммы рассеяния

В одном графике можно представить несколько диаграмм рассеяния. Для этого в диалоговом окне *Рассеяния/Точки* (см. рис.5.1) щелкните на области Перекрывающаяся диаграмма рассеяния и кнопке *Задать*. В появившемся диалоговом окне могут быть заданы соответствующие X-Y пары переменных, которые должны быть представлены вместе. Значения, принадлежащие соответствующей паре, на диаграмме будут отмечены одной определенной маркировкой (рис.5.5).

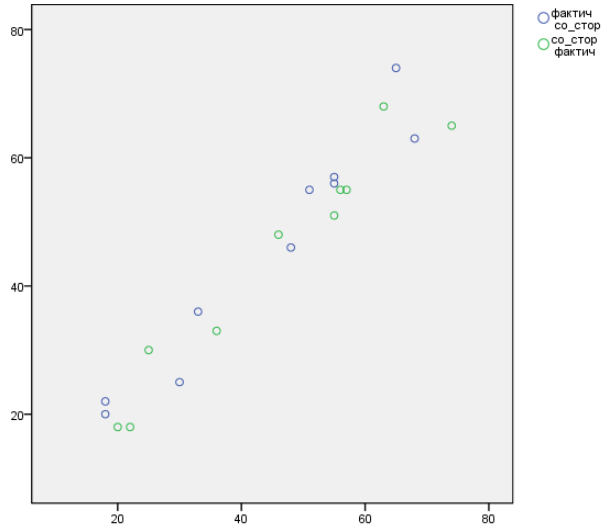


Рис.5.5. Перекрывающаяся диаграмма рассеяния

Этот метод имеет смысл применять только тогда, когда речь идет о переменных с одними и теми же областями значений.

В диалоговом окне *Рассеяния/Точки* можно также построить трехмерные диаграммы рассеяния. Они строятся на основании значений трех переменных и поэтому включают три оси.

3. Коэффициент корреляции Пирсона

Он также носит название **линейного корреляционного коэффициента** и рассчитывается по следующей формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

где x_i и y_i - значения двух переменных,

$\bar{x}$ и $\bar{y}$ - их средние значения,

σ_x и σ_y - стандартные отклонения,

n – количество пар значений (объем выборки).

Проведем исследование примера, взятого из металлургического производства – сталеплавильного цеха. Определим, существует ли корреляционная зависимость между потерей энергии при плавке стали и содержанием серы (файл «сера_энергия»).

В данном примере энергетические потери определяются процентным содержанием серы в металле, т.е. выборка, показывающая количество серы, будет факторным признаком (независимая), а выборка потери энергии – результативным признаком (зависимая).

Вычисление коэффициента корреляции Пирсона в SPSS Statistics осуществляется так:

- Выберите в меню *Анализ/Корреляции/Парные*;
- Появится диалоговое окно *Парные корреляции*.
- Все переменные, для которых необходимо вычислить попарно коэффициент корреляции, перенесите по очереди в поле тестируемых переменных. Расчет коэффициента корреляции *Пирсона* является предварительной установкой, также как *двусторонняя* проверка значимости и *маркировка значимых корреляций* (рис.5.6).
- Расчет начнется после нажатия *OK*.

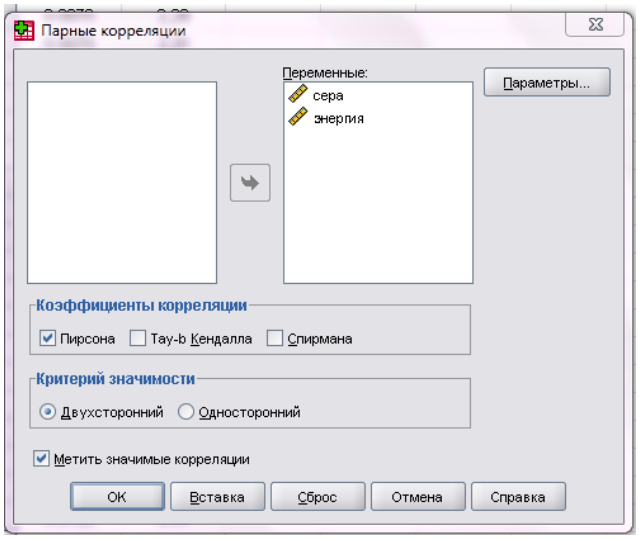


Рис.5.6. Диалоговое окно «Парные корреляции»

Для нашего примера будет сформирована корреляционная матрица размерностью 2×2 (по количеству тестируемых переменных) (рис.5.7).

Корреляции

		сера	энергия
сера	Корреляция Пирсона	1	,788**
	Знч.(2-сторон)		,000
	N	76	76
энергия	Корреляция Пирсона	,788**	1
	Знч.(2-сторон)	,000	
	N	76	76

**. Корреляция значима на уровне 0.01 (2-сторон.).

Рис.5.7. Корреляционная матрица Пирсона

Полученные результаты содержат: корреляционный коэффициент Пирсона r, количество использованных пар значений переменных (N=76) и вероятность ошибки p, соответствующая предположению о ненулевой корреляции. В приведенном примере существует сильная корреляция (r=0,788), поэтому все коэффициенты являются сверхзначимыми (p<0,001). Следовательно, маркировка корреляции, приведенная ниже таблицы, должна

была бы состоять из трех звездочек, которыми обозначается уровень p=0,001.

При помощи щелчка на кнопке *Параметры* в диалоговом окне *Парные корреляции* (рис.5.8) можно организовать расчет среднего значения и стандартного отклонения. дополнительно могут выводиться отклонения произведений моментов (значение числителя в формуле для коэффициента корреляции) и элементы ковариационной матрицы (числитель, деленный на n-1).

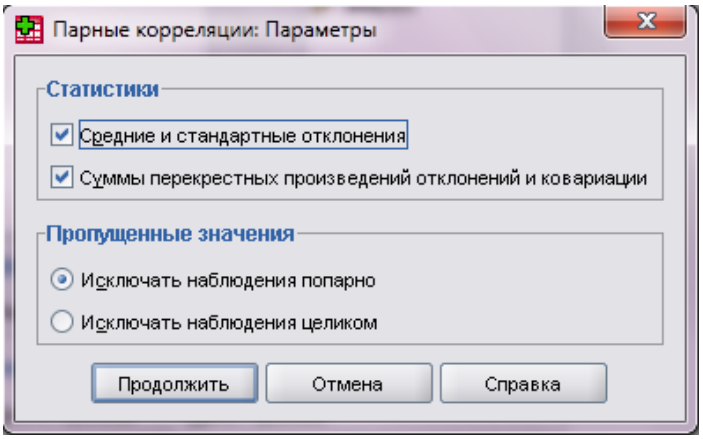


Рис.5.8. Диалоговое окно «Парные корреляции: Параметры»

4. Ранговые коэффициенты корреляции по Спирману и Кендаллу

Для переменных, принадлежащих к порядковой шкале, или для переменных, не подчиняющихся нормальному распределению, а также для переменных с интервальной шкалой вместо коэффициента Пирсона рассчитывается ранговая корреляция по Спирману. Для этого отдельным значениям переменных присваиваются ранговые места, которые впоследствии обрабатываются с помощью соответствующих формул. Чтобы выявить ранговую корреляцию, уберите в диалоговом окне *Парные корреляции* метку для расчета корреляции Пирсона, установленную по умолчанию. Вместо этого активируйте расчет корреляции *Спирмана* и нажмите *ОК*.

Для переменных «*сера-энергия*» получим корреляционную матрицу (рис.5.9).

Корреляции

			сера	энергия
ро Спирмена	сера	Коэффициент корреляции	1,000	,793**
		Знч. (2-сторон)	.	,000
		N	76	76
	энергия	Коэффициент корреляции	,793**	1,000
		Знч. (2-сторон)	,000	.
		N	76	76

**. Корреляция значима на уровне 0.01 (2-сторонняя).

Рис.5.9. Корреляционная матрица Спирмена

Вы видите, что корреляция между расходом энергии и содержанием серы высокозначима, т.е. повышение концентрации серы в металле требует больше энергии (положительная связь).

Корреляция r Пирсона между переменными (r=0,788) меньше ранговой корреляции Спирмана (0,793). Это значит, что рассмотрение рангов (а не самих наблюдений) в действительности улучшает оценку зависимости между переменными, т.к. «подавляет» случайную изменчивость и уменьшает воздействия выбросов. Чем больше объем располагаемых данных, тем более заметна разница между коэффициентами.

Ещё одним вариантом ранговой корреляции является коэффициент Кендалла, расчет которого вызывается в диалоговом окне *Парные корреляции*. Чаще всего коэффициенты Кендалла значительно ниже корреляционных коэффициентов Спирмана. Убедитесь в этом самостоятельно и проанализируйте результаты.

5. Частная корреляция

Если исследовать достаточно большую совокупность мужчин и сопоставить размер их обуви с уровнем образованности, то между этими двумя переменными можно заметить хоть и небольшую, но в то же время значимую корреляцию. Это пример так называемой ложной корреляции. Здесь статистически значимый коэффициент корреляции является не проявлением некоторой причинной связи между двумя рассматриваемыми переменными, а в большей степени обусловлен некоторой третьей переменной.

В рассматриваемом примере (файл «обувь_образов») такой переменной является рост. С одной стороны, существует некоторая незначительная корреляция между ростом и уровнем образованности, а с другой – вполне объяснимая и логичная связь между ростом и размером обуви. Вместе эти две корреляции приводят к ложной корреляции. Для исключения одной такой искажающей переменной необходим расчет так называемой частной корреляции.

Если присвоить коррелирующим переменным индексы 1 и 2, а искажающей переменной – индекс 3, и попарно рассчитать корреляционный коэффициент (Пирсона) r_{12}, r_{13} и r_{23}, то для частных коэффициентов получим:

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Расчет частных корреляций в SPSS Statistics производится в меню

Анализ/Корреляции/Частные.

В появившемся диалоговом окне *Частные корреляции* перенесите коррелирующие переменные в поле признаков, а искажающую переменную – в поле контрольных переменных и оставьте предварительную установку для двухстороннего теста значимости (рис.5.10).

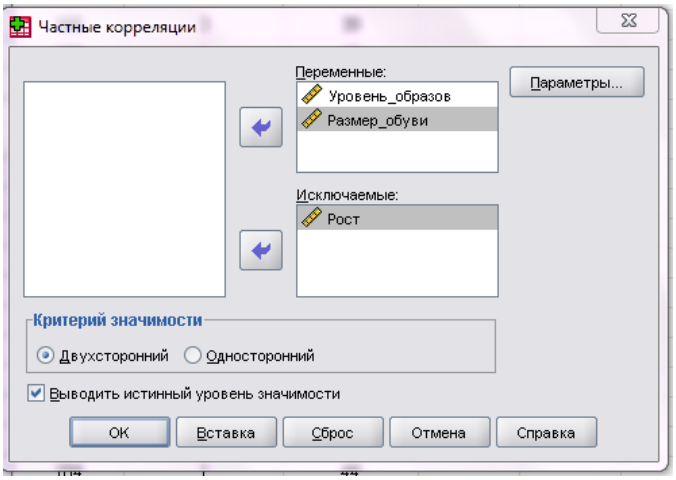


Рис.5.10. Диалоговое окно «Частные корреляции»

Кнопкой *Параметры* наряду с традиционной обработкой пропущенных значений, можно организовать расчет среднего значения, стандартного отклонения и вывод «корреляций нулевого порядка» (т.е. простых корреляционных коэффициентов). Нажмите *OK*.

Появятся результаты (рис.5.11):

- частный корреляционный коэффициент;
- число степеней свободы (число наблюдений минус 3);
- уровень значимости.

Описательные статистики

	Среднее	Стд. Отклонение	N
Уровень_образов	1,74	,806	19
Размер_обуви	43,37	1,535	19
Рост	180,95	8,370	19

Корреляции

Контрольные переменные			Уровень_образов	Размер_обуви	Рост
-ничего-[a]	Уровень_образов	Корреляция	1,000	-,456	-,159
		Значимость (2-сторон.)	.	,050	,516
		ст.св.	0	17	17
	Размер_обуви	Корреляция	-,456	1,000	,425
		Значимость (2-сторон.)	,050	.	,069
		ст.св.	17	0	17
	Рост	Корреляция	-,159	,425	1,000
		Значимость (2-сторон.)	,516	,069	.
		ст.св.	17	17	0
Рост	Уровень_образов	Корреляция	1,000	-,435	
		Значимость (2-сторон.)	.	,071	
		ст.св.	0	16	
	Размер_обуви	Корреляция	-,435	1,000	
		Значимость (2-сторон.)	,071	.	
		ст.св.	16	0	

a. Ячейки содержат корреляции нулевого порядка (Пирсон).

Рис.5.11. Матрица частной корреляции

6. Внутриклассовый коэффициент корреляции (Intraclass Correlation Coefficient (ICC))

Внутриклассовый коэффициент корреляции (ICC) принимает значения в диапазоне между –1 и +1. Он применяется в качестве меры связанности с том случае, когда согласованность двух признаков должна быть проверена не так, как при расчете рассмотренных выше корреляционных коэффициентов, относительно её общей направленности («чем больше одна переменная, тем больше вторая»), а также и относительно средних уровней обеих переменных.

Таким образом, расчет ICC считается уместным только тогда, когда обе переменные имеют приблизительно одинаковый уровень значений. Подобная ситуация вероятнее всего возникает в случае, когда одной и той же величине дается двоякая оценка.

Рассмотрим расчет ICC на данных типичного примера. В файле «возраст» находятся две переменные (табл.5.1): фактический возраст и возраст по оценке со стороны.

Таблица 5.1

Данные для расчета внутриклассового коэффициента корреляции

Фактический возраст	Возраст по оценке со стороны
22	18
55	51
56	55
57	55
63	68
74	65
25	30
36	33
46	48
20	18

Если произвести расчет корреляции по Пирсону, то получим значение r=0,973. Определим теперь ICC:

- выберите в меню *Анализ/Шкалирование/Анализ пригодности*;
- перенесите обе переменные в список объектов;
- через кнопку *Статистики* активируйте опцию *Внутриклассовые коэффициенты корреляции*;
- в качестве модели выберите *Однократная случайная*, которая соответствует традиционному расчету ICC;
- оставьте предварительно установленный 95% доверительный интервал и подтвердите нажатием *Продолжить* и *OK*.

В окне просмотра *Вывод* появятся результаты (рис.5.12).

Сводка обработки наблюдений

		N	%
Наблюдения	Валидные	10	100,0
	Исключенные[a]	0	,0
	Итого	10	100,0

а. Сплошное исключение основано на всех переменных в процедуре.

Статистики пригодности

Альфа Кронбаха	Количество пунктов
,986	2

Внутриклассовый коэффициент корреляции

	Внутриклассовая корреляция	95% доверительный интервал		F-критерий с истинным значением 0			
		Нижняя граница	Верхняя граница	Значение	ст.св.1	ст.св.2	Знач.
Простые меры	,973	,902	,993	72,952	9	10	,000
Средние меры	,986	,948	,997	72,952	9	10	,000

Однофакторная модель случайных эффектов, в которой эффекты испытуемых являются случайными.

Рис.5.12. Результаты внутриклассового коэффициента корреляции

Результаты обычного расчета ICC находятся под заголовком «Внутриклассовая корреляция». Значение коэффициента составило ICC=0,973, которое с 95% доверительным интервалом принадлежит к диапазону от 0,902 до 0,993. В данном случае это значение не просто очень близко, а совпадает с корреляционным коэффициентом Пирсона.

Ещё одним типичным примером для расчета ICC является определение связей между фактическим весом и весом по оценке со стороны или фактическим и оценочным ростом.

7. Общие задания

1. Для данных файла «лазер.sav» (см. лабораторную работу №3) установите, между какими переменными имеется корреляционная зависимость, и какие, при этом, значения принимает коэффициент корреляции Пирсона.
2. Для того же примера определите коэффициенты ранговой корреляции.
3. Отобразите исходные данные в виде простых диаграмм рассеяния и матричных диаграмм.
4. Определите, для каких переменных указанного файла возможно построить наложенную диаграмму рассеяния. Постройте её, а также трехмерную диаграмму рассеяния.
5. Рассчитайте ICC по данным фактического и оценочного роста студентов своей группы.

6. Выполните расчет всех корреляционных коэффициентов, которые подходят для вашего индивидуального задания.
7. Сделайте вывод по полученным результатам.
8. Представьте графически исходные данные.

8. Требования к отчету

Отчет по лабораторной работе №5 предоставляется в письменном виде и должен содержать:

1. Ответы на контрольные вопросы № 1 - 4, 7.
2. Анализ результатов, полученных после выполнения практических заданий.
3. Исходные данные по заданию №5 и результат его выполнения.
4. Исходные данные и результаты выполнения индивидуального практического задания.
5. Выводы по полученной информации.

9. Индивидуальные задания

1÷2. Имеются данные о деятельности крупнейших компаний США в 1996 г.

№ п/п	Чистый доход, млрд. долл. США, y	Оборот капитала, млрд. долл. США, x_1	Использованный капитал, млрд. долл. США, x_2	Численность служащих, тыс. чел., x_3	Рыночня капитлизация компании, млрд. долл. США, x_4
1	0,9	31,3	18,9	43,0	40,9
2	1,7	13,4	13,7	64.7	40,5
3	0,7	4,5	18,5	24,0	38,5
4	1,7	10,0	4,8	50,2	38,9
5	2,6	20,0	21,8	106,0	37,3
6	1,3	15,0	5,8	96,6	26,5
7	4,1	137,1	99,0	347,0	37,0
8	1,6	17,9	20,1	85,6	36,8
9	6,9	165,4	60,6	745,0	36.3
10	0,4	2,0	1,4	4,1	35.3
11	1,3	6,8	8,0	26,8	35,3
12	1,9	27,1	18,9	41,7	35,0
13	1,9	13,4	13,2	61,8	26,2
14	1,4	9,8	12,6	212,0	33,1
15	0,4	19,5	12,2	105,0	32,7
16	0,8	6,8	3,2	33,5	32,1
17	1,8	27,0	13.0	142,0	30,5
18	0,9	12,4	6,9	96,0	29,8
19	1,1	17,7	15,0	140,0	25,4
20	1,9	12,7	11,9	59,3	29,3

21	-0,9	21,4	1,6	131,0	29,2
22	1,3	13,5	8,6	70,7	29,2
23	2,0	13,4	11.5	65,4	29,1
24	0,6	4,2	1,9	23.1	27,9
25	0,7	15,5	5,8	80,8	27,2

3÷4. По данным, представленным в табл. 2.19, изучается зависимость индекса человеческого развития y от переменных:

X_1 - ВВП 1997 г., % к 1990 г.;

X_2 - расходы на конечное потребление в текущих ценах, % к ВВП;

X_3 - расходы домашних хозяйств, % к ВВП;

X_4 - валовое накопление, % к ВВП;

Страна	У	X_1	X_2	X_3	X_4
Австрия	0,904	115,0	75,5	56,1	25,2
Австралия	0,922	123,0	78,5	61,8	21,8
Белоруссия	0,763	74,0	78,4	59,1	25,7
Бельгия	0,923	111,0	77,7	63,3	17,8
Великобритания	0,918	113,0	84,4	64,1	15,9
Германия	0,906	110,0	75,9	57,0	22,4
Дания	0,905	119,0	76,0	50,7	20,6
Индия	0,545	146,0	67,5	57,1	25,2
Испания	0,894	113,0	78,2	62,0	20,7
Италия	0,900	108,0	78,1	61,8	17,5
Канада	0,932	113,0	78,6	58,6	19,7
Казахстан	0,740	71,0	84,0	71,7	18,5
Китай	0,701	210,0	59,2	48,0	42,4
Латвия	0,744	94,0	90,2	63,9	23,0
Нидерланды	0,921	118,0	72,8	59,1	20,2
Норвегия	0,927	130,0	67,7	47,5	25,2
Польша	0,802	127,0	82,6	65,3	22,4
Россия	0,747	61,0	74,4	53,2	22,7
США	0,927	117,0	83,3	67,9	18,1
Украина	0,721	46,0	83,7	61,7	20,1
Финляндия	0,913	107,0	73,8	52,9	17,3
Франция	0,918	110,0	79,2	59,9	16,8
Чехия	0,833	99,2	71,5	51,5	29,9
Швейцария	0,914	101,0	75,3	61,2	20,3
Швеция	0,923	105,0	79,0	53,1	14,1

5÷6. Имеются данные о продаже квартир на вторичном рынке жилья в Санкт-Петербурге на 01.05.2000 г.

№ п/п	y	X_1	X_2	X_3	X_4
1	13,0	37,0	21,5	6,5	20
2	16,5	60,0	27,0	22,4	10
3	17,0	60,0	30,0	15,0	10
4	15,0	53,0	26,2	13,0	15
5	14,2	35,0	19,0	9,0	8

6	10,5	30,3	17,5	5,6	15
7	23,0	43,0	25,5	8,5	5
8	12,0	30,0	17,8	5,5	10
9	15,6	35,0	18,0	5,3	3
10	12,5	32,0	17,0	6,0	5
11	11,3	31,0	18,0	5,5	10
12	13,0	33,0	19,6	7,0	5
13	21,0	53,0	26,0	16,0	5
14	12,0	32,2	18,0	6,3	20
15	11,0	31,0	17,3	5,5	15
16	11,0	36,0	19,0	8,0	5
17	22,5	48,0	29,0	8,0	15
18	26,0	55,5	35,0	8,0	10
19	18,5	48,0	28,0	8,0	10
20	13,2	44,1	30,0	6,0	25
21	25,8	80,0	51,0	13,0	10

Принятые в таблице обозначения:

Y - цена квартиры, тыс. долл.;

X_1 - общая площадь квартиры ($м^2$);

X_2 - жилая площадь квартиры ($м^2$);

X_3 - площадь кухни ($м^2$);

X_4 - расстояние от метро, минут пешком.

7÷8. Изучается зависимость средней ожидаемой продолжительности жизни от нескольких факторов по данным за 1995 г.

Страна	y	X_1	X_2	X_3	X_4
Мозамбик	47	3,0	2,6	2,4	113
Бурунди	49	2,3	2,6	2,7	98
Чад	48	2,6	2,5	2,5	117
Непал	55	4,3	2,5	2,4	91
Буркина-Фасо	49	2,9	2,8	2,1	99
Мадагаскар	52	2,4	3,1	3,1	89
Бангладеш	58	5,1	1,6	2,1	79
Гаити	57	3,4	2,0	1,7	72
Мали	50	2,0	2,9	2,7	123
Нигерия	53	4,5	2,9	2,8	80
Кения	58	5,1	2,7	2,7	58
Того	56	4,2	3,0	2,8	88
Индия	62	5,2	1,8	2,0	68
Бенин	50	6,5	2,9	2,5	95
Никарагуа	68	7,4	3,1	4,0	46
Гана	59	7,4	2,8	2,7	73
Ангола	47	4,9	3,1	2,8	124
Пакистан	60	8,3	2,9	3,3	90
Мавритания	51	5,7	2,5	2,7	96
Зимбабве	57	7,5	2,4	2,2	55
Гондурас	67	7,0	3,0	3,8	45
Китай	69	10,8	1,1	1,1	34
Камерун	57	7,8	2,9	3,1	56

Конго	51	7,6	2,9	2,6	90
Шри-Ланка	72	12,1	1,3	2,0	16
Египет	63	14,2	2,0	2,7	56
Индонезия	64	14,1	1,6	2,5	51
Филиппины	66	10,6	2,2	2,7	39
Морокко	65	12,4	2,0	2,6	55
Гватемала	66	12,4	2,9	3,5	44
Эквадор	69	15,6	2,2	3,2	36
Ямайка	74	13,1	1,0	1,8	13
Алжир	70	19,6	2,2	4,1	34
Парагвай	68	13,5	2,7	2,9	41
Тунис	69	18,5	1,9	3,0	39

Принятые в таблице обозначения:

Y - средняя ожидаемая продолжительность жизни при рождении, лет;

x_1 - ВВП в паритетах покупательной способности;

x_2 - темпы прироста населения по сравнению с предыдущим годом, %;

x_3 - темпы прироста рабочей силы по сравнению с предыдущим годом, %;

x_4 -коэффициент младенческой смертности, %.

9÷10. Изучается зависимость средней ожидаемой продолжительности жизни от нескольких факторов по данным за 1995 г.

Страна	y	X_1	X_2	X_3	X_4
Белоруссия	70	15,6	0,2	0,2	13
Перу	66	14,0	2,0	3,1	47
Тайланд	69	28,0	0,9	1,3	35
Панама	73	22,2	1,7	2,4	23
Турция	67	20,7	1,7	2,1	48
Польша	70	20,0	0,3	0,6	14
Словакия	72	13,4	0,3	0,7	11
Венесуэла	71	29,3	2,3	3,0	23
ЮАР	64	18,6	2,2	2,4	50
Мексика	72	23,7	1,9	2,8	33
Мавритания	71	49,0	1,3	1,8	16
Бразилия	67	20,0	1,5	1,6	44
Тринидад	72	31,9	0,8	1,8	13
Малайзия	71	33,4	2,4	2,7	12
Чили	72	35,3	1,5	2,1	12
Уругвай	73	24,6	0,6	1,0	18
Аргентина	73	30,8	1,3	2,0	22
Греция	78	43,4	0,6	0,9	8
Испания	77	53,8	0,2	1,0	7
Нов.Зеландия	76	60,6	1,4	1,5	7
Ирландия	77	58,1	0,5	1,7	6
Израиль	77	61,1	3,5	3,5	8
Австралия	77	70,2	1,1	1,4	6
Италия	78	73,7	0,2	0,1	7
Канада	78	78,3	1,3	1,0	6
Финляндия	76	65,8	0,5	0,1	5

Гонконг	79	85,1	1,6	1,3	5
Швеция	79	68,7	0,6	0,3	4
Нидерланды	78	73,9	0,7	0,6	6
Бельгия	77	80,3	0,4	0,5	8
Франция	78	78,0	0,5	0,8	6
Сингапур	76	84,4	2,0	1,7	4
Австрия	77	78,8	0,8	0,5	6
США	77	100,0	1,0	1,1	8
Дания	75	78,7	0,3	0,1	6
Япония	80	82,0	0,3	0,6	4
Швейцария	78	95,9	1,0	0,8	6

Принятые в таблице обозначения:

Y - средняя ожидаемая продолжительность жизни при рождении, лет;

X_1 - ВВП в паритетах покупательной способности;

X_2 - темпы прироста населения по сравнению с предыдущим годом, %;

X_3 - темпы прироста рабочей силы по сравнению с предыдущим годом, %;

X_4 - коэффициент младенческой смертности, %.

11÷12. Имеются данные о деятельности крупнейших компаний США в 1996 г.

№ п/п	Чистый доход, млрд. долл. США, y	Оборот капитала, млрд. долл. США, x_1	Использованный капитал, млрд. долл. США, x_2	Численность служащих, тыс. чел., x_3	Рыночня капитлизация компании, млрд. долл. США, x_4
1	0,9	31,3	18,9	43,0	40,9
2	1,7	13,4	13,7	64.7	40,5
3	0,7	4,5	18,5	24,0	38,5
4	1,7	10,0	4,8	50,2	38,9
5	2,6	20,0	21,8	106,0	37,3
6	1,3	15,0	5,8	96,6	26,5
7	4,1	137,1	99,0	347,0	37,0
8	1,6	17,9	20,1	85,6	36,8
9	6,9	165,4	60,6	745,0	36.3
10	0,4	2,0	1,4	4,1	35.3
11	1,3	6,8	8,0	26,8	35,3
12	1,9	27,1	18,9	41,7	35,0
13	1,9	13,4	13,2	61,8	26,2
14	1,4	9,8	12,6	212,0	33,1
15	0,4	19,5	12,2	105,0	32,7
16	0,8	6,8	3,2	33,5	32,1
17	1,8	27,0	13.0	142,0	30,5
18	0,9	12,4	6,9	96,0	29,8
19	1,1	17,7	15,0	140,0	25,4
20	1,9	12,7	11,9	59,3	29,3
21	-0,9	21,4	1,6	131,0	29,2

22	1,3	13,5	8,6	70,7	29,2
23	2,0	13,4	11.5	65,4	29,1
24	0,6	4,2	1,9	23.1	27,9
25	0,7	15,5	5,8	80,8	27,2

13÷14. По данным, представленным в табл. 2.19, изучается зависимость индекса человеческого развития y от переменных:

X_1 - ВВП 1997 г., % к 1990 г.;

X_2 - расходы на конечное потребление в текущих ценах, % к ВВП;

X_3 - расходы домашних хозяйств, % к ВВП;

X_4 - валовое накопление, % к ВВП;

Страна	У	X_1	X_2	X_3	X_4
Австрия	0,904	115,0	75,5	56,1	25,2
Австралия	0,922	123,0	78,5	61,8	21,8
Белоруссия	0,763	74,0	78,4	59,1	25,7
Бельгия	0,923	111,0	77,7	63,3	17,8
Великобритания	0,918	113,0	84,4	64,1	15,9
Германия	0,906	110,0	75,9	57,0	22,4
Дания	0,905	119,0	76,0	50,7	20,6
Индия	0,545	146,0	67,5	57,1	25,2
Испания	0,894	113,0	78,2	62,0	20,7
Италия	0,900	108,0	78,1	61,8	17,5
Канада	0,932	113,0	78,6	58,6	19,7
Казахстан	0,740	71,0	84,0	71,7	18,5
Китай	0,701	210,0	59,2	48,0	42,4
Латвия	0,744	94,0	90,2	63,9	23,0
Нидерланды	0,921	118,0	72,8	59,1	20,2
Норвегия	0,927	130,0	67,7	47,5	25,2
Польша	0,802	127,0	82,6	65,3	22,4
Россия	0,747	61,0	74,4	53,2	22,7
США	0,927	117,0	83,3	67,9	18,1
Украина	0,721	46,0	83,7	61,7	20,1
Финляндия	0,913	107,0	73,8	52,9	17,3
Франция	0,918	110,0	79,2	59,9	16,8
Чехия	0,833	99,2	71,5	51,5	29,9
Швейцария	0,914	101,0	75,3	61,2	20,3
Швеция	0,923	105,0	79,0	53,1	14,1

15÷16. Имеются данные о продаже квартир на вторичном рынке жилья в Санкт-Петербурге на 01.05.2000 г.

№ п/п	y	X_1	X_2	X_3	X_4
1	13,0	37,0	21,5	6,5	20
2	16,5	60,0	27,0	22,4	10
3	17,0	60,0	30,0	15,0	10
4	15,0	53,0	26,2	13,0	15
5	14,2	35,0	19,0	9,0	8
6	10,5	30,3	17,5	5,6	15

7	23,0	43,0	25,5	8,5	5
8	12,0	30,0	17,8	5,5	10
9	15,6	35,0	18,0	5,3	3
10	12,5	32,0	17,0	6,0	5
11	11,3	31,0	18,0	5,5	10
12	13,0	33,0	19,6	7,0	5
13	21,0	53,0	26,0	16,0	5
14	12,0	32,2	18,0	6,3	20
15	11,0	31,0	17,3	5,5	15
16	11,0	36,0	19,0	8,0	5
17	22,5	48,0	29,0	8,0	15
18	26,0	55,5	35,0	8,0	10
19	18,5	48,0	28,0	8,0	10
20	13,2	44,1	30,0	6,0	25
21	25,8	80,0	51,0	13,0	10

Принятые в таблице обозначения:

Y - цена квартиры, тыс. долл.;

X_1 - общая площадь квартиры ($м^2$);

X_2 - жилая площадь квартиры ($м^2$);

X_3 - площадь кухни ($м^2$);

X_4 - расстояние от метро, минут пешком.

17÷18. Изучается зависимость средней ожидаемой продолжительности жизни от нескольких факторов по данным за 1995 г.

Страна	y	X_1	X_2	X_3	X_4
Мозамбик	47	3,0	2,6	2,4	113
Бурунди	49	2,3	2,6	2,7	98
Чад	48	2,6	2,5	2,5	117
Непал	55	4,3	2,5	2,4	91
Буркина-Фасо	49	2,9	2,8	2,1	99
Мадагаскар	52	2,4	3,1	3,1	89
Бангладеш	58	5,1	1,6	2,1	79
Гаити	57	3,4	2,0	1,7	72
Мали	50	2,0	2,9	2,7	123
Нигерия	53	4,5	2,9	2,8	80
Кения	58	5,1	2,7	2,7	58
Того	56	4,2	3,0	2,8	88
Индия	62	5,2	1,8	2,0	68
Бенин	50	6,5	2,9	2,5	95
Никарагуа	68	7,4	3,1	4,0	46
Гана	59	7,4	2,8	2,7	73
Ангола	47	4,9	3,1	2,8	124
Пакистан	60	8,3	2,9	3,3	90
Мавритания	51	5,7	2,5	2,7	96
Зимбабве	57	7,5	2,4	2,2	55
Гондурас	67	7,0	3,0	3,8	45
Китай	69	10,8	1,1	1,1	34
Камерун	57	7,8	2,9	3,1	56
Конго	51	7,6	2,9	2,6	90

Шри-Ланка	72	12,1	1,3	2,0	16
Египет	63	14,2	2,0	2,7	56
Индонезия	64	14,1	1,6	2,5	51
Филиппины	66	10,6	2,2	2,7	39
Морокко	65	12,4	2,0	2,6	55
Гватемала	66	12,4	2,9	3,5	44
Эквадор	69	15,6	2,2	3,2	36
Ямайка	74	13,1	1,0	1,8	13
Алжир	70	19,6	2,2	4,1	34
Парагвай	68	13,5	2,7	2,9	41
Тунис	69	18,5	1,9	3,0	39

Принятые в таблице обозначения:

Y - средняя ожидаемая продолжительность жизни при рождении, лет;

x_1 - ВВП в паритетах покупательной способности;

x_2 - темпы прироста населения по сравнению с предыдущим годом, %;

x_3 - темпы прироста рабочей силы по сравнению с предыдущим годом, %;

x_4 -коэффициент младенческой смертности, %.

19÷20. Изучается зависимость средней ожидаемой продолжительности жизни от нескольких факторов по данным за 1995 г.

Страна	y	X_1	X_2	X_3	X_4
Белоруссия	70	15,6	0,2	0,2	13
Перу	66	14,0	2,0	3,1	47
Тайланд	69	28,0	0,9	1,3	35
Панама	73	22,2	1,7	2,4	23
Турция	67	20,7	1,7	2,1	48
Польша	70	20,0	0,3	0,6	14
Словакия	72	13,4	0,3	0,7	11
Венесуэла	71	29,3	2,3	3,0	23
ЮАР	64	18,6	2,2	2,4	50
Мексика	72	23,7	1,9	2,8	33
Мавритания	71	49,0	1,3	1,8	16
Бразилия	67	20,0	1,5	1,6	44
Тринидад	72	31,9	0,8	1,8	13
Малайзия	71	33,4	2,4	2,7	12
Чили	72	35,3	1,5	2,1	12
Уругвай	73	24,6	0,6	1,0	18
Аргентина	73	30,8	1,3	2,0	22
Греция	78	43,4	0,6	0,9	8
Испания	77	53,8	0,2	1,0	7
Нов.Зеландия	76	60,6	1,4	1,5	7
Ирландия	77	58,1	0,5	1,7	6
Израиль	77	61,1	3,5	3,5	8
Австралия	77	70,2	1,1	1,4	6
Италия	78	73,7	0,2	0,1	7
Канада	78	78,3	1,3	1,0	6
Финляндия	76	65,8	0,5	0,1	5
Гонконг	79	85,1	1,6	1,3	5

Швеция	79	68,7	0,6	0,3	4
Нидерланды	78	73,9	0,7	0,6	6
Бельгия	77	80,3	0,4	0,5	8
Франция	78	78,0	0,5	0,8	6
Сингапур	76	84,4	2,0	1,7	4
Австрия	77	78,8	0,8	0,5	6
США	77	100,0	1,0	1,1	8
Дания	75	78,7	0,3	0,1	6
Япония	80	82,0	0,3	0,6	4
Швейцария	78	95,9	1,0	0,8	6

Принятые в таблице обозначения:

Y - средняя ожидаемая продолжительность жизни при рождении, лет;

X_1 - ВВП в паритетах покупательной способности;

X_2 - темпы прироста населения по сравнению с предыдущим годом, %;

X_3 - темпы прироста рабочей силы по сравнению с предыдущим годом, %;

X_4 - коэффициент младенческой смертности, %.

10. Контрольные вопросы

1. Что называют корреляцией?
2. Для каких переменных можно применять расчет корреляций по Пирсону? Приведите формулу для расчета этого коэффициента.
3. Когда применяют ранговую корреляцию?
4. Дайте определение дихотомической переменной.
5. Какие диаграммы рассеяния можно построить в SPSS Statistics и как это сделать?
6. Объясните методику построения трехмерных диаграмм рассеяния в SPSS Statistics.
7. Приведите формулу для расчета частного корреляционного коэффициента и объясните, в каких случаях он используется.
8. Как в SPSS Statistics определить внутриклассовый коэффициент корреляции?

Лабораторная работа №6

Тема: Регрессионный анализ. Анализ временных рядов. Подбор тренда.

Цель: Сформировать практические навыки регрессионного анализа и обработки временных рядов.

1. Введение

Если расчёт корреляции характеризует силу связи между переменными, то регрессионный анализ служит для определения вида этой связи и дает возможность для прогнозирования значения одной (зависимой) переменной, отталкиваясь от значения другой независимой переменной.

Простая линейная регрессия является простейшей и применяется чаще всех остальных видов. Для проведения линейного регрессионного анализа зависимая переменная должна иметь интервальную (или порядковую) шкалу.

Временной ряд - это последовательность чисел; его элементы - это значения некоторого протекающего во времени процесса. Они измерены в последовательные моменты времени, обычно через равные промежутки. Как правило, составляющие временной ряд числа, - элементы временного ряда, нумеруют в соответствии с номером момента времени, к которому они относятся (например, x_1, x_2, x_3 и т.д.). Таким образом, порядок следования элементов временного ряда весьма существен.

Чаще всего значения временного ряда получаются непосредственной записью значений некоторого процесса через определенные промежутки времени. Например, если ежесуточно в определенное время записывать показания термометра, то получится временной ряд со значениями температуры в том месте, в котором находится термометр. Иногда значения элементов временного ряда получаются накапливанием некоторых данных за определенный промежуток времени (например, суммарное число посетителей магазина за день), усреднением (средняя температура за день) и т.д.

Целью прикладного статистического анализа временных рядов является построение математической модели ряда, с помощью которой можно объяснить поведение ряда и осуществить прогноз его дальнейшего поведения.

Анализ временного ряда обычно начинается с построения и изучения его графика. Затем обычно пробуют выделить – во временном ряде тренд, сезонные и периодические компоненты. После их исключения временной ряд

должен стать стационарным, т.е. таким, вероятностные свойства которого не изменяются с течением времени.

Тренд – это плавно изменяющаяся во времени компонента, которая описывает долговременные процессы.

Сезонная составляющая описывает поведение, изменяющееся в течение заданного периода (суток, месяца, года, десятка лет и т.д.). Например, пик пассажироперевозок на работу или дачи по дням, объем продаж подарков в декабре – это сезонные компоненты. Главная идея подхода к анализу сезонной компоненты заключается в переходе от сравнения всех значений временного ряда между собой - к сравнению значений через определенный промежуток времени.

Циклическая компонента описывает длительные периоды относительного подъема и спада. Она состоит из циклов, которые меняются по амплитуде и протяженности и представляют собой нечто среднее между сезонной компонентой и трендом.

Для оценки и удаления трендов из временных рядов чаще всего используется метод наименьших квадратов.

Значения временного ряда x_t рассматривают как отклик (зависимую переменную), а время t - как фактор, влияющий на отклик (независимую переменную):

$$x_{ti} = f(t_i, \theta) + \varepsilon \qquad \text{i = 1,2,...., n}$$

где f — функция тренда (она обычно предполагается гладкой),

θ — неизвестные нам параметры (параметры модели временного ряда),

ε_i — независимые и одинаково распределенные случайные величины, распределение которых мы предполагаем нормальным.

Метод наименьших квадратов состоит в том, что мы выбираем функцию тренда так, чтобы:

$$\sum [x_i - f(t_i, \theta)]^2 \to \min_{\theta}$$

2. Простая линейная регрессия

Линейная связь выражается прямой вида

$$y = b \cdot x + a$$

где b - регрессионный коэффициент,

a - смещение по оси ординат.

Смещение по оси ординат соответствует точке на оси y (вертикальной оси), где прямая регрессии пересекает эту ось. Коэффициент регрессии b указывает на угол наклона прямой.

При проведении простой линейной регрессии основной задачей

является определение параметров b и a. Оптимальным решением этой задачи является такая прямая, для которой сумма квадратов вертикальных расстояний до отдельных точек данных является минимальной (МНК).

Расчет линейного уравнения регрессии проведем на примере данных файла «сера_энергия».

Чтобы вызвать регрессионный анализ в SPSS Statistics, выберите в меню пункт *Анализ/Регрессия/Линейная*.

Появится диалоговое окно *Линейная регрессия*. Перенесите переменную «*энергия*» в поле для зависимых переменных и присвойте переменной «*сера*» статус независимой переменной. Ничего больше не меняя, начните расчет нажатием *ОК* (рис.6.1).

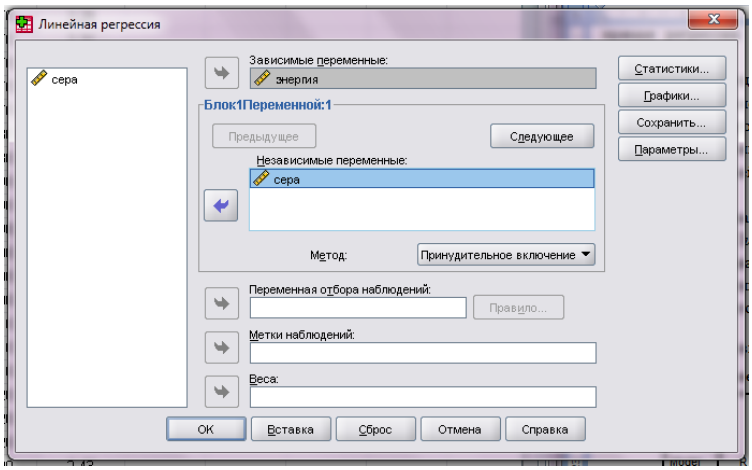


Рис.6.1. Диалоговое окно «Линейная регрессия»

Результат линейной регрессии появятся в окне *Вывод* (рис.6.2).

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,788[a]	,620	,615	,07510

a. Предикторы: (конст) сера

Дисперсионный анализ[b]

Модель		Сумма квадратов	ст.св.	Средний квадрат	Щ	Знч.
1	Регрессия	,682	1	,682	120,938	,000[a]
	Остаток	,417	74	,006		
	Всего	1,099	75			

a. Предикторы: (конст) сера

b. Зависимая переменная: энергия

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	2,309	,024		94,554	,000
	сера	14,411	1,310	,788	10,997	,000

a. Зависимая переменная: энергия

Рис.6.2. Результаты линейной регрессии

В третьей таблице выводятся коэффициент регрессии b и смещение по оси ординат a под именем «*константа*». То есть уравнение регрессии выглядит следующим образом:

энергия = 14,411 * сера + 2,309

В таблице *Сводка для модели* выведен коэффициент детерминации (R-квадрат). В нашем примере он равен 0,620. Эта величина характеризует качество регрессионной прямой, т.е. степень соответствия между регрессионной моделью и исходными данными. Мера определенности всегда лежит в диапазоне от 0 до 1. В этой же таблице под обозначением «*R*» показан корреляционный коэффициент Пирсона.

Принципиальный вопрос о том, может ли вообще имеющаяся связь между переменными рассматриваться как линейная, проще и нагляднее всего решать, глядя на соответствующую диаграмму рассеяния. Построение диаграмм рассеяния и нанесение на них регрессионной кривой рассмотрено в лабораторной работе №4.

3. Анализ временных рядов в пакете SPSS Statistics

Проведем анализ временного ряда для данных об урожайности зерновых культур с 1945 по 1989 гг. (в центнерах с гектара), представленных в файле «harvest».

Для построения графика временного ряда надо в пункте главного меню *Графика/Рассеяния/Точки/Простая диаграмма рассеяния*.

Присвоим значения переменных «*год*» и «*урожай*» осям X и Y соответственно. На полученном графике (рис.3) видно, что анализируемые данные содержат линейный тренд, т.е. тренд этого ряда может быть задан в виде прямой линии $tr_t = b_0 + b_1 * t$.

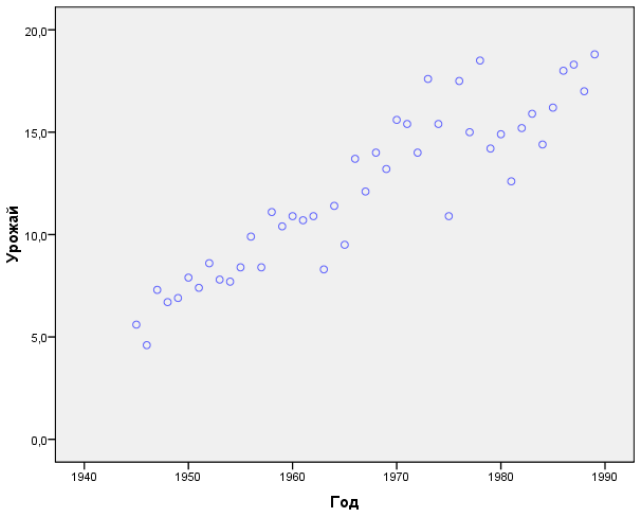


Рис.3. Диаграмма рассеяния временного ряда

В данной формуле в качестве независимой переменной фигурирует время t. Коэффициенты модели (b_0 и b_1) вычисляются с помощью метода наименьших квадратов по формулам:

$$b_0 = \bar{x} \quad (\text{где } \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t),$$

$$b_1 = \frac{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)\left(t - \frac{t(t+1)}{2}\right)}{\sum_{t=1}^{n}\left(t - \frac{t(t+1)}{2}\right)^2}$$

Для идентификации линейности тренда следует выбрать в пункте главного меню *Анализ/Регрессия/Подгонка кривой* (рис.6.4).

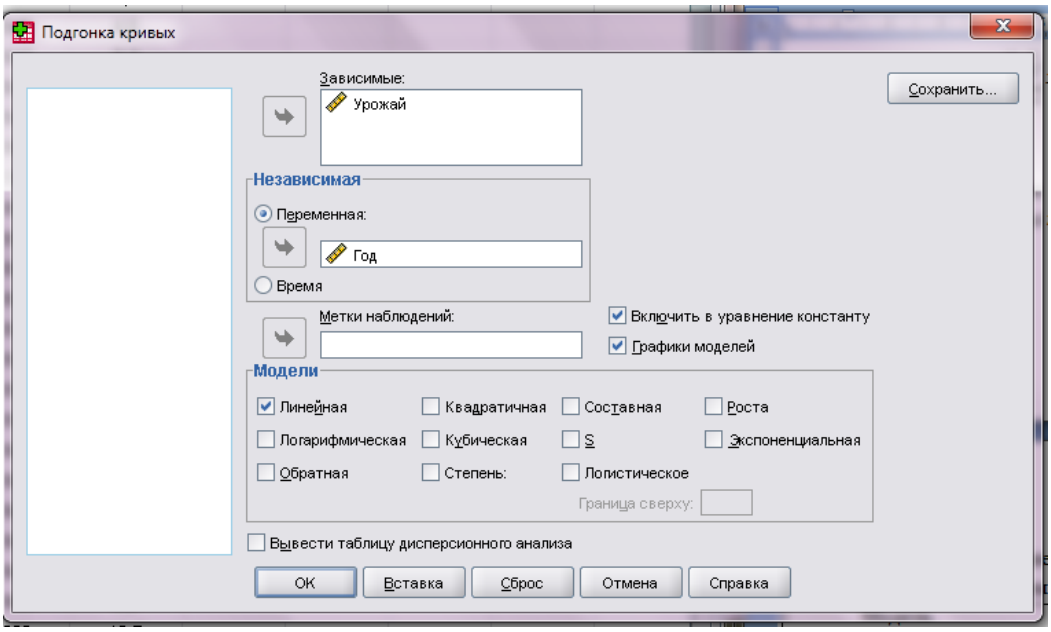


Рис.6.4. Диалоговое окно «Подгонка кривой»

В диалоговом окне *Подгонка кривой* можно выбрать одну из одиннадцати различных моделей, которым соответствуют формулы (табл.6.1).

В диалоговом окне *Подгонка кривой* надо:

- в поле *Зависимые* определить зависимую переменную (урожайность) и в поле *Независимая* - независимую (год);
- в области *Модели* выберите *линейную модель*;
- укажите *включение константы в уравнение*, установив флажок;
- нажмите *OK*.

Таблица 6.1

Модели кривых регрессионного анализа

Модель	Формула
Линейная	$y = b_0 + b_1 \cdot x$
Логарифмическая	$y = b_0 + b_1 \cdot \ln(x)$
Обратная	$y = b_0 + \frac{b_1}{x}$
Квадратичная	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2$
Кубическая	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$
Степенная	$y = b_0 \cdot x^{b_1}$
Показательная (составная)	$y = b_0 \cdot b_1^{\,x}$
S	$y = e^{b_0 + b_1 \cdot x}$
Логистическая	$y = \frac{1}{\frac{1}{u} + b_0 \cdot b_1^{\,x}}$
Роста	$y = e^{b_0 + b_1 \cdot x}$
Экспоненциальная	$y = b_0 \cdot e^{b_1 \cdot t}$

Замечание. При нажатии кнопки *Сохранить* происходит сохранение в виде отдельных переменных значения подобранной модели *Предсказанные значения*, *Остатки* и *Интервалы прогноза*, которые помещаются в электронную таблицу пакета.

Вывод результатов выполнения процедуры производится в окне *Вывод*. Таблицы содержат значения коэффициентов для выбранной модели тренда (рис.6.5).

Сводка для модели

R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
,915	,838	,834	1,609

Независимой переменной является Год.

Дисперсионный анализ

	Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
Регрессия	574,461	1	574,461	221,784	,000
Остаток	111,378	43	2,590		
Итого	685,839	44			

Независимой переменной является Год.

Коэффициенты

	Нестандартизованные коэффициенты		Стандартизованные коэффициенты		
	B	Стд. Ошибка	Бета	t	Знч.
Год	,275	,018	,915	14,892	,000
(Константа)	-528,950	36,338		-14,556	,000

Рис.6.5. Результаты выполнения процедуры «Подгонка кривой»

Результаты расчетов показывают, что линейная модель тренда объясняет примерно 84% (R-квадрат) общей вариации данных, а полученные оценки коэффициентов модели значимо отличаются от нуля (т.к. уровень значимости максимальный: Знч.<0,001). При этом коэффициент $b_0 = 5{,}289$ показывает среднюю урожайность зерновых в начальный (1945г.) момент времени рассматриваемого ряда. Значение коэффициента b_1 при переменной *Год* (при t=44 года) говорит о том, средний прирост урожайности за год равен примерно 0,275 ц/га.

Одновременно в окне просмотра появляется график ряда с подобранной кривой тренда (рис.6.6).

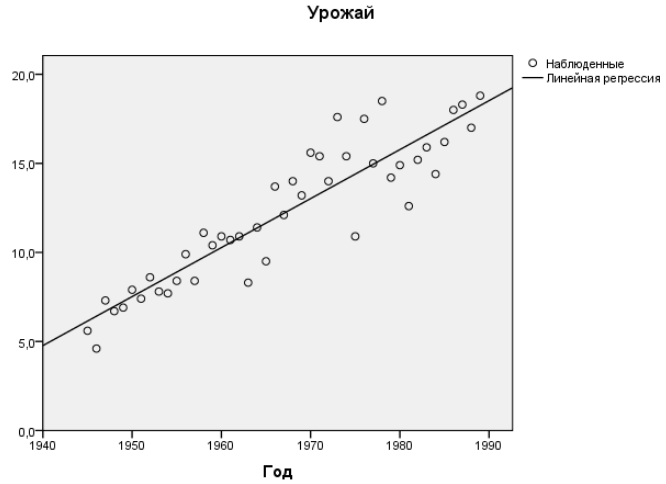


Рис.6.6. График регрессионного ряда с подобранной кривой тренда

4. Анализ остатков

Линия регрессии выражает наилучшее предсказание зависимой переменной (Y) по независимым переменным (X). Однако, природа редко (если вообще когда-нибудь) бывает полностью предсказуемой и обычно имеется существенный разброс наблюдаемых точек относительно подогнанной прямой. Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется **остатком.**

Поэтому дальнейший анализ модели связан с анализом остатков – выясняется, можно ли считать остатки некоррелированными и насколько их распределение согласуется с нормальным.

Замечание. Учитывая небольшую длину исследуемого ряда, вряд ли можно ожидать здесь высокой точности и достоверности результатов. Однако подобный анализ позволит понять, как далеко мы могли отклониться от условий применения метода наименьших квадратов для удаления тренда, и, тем самым, насколько можно верить полученным результатам.

Проверка коррелированности остатков. Одним из результатов работы процедуры *Подгонка кривых* является создание новой переменной EER_1, в

которой хранятся остатки подобранной модели. Для этого необходимо в диалоговом окне *Подгонка кривых/Сохранить* отметить функцию *Остатки*, затем *Продолжить* и *ОК*.

Далее, для выявления коррелированности остатков вычисляются оценки их автокорреляционной функции. Это можно сделать, вызвав меню *Анализ/Прогнозирование/Автокорреляции* (рис.6.7).

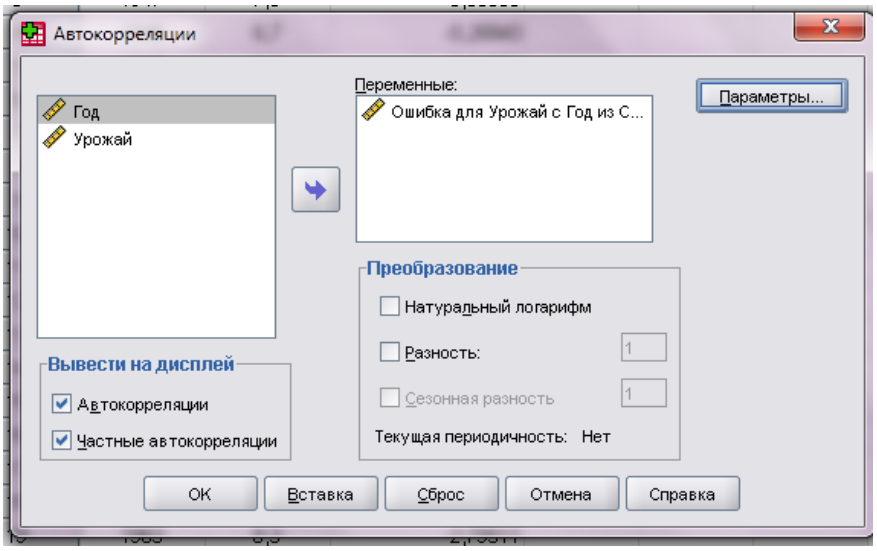


Рис.6.7. Диалоговое окно «Автокорреляции»

В окне *Автокорреляции* переменная ERR_1 со значениями остатков помещается в область *Переменные*. С помощью кнопки *Параметры* задайте *Максимальное число лагов* (шагов) равным 10, учитывая небольшую длину изучаемого ряда. Затем нажмите *Продолжить* и *ОК*. Программа выдает графики автокорреляционной функции (коррелограмму) (рис.6.8).

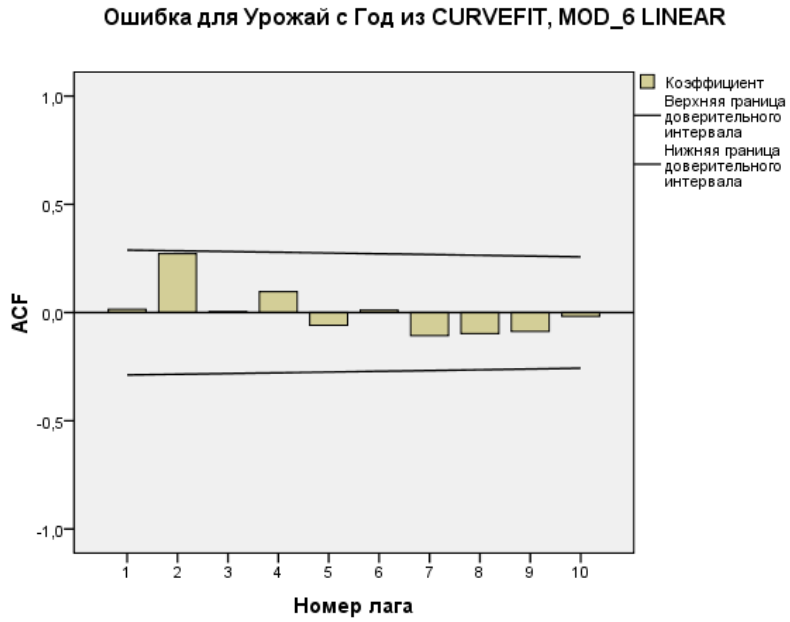


Рис.6.8. График автокорреляционной функции остатков для ряда урожайности зерновых

На основании графика можно говорить о некоррелированности остатков.

Остатки не коррелированны, если оценки лежат внутри доверительного интервала для нулевых значений автокорреляционной функции.

5. Общие задания

1. Получите задания у преподавателя.
2. Вычислите оценки параметров простой линейной регрессии.
3. Получите графики уравнения регрессии.
4. Для вашего временного ряда постройте график и подберите модель тренда.
5. Проведите анализ остатков.
6. Проанализируйте полученные результаты.

6. Требования к отчету

Отчет по лабораторной работе №6 предоставляется в письменном виде и должен содержать:

1. Ответы на контрольные вопросы № 1 – 6.
2. Анализ результатов, полученных после выполнения практических заданий.
3. Исходные данные и результаты выполнения индивидуального практического задания.
4. Выводы по полученной информации.

7. Индивидуальные задания

1÷2. Экспорт Украины за 1976 – 2004 гг. характеризуются данными, представленными в табл.

Год	Россия, млн руб./грв.
	Экспорт
1976	44
1977	47
1978	51
1979	56
1980	62
1981	67
1982	72
1983	79
1984	95
1985	117
1986	129
1987	146
1988	166

1989	204
1990	209
1991	236
1992	257
1993	281
1994	328
1995	366
1996	405
1997	431
1998	450
1999	498
2000	549
2001	523
2002	527
2003	590
2004	669

3÷4. Импорт Украины за 1976 – 2004 гг. характеризуются данными, представленными в табл.

Год	Импорт
1976	43
1977	46
1978	51
1979	56
1980	63
1981	71
1982	74
1983	80
1984	91
1985	131
1986	126
1987	144
1988	164
1989	206
1990	205
1991	247
1992	278
1993	280
1994	332
1995	386
1996	419
1997	412
1998	434
1999	496
2000	547
2001	510
2002	520
2003	584
2004	661

5÷6. Внешнеторговый оборот Украины за 1976 – 2004 гг. характеризуются данными, представленными в табл.

Год	Внешнеторговый оборот
1976	87
1977	93
1978	102
1979	112
1980	125
1981	138
1982	146
1983	159
1984	186
1985	248
1986	255
1987	290
1988	330
1989	410
1990	414
1991	483
1992	535
1993	561
1994	660
1995	752
1996	824
1997	843
1998	884
1999	994
2000	1096
2001	1033
2002	1047
2003	1174
2004	1330

7÷8. Экспорт России за 1976 – 2004 гг. характеризуются данными, представленными в табл.

Год	Экспорт
1976	202
1977	219
1978	239
1979	278
1980	306
1981	328
1982	352
1983	402
1984	483
1985	562
1986	609
1987	683
1988	846

1989	1116
1990	1065
1991	1266
1992	1474
1993	1540
1994	1798
1995	2026
1996	2286
1997	2640
1998	2924
1999	3337
2000	3479
2001	3367
2002	3477
2003	3900
2004	4498

9÷10. Импорт России за 1976 – 2004 гг. характеризуются данными, представленными в табл.

Год	Импорт
1976	195
1977	209
1978	221
1979	248
1980	283
1981	305
1982	337
1983	351
1984	400
1985	474
1986	533
1987	581
1988	633
1989	811
1990	1109
1991	1061
1992	1261
1993	1499
1994	1570
1995	1866
1996	2125
1997	2357
1998	2694
1999	2864
2000	3277
2001	3379
2002	3187
2003	3334
2004	3719

11÷12. Внешнеторговый оборот России за 19761 – 2004 гг. характеризуются данными, представленными в табл

Год	Внешнеторговый оборот
1976	395
1977	411
1978	440
1979	487
1980	561
1981	611
1982	665
1983	703
1984	802
1985	957
1986	1095
1987	1190
1988	1316
1989	1657
1990	2225
1991	2126
1992	2527
1993	2973
1994	3110
1995	3664
1996	4151
1997	4643
1998	5334
1999	5788
2000	6614
2001	6858
2002	6554
2003	6811
2004	7619

13÷14. Имеются поквартальные данные по розничному товарообороту России в 1995-1999 гг.

Номер квартала	Товарооборот, % к предыдущему периоду
1	100
2	93,9
3	96,5
4	101,8
5	107,8
6	96,3
7	95,7
8	98,2
9	104
10	99
11	98,8

12	101,9
13	113,1
14	98,4
15	97,3
16	102,1
17	97,6
18	83,7
19	84,3
20	88,4

15÷16. Имеются данные об объеме экспорта из Российской Федерации (млрд долл., цены Фондовой Общероссийской биржи (ФОБ)) за 1994-1999 гг.

Номер квартала	Экспорт, млрд долл., цены ФОБ
1	4087
2	4737
3	5768
4	6005
5	5639
6	6745
7	6311
8	7107
9	5741
10	7087
11	7310
12	8600
13	6975
14	6891
15	7527
16	7971
17	5875
18	6140
19	6248
20	6041
21	4626
22	6501
23	6284
24	6707

17÷18. Приводятся сведения об уровне среднегодовых цен на какао-бобы из Бразилии, амер. центы за фунт.

Год	Цена	Год	Цена	Год	Цена	Год	Цена
1970	9.40	1977	39.40	1984	49.40	1991	59.40
1971	3.50	1978	33.50	1985	43.50	1992	53.50
1972	6.20	1979	36.20	1986	46.20	1993	56.20
1973	28.50	1980	58.50	1987	68.50	1994	78.50
1974	53.40	1981	83.40	1988	93.40	1995	103.40
1975	36.60	1982	66.60	1989	76.60	1996	86.60
1976	57.00	1983	87.00	1990	97.00	1997	107.00

19÷20. Приводятся сведения об уровне среднегодовых цен на рис из Таиланда на рынках Бангкока, амер. доллары за метрическую тонну.

Год	Цена	Год	Цена	Год	Цена	Год	Цена
1970	123.00	1977	302.00	1984	312.00	1991	322.00
1971	110.00	1978	399.00	1985	409.00	1992	419.00
1972	130.00	1979	364.00	1986	374.00	1993	384.00
1973	276.00	1980	464.00	1987	474.00	1994	484.00
1974	522.00	1981	513.00	1988	523.00	1995	533.00
1975	343.00	1982	323.00	1989	333.00	1996	343.00
1976	234.00	1983	307.00	1990	317.00	1997	327.00

8. Контрольные вопросы

1. Для чего применяется регрессионный анализ?
2. Что такое временной ряд?
3. Как получают временные ряды? Приведите примеры.
4. Какова цель статистического анализа временных рядов?
5. С чего начинается анализ временных рядов?
6. Опишите метод наименьших квадратов.
7. Какие возможности предоставляет процедура Линейная регрессия?
8. Для чего предназначена процедура Подгонка кривой? Опишите ее работу.
9. Как проводится проверка коррелированности остатков?

Учебное издание

Информационные системы и технологии

Методические рекомендации по курсовому проектированию

П.л.5. Тираж 25 экз.