Исторический факультет
Московского государственого университета
им. М.В. Ломоносова

КОМПЬЮТЕРИЗОВАННЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ДЛЯ ИСТОРИКОВ

Учебное пособие

Под редакцией
Л.И. Бородкина, И.М. Гарсковой

МОСКВА
1999–2009

ББК 63.3 (2) 521
В 75

Авторы:
Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова, Т.Ф. Изместьева,
В.В. Лазарев, А.И. Тихонов

Компьютеризованный статистический анализ для историков / Под ред. Л.И. Бородкина и И.М. Гарсковой. – М.: 1999. – 187 с.: илл.

Пособие представляет собой методический материал по второй части курса "Информатика и математика", который читается студентам исторического факультета Московского государственного университета им. М.В. Ломоносова. В пособии дается изложение основных понятий и методов математической статистики и анализа данных, адаптированное с учетом опыта применения этих методов в исторических исследованиях. Большое внимание уделяется компьютерной реализации этих методов. Изложение иллюстрировано многочисленными примерами, основанными на материалах исторических источников и результатах исследований историков-квантификаторов. Каждая глава пособия завершается списком контрольных вопросов и заданий. Самостоятельную ценность представляет обширное приложение, содержащее около 40 таблиц статистических данных, которые могут использоваться для самостоятельной работы студентов и выполнения контрольных заданий

Пособие ориентировано не только на студентов-историков, но и на других читателей, интересующихся применением статистических методов в исторических исследованиях.

Рецензенты:
д.и.н., проф. С.Г. Кащенко,
д.т.н., проф. Е.В Бауман

ISBN 5-204-00125-5

© Коллектив авторов, 1999
© Исторический факультет МГУ, 1999

ПРЕДИСЛОВИЕ

Преподавание статистических методов на исторических факультетах имеет уже достаточно длительную традицию. Так, студентам-историкам Московского университета эта дисциплина преподается (по инициативе И.Д. Ковальченко) уже около 30 лет. С начала 1980-х годов занятия со студентами-историками по курсу "Основы математической статистики" включали элементы практикума на ЭВМ, когда с помощью удаленных терминалов, расположенных на историческом факультете, удавалось использовать возможности парка "больших" машин университетского вычислительного центра для обучения студентов статистическим методам [1]. Ситуация начала резко меняться в лучшую сторону на рубеже 80-х – 90-х годов, в связи с приходом "микрокомпьютерной волны". Появление на историческом факультете компьютерных классов создало новые возможности. Их реализации способствовало и то обстоятельство, что разработанные в середине 90-х гг. новые образовательные стандарты включали и цикл "Информатика и математика", обязательный для студентов исторических специальностей в университетах РФ.

Сейчас цикл дисциплин "Информатика и математика" на историческом факультете МГУ преподается студентам непосредственно после курса "Количественные методы в исторических исследованиях", читаемого в III семестре. Цикл "Информатика и математика" рассчитан на два семестра (программа цикла включена в homepage исторического факультета МГУ – http://www.hist.msu.ru). Лекции сопровождаются практическими занятиями; оба практикума были поставлены в 1991 г. и непрерывно совершенствуются на протяжении 90-х гг. [2]

---

[1] Представление о методических находках и проблемах компьютеризации обучения историков МГУ того времени дают выпущенные нами в тот период методические разработки: Бородкин Л.И., Васенин В.Г., Гарскова И.М., Изместьева Т.Ф. Использование вычислительной техники в учебном процессе на историческом факультете. Методическая разработка по курсу "Основы математической статистики". М., МГУ, 1985. – 72 С.; Они же. Компьютер в историческом исследовании. Учебно-методическая разработка. М., МГУ, 1986.

[2] Так, в 1996 г. в серии "10 новых учебников по историческим дисциплинам" вышло в свет учебное пособие (Историческая информатика / Отв. ред. Л.И. Бородкин и И.М. Гарскова. М., 1996. – 400 С.), в основном ориентированное на первую часть цикла; вопросы применения статистических методов анализа данных в этом пособии рассматривались в одной главе, что было явно недостаточно.

Курс математики является второй частью цикла "Информатика и математика". Этот курс специально ориентирован преимущественно на использование статистических методов и методов анализа данных в работе историка; здесь также затрагиваются методические вопросы математического моделирования исторических процессов и явлений. Курс включает лекции и практические занятия и завершается экзаменом.

В центре внимания практикума по данному курсу находятся методы математической статистики и анализа данных. Несколько особняком стоит последний раздел, посвященный математическому моделированию в социальных науках, и, в частности, в исторических исследованиях. В этой части курса студенты знакомятся с аналитическими, статистическими и имитационными моделями, обсуждают проблемы их адекватности, верификации, оценки параметров, а также корректности полученных на их основе содержательных результатов.

Курс ориентирован на учет специфики гуманитарного образования. Так, при изложении математических понятий и методов основное внимание уделяется их логической структуре. Обсуждение принципов построения математических моделей и статистических теорий проводится с учетом ограничений, порождаемых особенностями социально-гуманитарного знания.

Курс математики (с акцентом на методы статистики) для историков имеет свою специфику также и в том, что иллюстративный, учебный материал и тестовые задания базируются на фрагментах реальных источников. Кроме того, преподаватели и в постановке задач обычно выделяют наиболее характерные типы исследовательских проблем, с которыми сталкиваются историки (например, построение типологии или изучение динамики и т.д.).

Далее, помимо стандартных методов, которые есть практически в любом пакете статистических программ, в данном курсе имеется и компонента, методически ориентированная на особенности задач, возникающих в социально-гуманитарных исследованиях. Здесь можно назвать задачи многомерной нечеткой классификации (для решения которых используется оригинальное программное обеспечение, разработанное в Лаборатории исторической информатики) или задачи моделирования динамики исторических процессов с использованием методов теории самоорганизующихся систем (пока эта проблематика изучается на демонстрационном уровне).

Вновь подчеркнем, что методика преподавания курса особое внимание уделяет не чисто математическим аспектам, а скорее вопросам логики и корректности применения тех или иных методов, т.е. умению выбирать ме-

тодически верные способы решения конкретных исследовательских проблем, работая со стандартными пакетами статистических программ. Практические занятия ведут сотрудники Лаборатории исторической информатики в компьютерных классах исторического факультета, оснащенных современными компьютерами, соединенными в локальную сеть. Практические задания студенты выполняют, работая преимущественно в пакете STATISTICA.

Методическое пособие, предлагаемое вниманию читателя, состоит из введения, восьми глав (4 частей) и приложения.

Введение написано Л.И. Бородкиным и И.М. Гарсковой, главы 1, 4 и 5 – Е.Б. Беловой, И.М. Гарсковой и В.В. Лазаревым, главы 2 и 3 – И.М. Гарсковой, главы 6 и 7 – Л.И. Бородкиным и И.М. Гарсковой, глава 8 – Т.Ф. Изместьевой. Материалы для приложения подготовлены при участии А.И. Тихонова.

ВВЕДЕНИЕ

Применение компьютерных методов и технологий для статистической обработки массовых исторических источников имеет уже 30-летнюю традицию.

Появление в последние годы целого ряда новых (или усовершенствованных) статистических пакетов с большим набором методов и удобным интерфейсом может породить у историка иллюзию простоты использования даже сложных методов статистического анализа. В этой связи следует отметить, что статистические методы и подходы в изучении исторических источников требуют некоторых комментариев относительно пределов их корректного и эффективного использования [1].

Прежде всего, при работе с этими методами используются как равноправные и термин "*статистические методы*", и термин "*методы анализа данных*". При том, что "технологически" математическая статистика и анализ данных практически не различаются, следует все же указать, что оба эти подхода к анализу базируются на различных моделях получения данных и, соответственно, определяют различие в подходах к интерпретации полученных результатов.

В основе *математико-статистического* подхода – *вероятностная* модель, предполагающая, что имеющаяся статистическая совокупность представляет собой **выборку** из некоторой реальной или гипотетической **генеральной совокупности**, на которую и должны распространяться полученные выводы. Эта модель довольно хорошо соответствует данным, действительно представляющим собой выборочные совокупности. Например, бюджеты крестьянских хозяйств определенного селения представляют собой лишь часть более обширной совокупности бюджетов хозяйств данного региона, а личные карточки рабочих некоторого предприятия – часть личных карточек рабочих некоторой отрасли и т.п. Изучая такие совокупности документов, исследователь действительно стремится расширить свои выводы на весь регион или отрасль промышленности, и в этом случае оправдан математико-статистический подход.

Иная модель лежит в основе *анализа данных*: не предполагается, что изучаемая информация получена из более обширной генеральной совокупности, и полученные выводы интерпретируются без какого-либо расшири-

[1] Это представляется тем более целесообразным, что учебник "Количественные методы в исторических исследованиях", вышедший под ред. И.Д. Ковальченко в 1984 г., давно уже стал библиографической редкостью.

тельного толкования. Например, если изучается социально-профессиональная структура городского населения (скажем, для построения типологии городов в этом аспекте) и исследователь располагает соответствующими данными по всем городам, нет смысла расширять полученные результаты типологии на генеральную совокупность (ее просто нет, вернее, она совпадает с изучаемой, хотя и здесь сторонники вероятностного подхода утверждают, что имеющаяся совокупность – лишь один из возможных случайных результатов реализации некоторого исторического процесса). Не вдаваясь далее в существо этого вопроса, еще раз подчеркнем, что два упомянутых подхода используют один и тот же арсенал методов и различаются лишь на этапах постановки задачи и интерпретации результатов.

Не следует думать, что статистические методы пригодны лишь для анализа статистических источников, представляющих собой в исходном виде цифровой материал – статистические методы годятся и для работы с неколичественной по природе информацией. Откуда же берутся цифры в этом случае? Здесь мы подходим к основной характерной особенности статистических методов: они не имеют дело с отдельными случаями, объектами, индивидуумами – но всегда с совокупностями, группами, т.е. *массовым материалом*. Там и тогда, где и когда речь идет о **совокупности данных,** возможен статистический подсчет и, следовательно, применение статистических методов. Итак, мы имеем дело с совокупностью объектов, которые обладают некоторым набором признаков (показателей, характеристик). Одни показатели (они называются количественными) могут быть измерены для каждого объекта числом, для других это невозможно. Те же показатели, которые не могут быть измерены количественно, т.е. выражены числом для одного объекта (профессия человека, отраслевая принадлежность предприятия и т.п.), также довольно просто связать с количеством (частотой) встречаемости соответствующих значений в рамках определенной совокупности.

Теперь мы вплотную подошли к проблеме измерения показателей. Как уже стало ясно из предыдущего абзаца, *измерить – это значит связать с некоторым числом*. В связи с возможностью измерения все признаки принято делить на две большие группы: **количественные и качественные**. С количественными показателями допустимы все арифметические операции, для них разработано большинство статистических методов. Качественные признаки измеряются по *номинальной шкале,* что эквивалентно отнесению каждого объекта к одной из категорий по данному признаку. Подсчет количества или доли объектов, попадающих в ту или иную категорию данного признака, связывает с каждой из них число, т.е. на уровне совокупности происходит как бы превращение качества в количество: для группы людей число рабочих среди них – количественный показатель, с которым уже

можно совершать арифметические операции, точно так же, как, например, со средним возрастом в данной совокупности.

Номинальные признаки у отдельных объектов, однако, не могут быть измерены числом и поэтому не могут участвовать не только в арифметических операциях (ибо что есть средняя профессия двух индивидуумов?), но даже и в операциях сравнения: две разные профессии нельзя сравнить по принципу "больше-меньше". Несколько больше возможностей для таких сравнений предоставляют *ранговые* признаки (или признаки, измеряемые по *шкале порядка*); категории этих признаков упорядочены в виде некоей "табели о рангах", так что объекты, попадающие в разные категории, сравнимы между собой по принципу "лучше-хуже" (подобно категориям признака "образование" или всем понятным баллам экзаменационных оценок). Однако сравнение категорий (значений) качественного признака не позволяет выразить различие каким-либо реальным числом (ведь если один студент сдал экзамен на "двойку", а другой на "четверку", из этого не следует, что разница "2" означает, будто первый знает в два раза меньше второго или прочитал на две книги больше и т.п.). Тем не менее, нередко именно числами, величины которых соответствуют порядку рангов, обозначают категории ранговых признаков, создавая обманчивое впечатление, что это количественные показатели, но это не числа в прямом смысле слова, а именно коды, условные обозначения. Точно также числами обозначают иногда и категории номинальных признаков, но в этом случае даже сравнение их по величине ни о чем не говорит. Реальными числами для качественных показателей, как уже было сказано, являются количества или доли объектов, попадающих в отдельные категории, в данной совокупности. Для качественных признаков в статистике разработаны методы, основанные на этом способе измерения, и возможности работы с ними, разумеется, не ограничены подсчетом частот.

Учет этих "азбучных истин" статистики позволит, как мы полагаем, повысить культуру компьютерного анализа статистических данных в работе историка, осваивающего методы обработки массовых источников.

Переходя непосредственно к методам анализа статистических данных, введем несколько определений и обозначений. Обозначим число этих объектов или объем нашей совокупности n, тогда каждый признак X в этой совокупности принимает n значений: $x_1, x_2, \ldots, x_n$. В простейшем случае на объектах задано значение только одного признака, в этом случае совокупность называется одномерной, в противном случае мы имеем дело с многомерными данными.

ОСНОВНЫЕ МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Классическим для математической статистики подходом, как было сказано во введении к данной главе, является представление исходных данных как выборки из реальной или гипотетической генеральной совокупности. При этом все результаты интерпретируются как выборочные, и ставится задача их оценки в генеральной совокупности. Основные методы математической статистики можно отнести к двум ее разделам: **теории статистического оценивания параметров** и **теории проверки статистических гипотез.**

Основные понятия теории статистического оценивания

Идея статистического оценивания параметров генеральной совокупности по выборочным данным сводится к тому, что выборочная характеристика какого-либо параметра (например, среднего арифметического значения признака) является не точным, а приближенным значением – **оценкой** – этого же параметра в генеральной совокупности. Возникает вопрос: как сильно отклоняется эта оценка от истинного значения? В частности, нельзя ли указать такую величину *ошибки*, которая "практически достоверно" (т.е. с вероятностью, близкой к единице) гарантировала бы, что выборочная оценка не отличается от неизвестного значения более чем на величину этой ошибки? Или – что то же самое – нельзя ли указать вокруг выборочного значения параметра такой интервал, который бы с заданной (достаточно высокой) вероятностью – **доверительной вероятностью** – "накрывал" бы истинное значение этого параметра? Этот интервал в математической статистике называется **доверительным интервалом**; его величина зависит как от доверительной вероятности (т.е. надежности оценивания), так и от объема выборки.

Основные понятия теории статистической проверки гипотез

На разных стадиях статистического исследования возникает необходимость в формулировке и проверке некоторых предположений – **гипотез** – относительно природы или величины неизвестных параметров (например, предположения об отсутствии взаимосвязи между двумя признаками). Цель статистической проверки высказанной (*нулевой*) гипотезы состоит в выявлении того, противоречит или нет эта гипотеза имеющимся статистическим (выборочным) данным. Процедура сопоставления высказанной гипотезы с реальными выборочными данными проводится на основе того или иного **статистического критерия.** Статистический критерий представляет собой совокупность правил вычисления некоторой статистической характеристи-

ки гипотезы и проверки ее величины. Результат проверки может быть либо *отрицательным* (данные противоречат высказанной гипотезе), и тогда гипотеза отклоняется, либо *неотрицательным*, и гипотеза не отклоняется. Статистическая проверка и ее вывод носят вероятностный характер, т.е. вывод делается всегда с определенной степенью вероятности (достаточно большой), а шанс отклонить верную гипотезу или, наоборот, не отклонить неверную – не равен нулю (хотя предполагается достаточно малым). В теории статистической проверки гипотез очень важно, что можно оценить вероятность совершить ошибку и, таким образом, получить представление о надежности выводов. Вероятность ошибочного отклонения нулевой гипотезы принято называть **уровнем значимости;** эта величина обычно выбирается из некоторого стандартного набора (0,1; 0,05; 0,001 и др.) Особенно распространенной является величина уровня значимости, равная 0,05; это означает, что в среднем в пяти случаях из 100 мы можем ошибочно отвергнуть высказанную гипотезу на основании данного статистического критерия. Каждому уровню значимости соответствует **критическое значение** статистической характеристики, которое делит все множество значений этой характеристики на две области: **допустимых значений** и **критическую** (область значений статистической характеристики, вероятность появления которых меньше выбранного уровня значимости). Таким образом, критическая область содержит именно те значения статистической характеристики, которые мы считаем практически невозможными.

К основным типам гипотез, проверяемых в ходе статистической обработки данных, относятся:

- гипотезы *о типе закона распределения* признака или критерии согласия (чаще всего проверяется соответствие нормальному закону распределения);
- гипотезы *о числовых значениях параметров* совокупности (например, о нулевом значении коэффициента корреляции);
- гипотезы *о типе зависимости* признаков (например, о линейной зависимости) [1].

[1] Наиболее важным вопросом для исторического исследования, изучающего закономерности сложных явлений, является установление взаимосвязей. При этом существенно не установить наличие связи там, где ее на самом деле нет. Поэтому в историческом исследовании обычно проверяют гипотезы об отсутствии взаимосвязей. Однако часто историку приходится иметь дело не с выборкой, а с самой генеральной совокупностью – в этом случае параметры, вычисленные по статистическим данным, казалось бы, не требуют применения теории оценивания или теории проверки гипотез. Однако для задач установления связей или законов распределения проверка гипотез все же имеет смысл, т.к. выявляемые закономерности могут (особенно в малых по объему выборках) искажаться и затемняться действием слу-

ПАКЕТ STATISTICA

Несложные статистические методы можно, конечно, применять и "вручную". Однако в наше время, как правило, используются пакеты прикладных статистических программ, широко доступные пользователям персональных компьютеров и содержащие широкий набор методов, включая наиболее "продвинутые". Основной задачей данной главы является прежде всего пояснение тех методов, которые предлагают пользователям эти пакеты. Без правильного понимания методического аппарата невозможны ни правильный выбор соответствующих методов, ни корректная интерпретация массы результатов, которые пользователь получает при работе с каждым из этих методов.

Иллюстрация методов работы в этой главе будет ориентирована на статистический пакет STATISTICA для Windows. Этот программный продукт фирмы StatSoft полностью совместим со всеми возможностями оболочки Windows и по своему дизайну прекрасно соответствует системе Microsoft Office, отдельные элементы которой уже рассматривались в предыдущих главах. Особенно много сходства у пакета STATISTICA с табличным процессором Excel. Это не удивительно, поскольку именно "идеология" электронных таблиц положена в основу модуля организации данных (**Data Management**) в пакете STATISTICA. Возможности ввода, редактирования, кодировки, сортировки и т.п., которые так прекрасно выполняют табличные процессоры, наряду с богатейшим выбором типов графического представления данных – все это снимает обычные (и справедливые – что касается таких известнейших пакетов, как Statgraphics или SPSS в версиях для DOS) упреки в адрес статистических пакетов, уделяющих недостаточное внимание подготовке, организации и визуализации данных.

Перечислим коротко, что к числу возможностей организации данных в пакете STATISTICA относятся:

- ввод данных непосредственно в таблицу;
- экспорт данных из таких пакетов, как Lotus/Quattro, Excel, SPSS, dBASE, чтение обычных ASCII-файлов;
- добавление, удаление, перемещение, копирование и переименование строк и столбцов таблицы (объектов и признаков);
- создание новых признаков на основе исходных (подсчет процентов и долей, относительных и суммарных показателей и т.п.).

чайных причин. При этом гипотеза применяется не для распространения полученных выводов на некую более обширную генеральную совокупность, а для того, чтобы установить, насколько закономерными или же случайными являются полученные выводы для имеющихся в данной совокупности условий.

ЧАСТЬ I

СТАТИСТИЧЕСКОЕ ОПИСАНИЕ

ГЛАВА 1

ДЕСКРИПТИВНАЯ СТАТИСТИКА

Для более глубокого исследования материала необходимы обобщающие количественные показатели, раскрывающие общие свойства статистической совокупности. Эти показатели, во-первых, дают общую картину, показывают тенденцию развития процесса или явления, нивелируя случайные индивидуальные отклонения, во-вторых, позволяют сравнивать различные совокупности и, наконец, используются во всех разделах математической статистики при более полном и сложном анализе статистического материала. Статистические характеристики описывают параметры т.н. эмпирического распределения признака.

1.1. ОСНОВНЫЕ СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ

Эти параметры можно разделить на две основные группы: меры среднего уровня и меры рассеяния (разброса).

1.1.1. Меры среднего уровня

К ним относятся:

- *среднее* (арифметическое) значение – сумма всех значений, отнесенная к общему числу наблюдений (принятые обозначения: Mean или $\overline{X}$), т.е. средним арифметическим значением признака X называется величина

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

где x_i – значение признака у *i*-го объекта, *n* – число объектов в совокупности.

- *минимум* – минимальное значение переменной (Min)
- *максимум* – максимальное значение переменной (Max)

- *мода* – наиболее часто встречающееся значение переменной (M)
- *медиана* – среднее по порядку значение (принятые обозначения: Median, m). Медиана – это "срединное" значение признака в том смысле, что у половины объектов совокупности значения этого признака меньше, а у другой половины – больше медианы. Вычислить медиану можно таким образом: упорядочить все значения признака по возрастанию (убыванию) и найти число в этом вариационном ряду, которое либо имеет номер $(n+1)/2$ – в случае нечетного n, либо находится посередине между числами с номерами $n/2$ и $(n+2)/2$ – в случае четного n [1].

Не все из перечисленных характеристик можно вычислять для качественных признаков. Если признак качественный и номинальный, то для него можно найти только моду (ее значением будет название наиболее часто встречающейся категории номинального признака). Если признак ранговый, то кроме моды для него можно найти еще и медиану, а также минимум и максимум. Однако среднее арифметическое значение можно вычислять только для количественных признаков.

В случае количественных данных все характеристики среднего уровня, очевидно, измеряются в тех же единицах, что и сам исходный признак. Если все значения исходного признака изменятся в несколько раз или на некоторое число, то же самое произойдет и со всеми средними величинами для этого признака.

1.1.2. Меры рассеяния

К ним относятся:

- *среднее квадратическое или стандартное отклонение* – мера разброса значений признака около среднего арифметического значения (принятые обозначения: Std.Dev. (*standard deviation*), σ или s). Величина этого отклонения вычисляется по формуле

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}\ .$$

- *дисперсия признака* (σ^2 или s^2)

[1] Упомянем также квартили, разбивающие ранжированный ряд значений признака на 4 части по 25% значений в каждой. Квартили при этом называются нижней, средней и верхней (при этом, очевидно, средняя квартиль совпадает с медианой). Аналогично можно ввести децили, разбивающие вариационный ряд значений на группы по 10% чисел и другие квантили - числа, разбивающие упорядоченную совокупность значений признака на равные по объему части.

• *коэффициент вариации* – отношение стандартного отклонения к среднему арифметическому, выраженное в процентах (обозначается в статистике буквой *V*). Коэффициент вычисляется по формуле: $V = \frac{s}{\bar{x}} \times 100\%$.

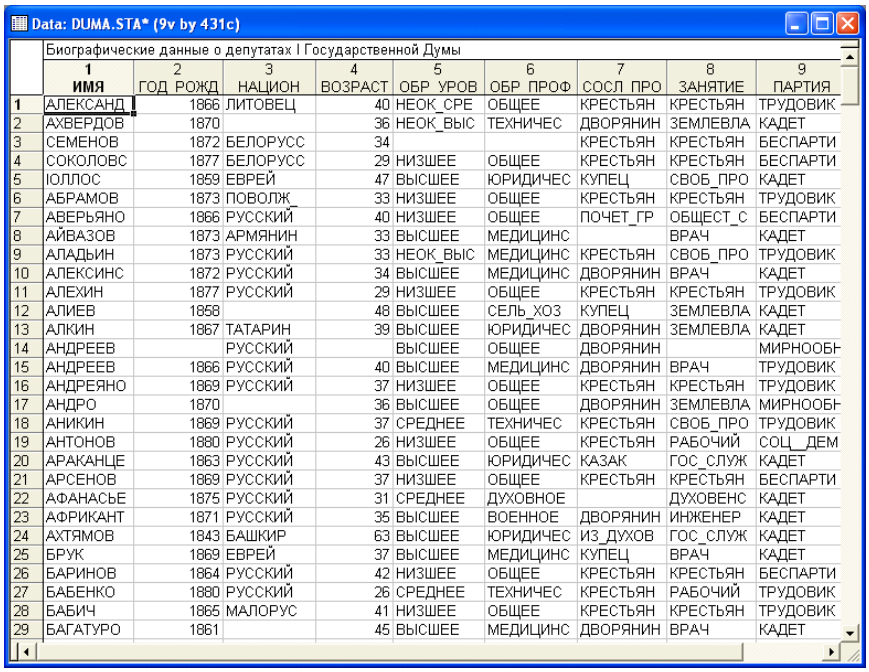
Прежде всего отметим, что *все* меры разброса можно вычислять только для количественных признаков. Все они показывают, насколько сильно варьируют значения признака (а точнее – их отклонения от среднего) в данной совокупности. Чем меньше значение меры разброса, тем ближе значения признака у всех объектов к своему среднему значению, а значит, и друг к другу. Если величина меры разброса равна нулю, значения признака у всех объектов одинаковы.

Наиболее часто используется среднее квадратическое (или стандартное) отклонение *s*. Оно измеряется, как и среднее арифметическое, в тех же единицах, что и сам исходный признак. Заметьте, что при изменении всех значений признака в несколько раз, точно так же изменится и стандартное отклонение, однако если все значения признака увеличить (уменьшить) на некоторую величину, его стандартное отклонение *не изменится*. Наряду со стандартным отклонением часто пользуются дисперсией, равной его квадрату, однако на практике она является менее удобной мерой, поскольку единицы измерения дисперсии не соответствуют единицам измерения признака (попробуйте представить рубли или тонны в квадрате!).

Смысл коэффициента вариации состоит в том, что он, в отличие от *s*, измеряет не абсолютную, а относительную меру разброса значений признака в статистической совокупности. Дело в том, что сравнение распределения отдельных признаков на основании обобщенных характеристик, таких, как среднее арифметическое значение и стандартное отклонение, затруднительно во многих случаях, например, когда эти признаки измеряются в разных единицах. Но даже если признаки и имеют одинаковый смысл, прямое сравнение возможно лишь для средних арифметических значений, но не для стандартных отклонений. Например, в одной группе среднее квадратическое отклонение по доходу равно 400 руб., а во второй – 2000 руб., то есть в 5 раз больше, чем в первой. Можно ли сделать вывод, что первая группа гораздо более однородна по величине дохода, чем вторая, – или следует обратить внимание на то, что и средние значения показателей неодинаковы? Если учесть, что средний доход в первой группе – 800 руб., а во второй – 8000 руб., то получим, что в первой группе V = 50%, а во второй V = 25%, т.е. в относительном измерении как раз вторая группа значительно более однородна .

Пример 1.1. Рассмотрим таблицу, созданную на основе базы данных по депутатам 1-й Государственной думы 1906 г. (файл Duma.sta). Эта таблица содержит количественную переменную "возраст".

Для получения дескриптивной статистики запустим программу STATISTICA и с помощью команды **Открыть** раздела **Файл** главного меню откроем файл с именем "Duma", который будет представлен отдельным окном (см. рис. 1.1)



Data: DUMA.STA* (9v by 431c)

Биографические данные о депутатах I Государственной Думы

	1 ИМЯ	2 ГОД_РОЖД	3 НАЦИОН	4 ВОЗРАСТ	5 ОБР_УРОВ	6 ОБР_ПРОФ	7 СОСЛ_ПРО	8 ЗАНЯТИЕ	9 ПАРТИЯ
1	АЛЕКСАНД	1866	ЛИТОВЕЦ	40	НЕОК_СРЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	ТРУДОВИК
2	АХВЕРДОВ	1870		36	НЕОК_ВЫС	ТЕХНИЧЕС	ДВОРЯНИН	ЗЕМЛЕВЛА	КАДЕТ
3	СЕМЕНОВ	1872	БЕЛОРУСС	34			КРЕСТЬЯН	КРЕСТЬЯН	БЕСПАРТИ
4	СОКОЛОВС	1877	БЕЛОРУСС	29	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	БЕСПАРТИ
5	ЮЛЛОС	1859	ЕВРЕЙ	47	ВЫСШЕЕ	ЮРИДИЧЕС	КУПЕЦ	СВОБ_ПРО	КАДЕТ
6	АБРАМОВ	1873	ПОВОЛЖ_	33	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	ТРУДОВИК
7	АВЕРЬЯНО	1866	РУССКИЙ	40	НИЗШЕЕ	ОБЩЕЕ	ПОЧЕТ_ГР	ОБЩЕСТ_С	БЕСПАРТИ
8	АЙВАЗОВ	1873	АРМЯНИН	33	ВЫСШЕЕ	МЕДИЦИНС		ВРАЧ	КАДЕТ
9	АЛАДЬИН	1873	РУССКИЙ	33	НЕОК_ВЫС	МЕДИЦИНС	КРЕСТЬЯН	СВОБ_ПРО	ТРУДОВИК
10	АЛЕКСИНС	1872	РУССКИЙ	34	ВЫСШЕЕ	МЕДИЦИНС	ДВОРЯНИН	ВРАЧ	КАДЕТ
11	АЛЕХИН	1877	РУССКИЙ	29	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	ТРУДОВИК
12	АЛИЕВ	1858		48	ВЫСШЕЕ	СЕЛЬ_ХОЗ	КУПЕЦ	ЗЕМЛЕВЛА	КАДЕТ
13	АЛКИН	1867	ТАТАРИН	39	ВЫСШЕЕ	ЮРИДИЧЕС	ДВОРЯНИН	ЗЕМЛЕВЛА	КАДЕТ
14	АНДРЕЕВ		РУССКИЙ		ВЫСШЕЕ	ОБЩЕЕ	ДВОРЯНИН		МИРНООБН
15	АНДРЕЕВ	1866	РУССКИЙ	40	ВЫСШЕЕ	МЕДИЦИНС	ДВОРЯНИН	ВРАЧ	ТРУДОВИК
16	АНДРЕЯНО	1869	РУССКИЙ	37	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	ТРУДОВИК
17	АНДРО	1870		36	ВЫСШЕЕ	ОБЩЕЕ	ДВОРЯНИН	ЗЕМЛЕВЛА	МИРНООБН
18	АНИКИН	1869	РУССКИЙ	37	СРЕДНЕЕ	ТЕХНИЧЕС	КРЕСТЬЯН	СВОБ_ПРО	ТРУДОВИК
19	АНТОНОВ	1880	РУССКИЙ	26	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	РАБОЧИЙ	СОЦ__ДЕМ
20	АРАКАНЦЕ	1863	РУССКИЙ	43	ВЫСШЕЕ	ЮРИДИЧЕС	КАЗАК	ГОС_СЛУЖ	КАДЕТ
21	АРСЕНОВ	1869	РУССКИЙ	37	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	БЕСПАРТИ
22	АФАНАСЬЕ	1875	РУССКИЙ	31	СРЕДНЕЕ	ДУХОВНОЕ		ДУХОВЕНС	КАДЕТ
23	АФРИКАНТ	1871	РУССКИЙ	35	ВЫСШЕЕ	ВОЕННОЕ	ДВОРЯНИН	ИНЖЕНЕР	КАДЕТ
24	АХТЯМОВ	1843	БАШКИР	63	ВЫСШЕЕ	ЮРИДИЧЕС	ИЗ_ДУХОВ	ГОС_СЛУЖ	КАДЕТ
25	БРУК	1869	ЕВРЕЙ	37	ВЫСШЕЕ	МЕДИЦИНС	КУПЕЦ	ВРАЧ	КАДЕТ
26	БАРИНОВ	1864	РУССКИЙ	42	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	БЕСПАРТИ
27	БАБЕНКО	1880	РУССКИЙ	26	СРЕДНЕЕ	ТЕХНИЧЕС	КРЕСТЬЯН	РАБОЧИЙ	ТРУДОВИК
28	БАБИЧ	1865	МАЛОРУС	41	НИЗШЕЕ	ОБЩЕЕ	КРЕСТЬЯН	КРЕСТЬЯН	ТРУДОВИК
29	БАГАТУРО	1861		45	ВЫСШЕЕ	МЕДИЦИНС	ДВОРЯНИН	ВРАЧ	КАДЕТ

Рис. 1.1. Окно данных

Чтобы выбрать вариант анализа, обратимся к разделу **Анализ** главного меню программы и в раскрывшемся списке выберем первый модуль – **Основные статистики и таблицы** (рис. 1.2). Откроется диалоговое окно со списком разделов Основной статистики, из которого надо снова выбрать первый – **Описательные статистики** (рис. 1.3).

Открывшее диалоговое окно в левом верхнем углу содержит графическую кнопку **Переменные**, нажав которую, можно выбрать анализируемые признаки. Выберем "возраст" и рассмотрим подробнее наиболее важные компоненты упомянутого диалогового окна (рис. 1.4).

Во-первых, окно содержит ряд вкладок, из которых по умолчанию открыта вкладка простой дескриптивной статистики (**Быстрый**).

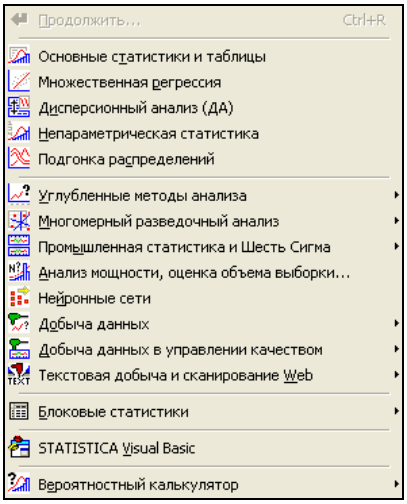


Рис. 1.2. Выбор модуля

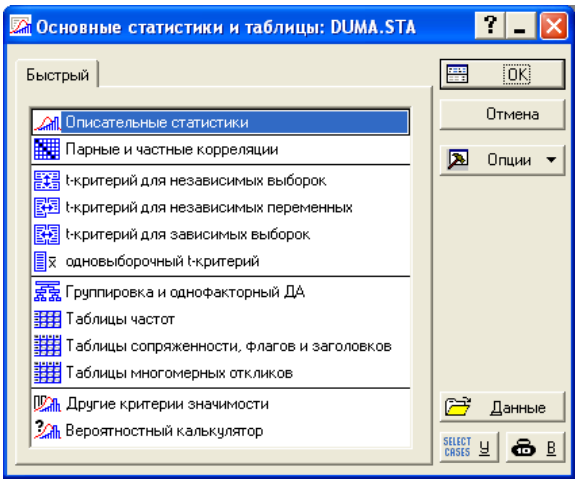


Рис. 1.3. Выбор раздела Основной статистики

Простая дескриптивная статистика прежде всего позволяет увидеть таблицу наиболее общих статистических характеристик исходных данных (графическая кнопка **Подробные описательные статистики**). В набор этих характеристик, предлагаемый программой по умолчанию, входят:

Valid N – число наблюдений, не содержащих пропусков (MD – *Missing Data*) в данной переменной, т.е. в нашем случае – число депутатов, у которых известен возраст; *Mean* – среднее арифметическое; *Standard Deviation* (среднее квадратическое отклонение), а также минимум и максимум (рис. 1.5).

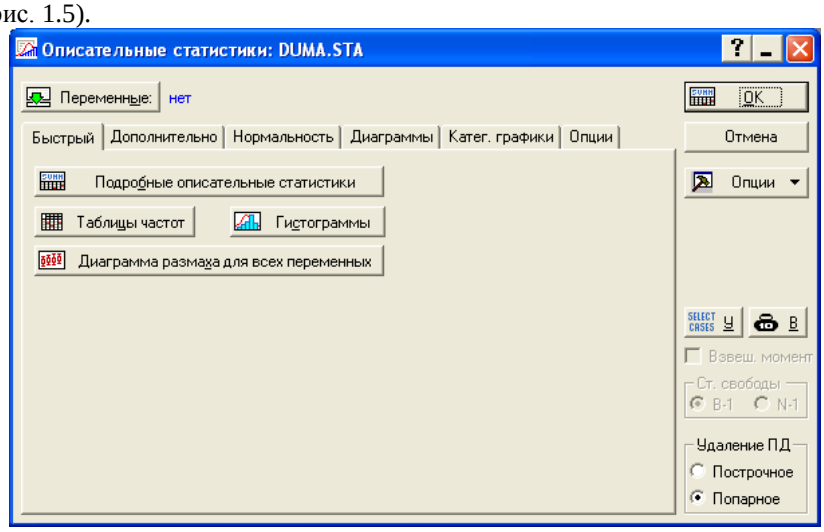


Рисунок 1.4. Диалоговое окно простой дескриптивной статистики

	Описательные статистики (DUMA.STA)				
Переменная	N набл.	Среднее	Минимум	Максимум	Стд. откл.
ВОЗРАСТ	425	40,90	26	66	8,56

Рис. 1.5. Простая дескриптивная статистика для переменной "возраст"

Кроме того, программа позволяет получить более детальную дескриптивную статистику. Выбрать нужный набор статистических характеристик позволяет вкладка **Дополнительно**. При переходе на эту вкладку вы увидите диалоговое окно (рис. 1.6), в котором можно пометить необходимые нам дополнительные характеристики.

Добавим к стандартному набору еще несколько характеристик:

Медиану;

Моду;

Нижний/верхний квартили;

Квартильный размах (разность между верхним и нижним квартилями), т.е. диапазон значений, в который попадает половина наблюдений, ближайших к медиане;

Размах – разность между минимальным и максимальным значениями признака.

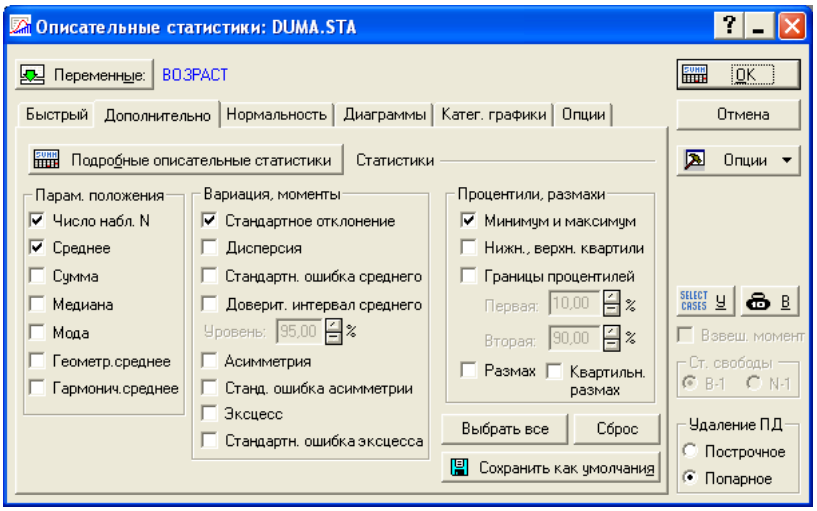


Рисунок 1.6. Диалоговое окно детальной дескриптивной статистики

Нажав после выбора графическую кнопку **Sподробные описательные статистики**, получим результат в виде таблицы, представленной на рис. 1.7.

	Описательные статистики (DUMA.STA)											
Переменная	N набл.	Среднее	Медиана	Мода	**Частота моды**	Минимум	Максимум	Нижняя Квартиль	Верхняя Квартиль	Размах	Квартиль Размах	Стд.откл.
ВОЗРАСТ	425	40,90	40	37	23	26	66	34	46	40	12	8,56

Рис. 1.7. Детальная дескриптивная статистика для переменной "возраст"

Итак, средний возраст депутатов – почти 41 год при среднем квадратическом отклонении около 8,6 года.

Поскольку коэффициент вариации в таблице на рис. 1.5 (в программе Statistica не предусмотрено его вычисление) отсутствует, подсчитаем его, пользуясь имеющимися данными. В нашем случае коэффициент вариации равен 8,6/40,9*100% ≈ 21%.

Качественный вывод из проделанных вычислений: большинство депутатов – люди среднего возраста, причем большинство депутатов имеют возраст в диапазоне 32-50 лет (или средний возраст 41 при относительном отклонении 21%.). При этом половина депутатов (интервал между нижним и верхним квартилями) имеет возраст от 34 до 46 лет.

Сравните среднее арифметическое значение и медиану. Среднее значение переменной "возраст" – около 41 года, т.е. больше, чем медиана (40).

Как видим, разница невелика и может быть связана с тем, что в наших данных присутствует небольшая асимметрия: диапазон значений признака (26 – 66 лет) смещен вправо относительно среднего, т.к. среди депутатов, возраст которых сильно отличается от среднего, преобладают пожилые, а не молодые люди (значения их возраста отклоняются далеко вправо по оси возрастов от среднего).

1.2. ЧАСТОТНЫЕ РАСПРЕДЕЛЕНИЯ

При работе с пакетом STATISTICA всегда предполагается, что исходные данные имеют вид таблицы "объекты-признаки", т.е. каждый признак (как количественный, так и качественный) задается для каждого объекта. Однако чем больше объем совокупности, тем чаще повторяются значения признаков у разных объектов. Например, в таблице Duma встречаются люди одинакового возраста, с одинаковыми профессиями или уровнем образования и т.д. Поэтому кроме средних величин изучают **распределения** признаков, которые дают информацию о том, сколько раз встречаются различные значения признаков, т.е. каковы их **частоты**. Кроме того, более адекватное представление о распределении дают упорядоченные значения признака (если речь идет о количественных или ранговых признаках).

Таким образом, мы приходим к идее **вариационного ряда** – для каждого признака это упорядоченный ряд значений, которые встречаются в исходных данных, с указанием их частот. Частоты могут выражаться как абсолютными числами (количества объектов), так и относительными (доли или проценты).

Вариационные ряды можно строить как для качественных, так и для количественных признаков. Для номинальных признаков порядок категорий в вариационном ряду не имеет значения, а для ранговых и количественных признаков значения упорядочивают. При этом для количественных признаков различают две ситуации: а) признак может принимать любые значения из некоторого диапазона (является непрерывным); б) признак может принимать только конечное число отдельных значений (является дискретным). Так, возраст в предыдущем примере может принимать только целые значения.

Значения непрерывных (а также и дискретных признаков, если число отдельных значений достаточно велико) принято группировать в интервалы, т.к. учет всех различных значений привел бы к слишком длинному вариационному ряду, в котором, к тому же, частоты были бы очень небольшими. Так, значения возраста часто группируют в интервалы длиной 5 лет: 20-25, 25-30 и т.д.

1.2.1. Частотные распределения количественных признаков

Для построения простых частотных таблиц в программе STATISTICA можно использовать графическую кнопку **Таблицы частот** на вкладке **Быстрый** (рис. 1.4). Если щелкнуть по соответствующей графической кнопке, программа выдает таблицу с параметрами по умолчанию (на рис. 1.8 показан результат построения частотного распределения признака "возраст" для файла Duma с числом интервалов, равным по умолчанию 10).

Как уже отмечалось, такие таблицы могут быть построены для любой переменной (как количественной, так и качественной), но значения количественной переменной можно естественным образом выстроить по порядку и кроме частот (*Count*) или долей (*Percent*) всех значений (или интервалов) подсчитать кумулятивные (т.е. накопленные) частоты (*Cumulative Count*) или доли (*Cumul.* %) значений признака. При этом подсчет долей (процентов) значений признака можно вести относительно только известных значений (*of Valid*) или относительно всех значений признака (*of All*).

Кумулятивная частота – это частота встречаемости текущего значения вместе со всеми предшествующими в упорядоченном ряду. Следовательно, кумулятивная доля показывает вклад текущего значения в общее количество. Зная кумулятивные частоту и долю, мы сразу можем сказать, содержит ли некоторая часть совокупности доминирующую группу значений или нет. Например, переменная может содержать несколько десятков различных значений, а совокупная встречаемость первых пяти из них составлять более половины.

	Таблица частот: BOZRACT (DUMA.STA) K-C d=,06865, p<,05 ;Лиллиефорса p<,01					
Группа	Частота	Кумул. частота	Процент допуст.	Кумул. % допуст.	% всех наблюд.	**Кумул. % от всех**
20,00000<x<=25,00000	0	0	0,00	0,00	0,00	0,00
25,00000<x<=30,00000	40	40	9,41	9,41	9,28	9,28
30,00000<x<=35,00000	86	126	20,24	29,65	19,95	29,23
35,00000<x<=40,00000	97	223	22,82	52,47	22,51	51,74
40,00000<x<=45,00000	82	305	19,29	71,76	19,03	70,77
45,00000<x<=50,00000	67	372	15,76	87,53	15,55	86,31
50,00000<x<=55,00000	24	396	5,65	93,18	5,57	91,88
55,00000<x<=60,00000	18	414	4,24	97,41	4,18	96,06
60,00000<x<=65,00000	8	422	1,88	99,29	1,86	97,91
65,00000<x<=70,00000	3	425	0,71	100,00	0,70	98,61
Пропущ.	6	431	1,41		1,39	100,00

Рис. 1.8. Частотное распределение для признака "возраст"

Более широкий спектр возможностей предлагает, однако, специальный раздел модуля **Основные статистики и таблицы**, который называется **Таблицы частот** (список разделов модуля **Основные статистики**… см. на

рис. 1.3). Выбрав этот раздел, мы переходим в диалоговое окно частотных распределений, показанное на рис. 1.9.

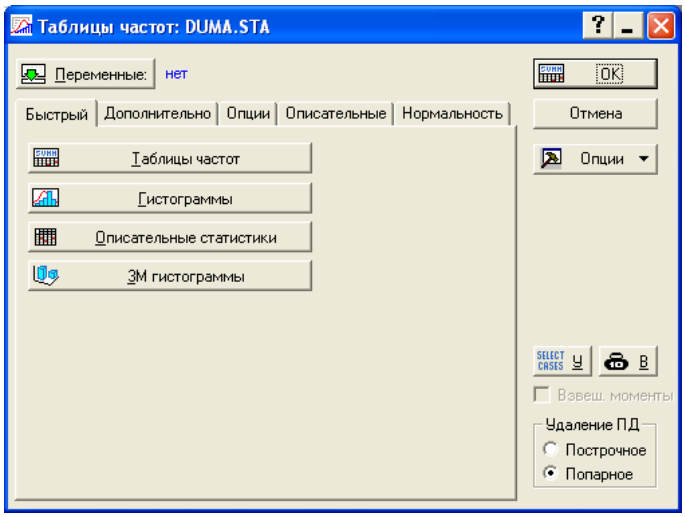


Рис. 1.9. Диалоговое окно частотных распределений

Вкладка **Опции** позволяет определить, какие из возможных частотных характеристик будут выводиться на экран (по умолчанию это простые и кумулятивные частоты и проценты – см. рис. 1.10)

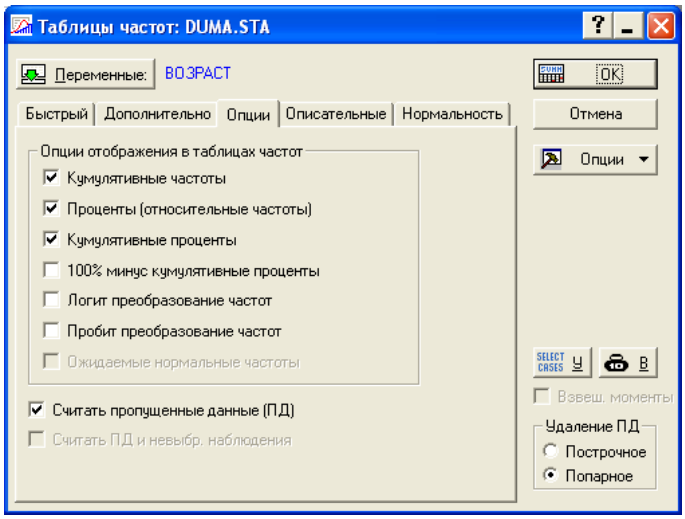


Рис. 1.10. Определение вывода на экран параметров частотной таблицы

Если перейти на вкладку **Быстрый** диалогового окна **Таблицы частот**, и нажать графическую кнопку **Таблицы частот**, то для количественного признака будет построена частотная таблица всех отдельных (дискретных) значений признака (по умолчанию учитываются все отдельные значения). Но это не всегда удобно, т.к. их может быть слишком много. Поэтому на основании таблиц частот отдельных значений количественного признака можно правильно подобрать размеры и число интервалов при переходе от дискретного к интервальному вариационному ряду. Это способствует более компактному представлению о распределении признака.

Изменить параметры частотной таблицы позволяет вкладка **Дополнительно,** где можно задавать параметры группировки значений признака:

- все различные значения – по умолчанию для количественных признаков;
- размер шага;
- число равных интервалов (диапазон значений переменной делится на заданное число интервалов, и границы интервалов получаются точными);
- приблизительное число интервалов (строятся интервалы с округленными значениями границ – часто эти значения оканчиваются на 0 или на 5).

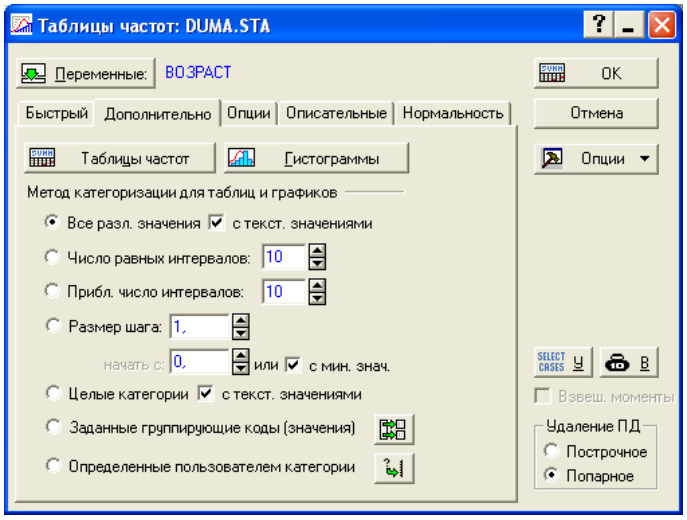


Рис. 1.11. Методы категоризации значений признака

Пример 1.2. Вновь обратимся к таблице Duma.sta и построим частотные таблицы для признака "возраст", выбирая либо длину интервала, либо

число интервалов. Напомним, что в блоке **Метод категоризации для таблиц и графиков** есть возможность построения дискретных вариационных рядов. Для того, чтобы построить такой ряд, надо оставить переключатель в позиции **Все различные значения** (рис. 1.11). Однако для выбранного признака лучше строить интервальные ряды.

а) Построим интервальный вариационный ряд, задавая размер возрастных групп, например, 5 лет (в поле **Размер шага**, т.е. размер интервала вкладки **Дополнительно** диалогового окна **Таблицы частот** надо поставить число 5). Щелчок по графической кнопке **Таблицы частот** дает результат, показанный на рис. 1.12.

При этом способе группировки число групп заданного размера вычисляется автоматически, а за нижнюю границу первого интервала берется минимальное значение признака.

Мы видим, что первые четыре интервала на рис. 1.12 в совокупности покрывают 70% значений возраста.

	Таблица частот: ВОЗРАСТ (DUMA.STA)			
От До	Частота	Кумул. частота	Процент	Кумул. процент
26,00000<=x<31,00000	40	40	9,28	9,3
31,00000<=x<36,00000	86	126	19,95	29,2
36,00000<=x<41,00000	97	223	22,51	51,7
41,00000<=x<46,00000	82	305	19,03	70,8
46,00000<=x<51,00000	67	372	15,55	86,3
51,00000<=x<56,00000	24	396	5,57	91,9
56,00000<=x<61,00000	18	414	4,18	96,1
61,00000<=x<66,00000	8	422	1,86	97,9
66,00000<=x<71,00000	3	425	0,70	98,6
Пропущ.	6	431	1,39	100,0

Рис. 1.12. Интервальный ряд распределения по возрасту с длиной интервала 5 лет

б) Теперь зададим число групп, равное пяти, в поле **Приблизительное число интервалов**. Результат приведен на рис. 1.13.

	Таблица частот: ВОЗРАСТ (DUMA.STA)			
От До	Частота	Кумул. частота	Процент	Кумул. процент
20,00000<x<=30,00000	40	40	9,28	9,3
30,00000<x<=40,00000	183	223	42,46	51,7
40,00000<x<=50,00000	149	372	34,57	86,3
50,00000<x<=60,00000	42	414	9,74	96,1
60,00000<x<=70,00000	11	425	2,55	98,6
70,00000<x<=80,00000	0	425	0,00	98,6
Пропущ.	6	431	1,39	100,0

Рис. 1.13. Интервальный ряд распределения по возрасту с числом интервалов, равным 5

При этом способе группировки программа вычисляет автоматически размер интервала, помещая минимум и максимум в центре первого и последнего интервалов, соответственно.

В данном случае уже первые две группы дают в совокупности более 50% всех депутатов.

1.2.2. Частотные распределения качественных признаков

Напомним, что для качественных признаков категории в частотных распределениях играют ту же роль, что и интервалы для количественных признаков, т.е. можно считать абсолютные и относительные частоты категорий. Однако подсчет кумулятивных частот имеет смысл для качественного признака лишь в том случае, если его категории упорядочены, т.е. если он является ранговым.

Пример 1.3. Вернемся к таблице данных по депутатам 1-й Государственной думы (файл Duma.sta). В этой таблице большинство признаков являются качественными, причем номинальными (за исключением уровня образования – это ранговый признак). Построим частотное распределение признака "уровень образования". Поскольку эта переменная является по существу ранговой, процедура построения частотных распределений (как и в случае количественного признака) может включать не только обычные, но и накопленные (кумулятивные) частоты: количества и доли депутатов, имеющих уровень образования не ниже или не выше данного (в зависимости от порядка категорий).

Если попробовать построить частотное распределение признака "уровень образования" по аналогии с признаком "возраст" (в блоке **Метод категоризации для таблиц и графиков** надо в этом случае выбрать **Целые категории**), то окажется, что таблица результатов содержит категории признака в произвольном порядке, а не по возрастанию или убыванию уровня образования. Это происходит потому, что в пакете STATISTICA всем текстовым данным ставятся в соответствие числовые коды, которые и используются при всех операциях с данными. Очевидно, что программа не знает смысла названий текстовых данных и поэтому производит оцифровку категорий качественных переменных произвольным образом.

Например, если в таблице данных дважды щелкнуть на имени переменной "уровень образования", а потом нажать графическую кнопку **Текстовые метки**, можно видеть что категория образования "высшее" имеет код 104, "неоконченное высшее" – код 102, "среднее" – код 105, "неоконченное среднее" – код 101, "низшее" – код 103, "малограмотный" – код 106 и "неграмотный" – код 107. Для того чтобы восстановить естественный порядок

категорий признака (т.е. "неграмотный" обозначить кодом "1", "малограмотный" – кодом "2", "низшее" – кодом "3" и т.д.), необходимо либо перекодировать значения, либо создать новую переменную, которая будет содержать числовые ранги. Перекодировку категорий рангового признака можно выполнить непосредственно в окне **Текстовые метки** (см. рис. 1.14).

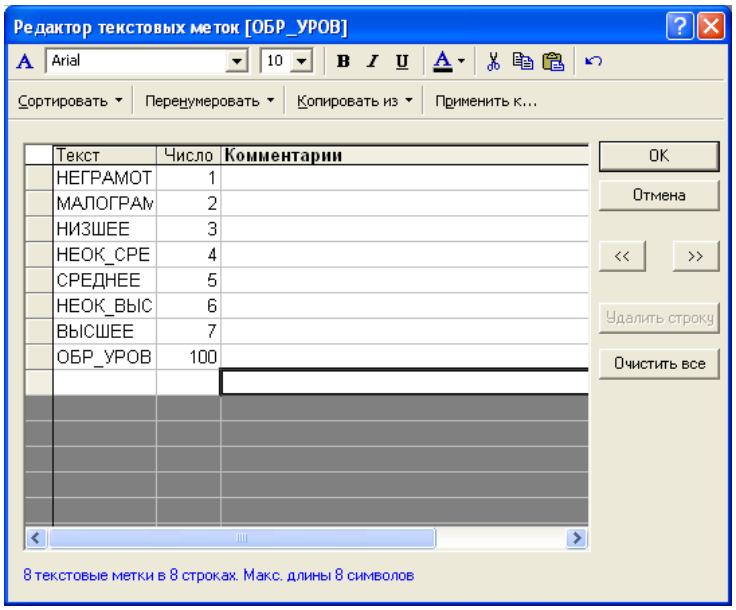


Рис. 1.14. Перекодировка категорий ранговой переменной "уровень образования"

Поскольку порядок рангов можно поменять на обратный, то интересно также добавить к частотной таблице колонку **100% минус кумулятивные проценты** (т.е. кумулятивные относительные частоты в обратном порядке). Для этого на вкладке **Опции** диалогового окна **Таблицы частот** надо пометить соответствующее поле (поставить флажок). Щелчок по графической кнопке **Таблицы частот** дает готовую частотную таблицу.

В табл. 1.1 приведен результат работы модуля **Таблицы частот** с перекодированной переменной "уровень образования".

Теперь рассмотрим построение частотных распределений *номинальных* качественных признаков. Здесь категории не могут быть упорядочены, и кумулятивные показатели теряют смысл. Для того, чтобы отключить по-

строение кумулятивных показателей, на вкладке **Опции** диалогового окна **Таблицы частот** надо "снять" соответствующие флажки: **Кумулятивные частоты, Кумулятивные проценты** и **100 минус кумулятивные проценты**. При этом надо оставить "включенными" **Проценты (относительные частоты)** для того, чтобы в таблице результатов присутствовала колонка с относительными частотами (процентными долями) категорий признака, а не только с абсолютными частотами.

Таблица 1.1. Частотное распределение признака "уровень образования"

Категория	Числовой ранг	Абсол. частота	Кумулят. абс. частота	%	Кумулят. процент	Кум. проц. в обрат. порядке
неграмот.	1	2	2	0,47	0,47	100,00
малограм.	2	17	19	3,95	4,42	99,53
низшее	3	114	133	26,51	30,93	95,58
неок. средн.	4	10	143	**2,33**	**33,26**	**69,07**
среднее	5	48	191	11,17	44,42	66,74
неок. высш.	6	14	205	3,26	47,67	55,58
высшее	7	225	430	52,33	100,00	52,32

Примечание. Числа, выделенные жирным шрифтом в табл. 1.1, указывают соответственно долю депутатов:
– с неоконченным средним образованием;
– с неоконченным средним образованием или более низким (образование ниже "среднего");
– с неоконченным средним образованием или более высоким (образование выше "низшего").

Пример 1.4. Построим частотное распределение признака "профиль образования" в таблице Duma (результаты представлены в табл. 1.2 и для удобства упорядочены по алфавиту).

Таблица 1.2. Частотное распределение признака "Профиль образования"

Категория	Абсолютная частота	Процент
военное	22	5,1
гуманитарное	14	3,2
духовное	26	6,0
естественнонаучное	17	3,9
медицинское	30	6,9
общее	129	29,9
педагогическое	16	3,7
разное	5	1,1
сельскохозяйственное	17	3,9
техническое	24	5,5

экономическое	3	0,6
юридическое	73	16,9
Missing	55	12,7

Примечание. Как обычно, в строке *Missing* (*пропущенные данные*) подсчитано число и доля депутатов, для которых нет сведений о профиле образования.

1.3. ВИЗУАЛИЗАЦИЯ ДАННЫХ

Графическое изображение частотного распределения называется **гистограммой**. Гистограмма показывает зависимость частоты встречаемости признака от соответствующего интервала группировки. Разумеется, вид гистограммы существенно зависит от количества интервалов: чем больше интервалов и чем меньше длина каждого из них, тем более четко выступают характерные черты распределения: *симметричность, унимодальность (одновершинность)* и т.п. Гистограмма также показывает *моду* распределения.

Гистограмма с параметрами по умолчанию доступна при нажатии графической кнопки **Гистограммы** на вкладке **Быстрый** (для количественных признаков при этом выбирается интервальный ряд с числом интервалов, равным 10 и границами интервалов, заканчивающимися на 0 или 5). Можно также получить гистограмму прямо из исходной таблицы: выделив нужный признак и щелкнув правой кнопкой мыши, выбрать в появившемся контекстном меню **Графики исходных данных | Гистограмма.**

Для получения более сложных гистограмм на вкладке **Дополнительно** в диалоговом окне **Таблицы частот** надо задать те же параметры группировки, что и для построения частотных распределений, а затем нажать на графическую кнопку **Гистограммы**. Наконец, все возможности построения графиков доступны из пункта **Графика** главного меню программы.

Пример 1.5. Построим гистограммы распределения депутатов I Государственной думы (файл Duma.sta) по возрасту, сначала задавая длину интервала группировки (пять лет), а затем – число интервалов группировки (пять).

Воспользовавшись любым из указанных выше способов, построим обе гистограммы (результаты показаны на рис. 1.15).

На рис. 1.15 числа на границах колонок по горизонтальной оси показывают интервалы значений возраста. По вертикальной оси откладывается число наблюдений (объектов) – в нашем случае, депутатов Думы – в соответствующей возрастной группе.

Заметим, что в данном случае мы имеем одномодальное (одновершинное) распределение, которое не является вполне симметричным.

а) длина интервала = 5 б) число интервалов = 5

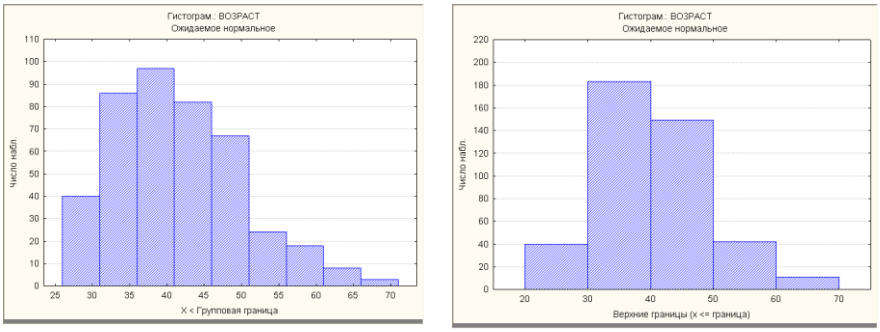


Рис. 1.15. Гистограммы распределения по возрасту (файл Duma.sta)

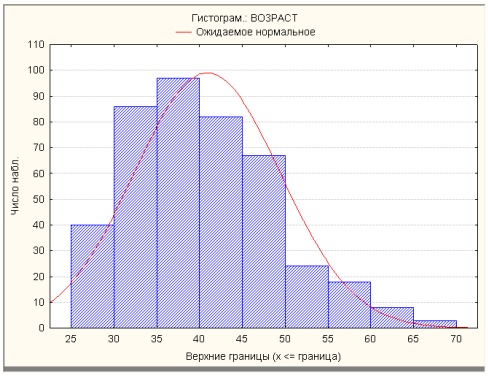


Рис. 1.16

Часто гистограмма используется в статистических пакетах для сопоставления распределения с *нормальным* (для проверки гипотезы о том, что значения данного признака *распределены по нормальному закону* – очень важному в теории вероятностей типу распределения) [1]. С этой целью на изображение реальной гистограммы накладывается теоретически вычисленный на осно-

[1] Среди всех вероятностных распределений есть такие, которые особенно часто используются на практике, они хорошо изучены. Особую роль играет т.н. **нормальное распределение**, которое часто реализуется во многих ситуациях, в которых на поведение случайной величины влияет большое количество независимых случайных факторов, среди которых нет сильно выделяющихся. Нормальное распределение можно изобразить графически в виде симметричной одновершинной кривой, напоминающей по форме колокол. Высота (ордината) каждой точки этой кривой показывает, как часто встречается соответствующее значение. Эти ординаты обобщают введенное ранее понятие частоты вариационного ряда. Форма нормальной кривой и положение ее на оси абсцисс полностью определяются двумя параметрами: средним арифметическим значением и средним квадратическим отклонением. Вершина кривой соответствует среднему арифметическому значению, т.е. наиболее часто встречаются значения, близкие к среднему, а по мере удаления от него частота падает. Более подробно нормальное распределение рассматривается в главе 2.

ве среднего арифметического и среднего квадратического отклонения график нормального распределения (см. рис. 1.16).

Аналогично строятся гистограммы и для качественных признаков. Для большинства из них, однако, порядок групп (категорий) не имеет значения, Поэтому для них гораздо интереснее выглядят графики другого типа, а именно – круговые диаграммы, которые отображают долю каждой категории признака в виде соответствующего сектора круга.

Пример 1.6. Вернемся к нашим данным и построим круговую диаграмму для признака "профиль образования". Выберем команду **2М Графики | Круговые диаграммы** в пункте **Графика** основного меню программы. Откроется диалоговое окно, представленное на рис. 1.17.

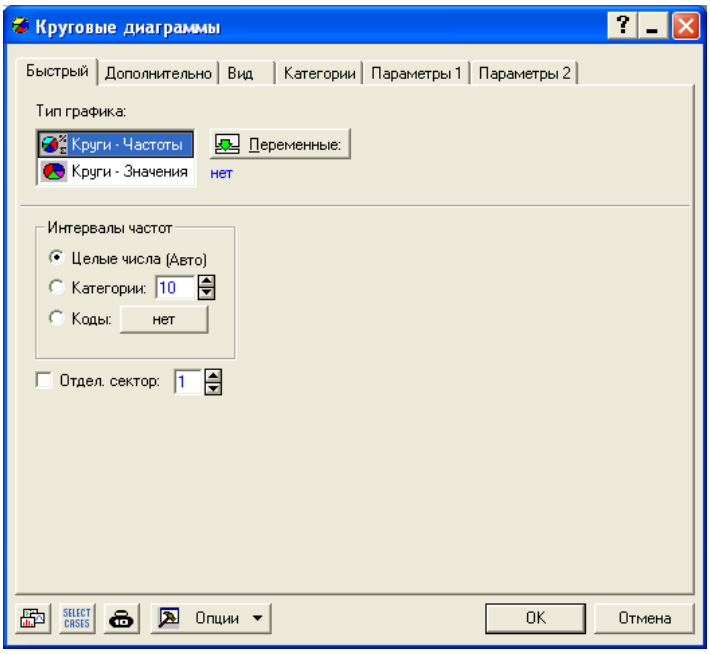


*Рис. 1.17. Диалоговое окно построения круговых диаграмм (вкладка **Быстрый**)*

В диалоговом окне на вкладке **Быстрый** (см. рис. 1.17) или вкладке **Дополнительно** (см. рис. 1.18) в блоке **Тип графика** обязательно надо указать **Частоты**, а не **Значения** (*исходные*). В блоке **Интервалы частот** надо выбрать **Целые числа (Авто)**, что соответствует работе с номинальными данными, а в блоке **Условные обозначения** (только на вкладке **Дополнительно**) – **Текст и процент**, чтобы видеть не только названия кате-

горий, но и частоты их встречаемости в совокупности объектов. Щелкнув по графической кнопке **OK**, получим результат, показанный на рис. 1.19.

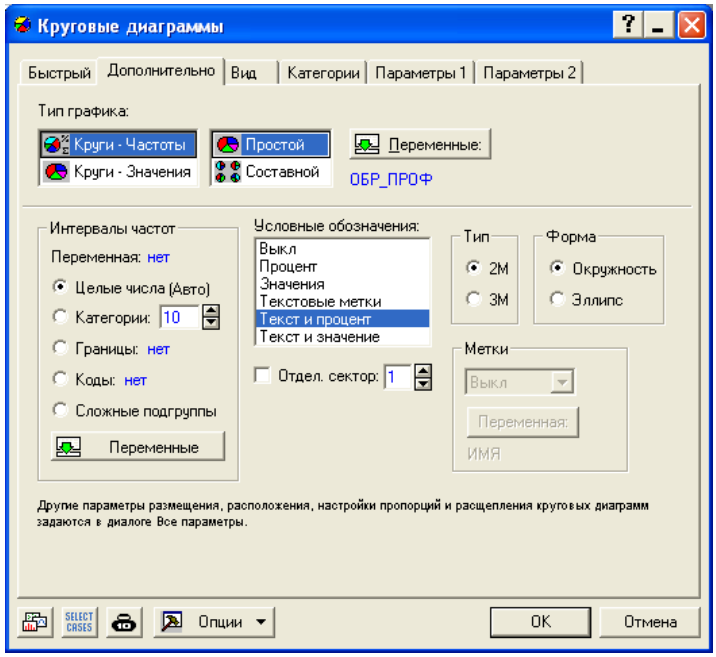


*Рис. 1.18. Диалоговое окно построения круговых диаграмм (вкладка **Дополнительно**)*

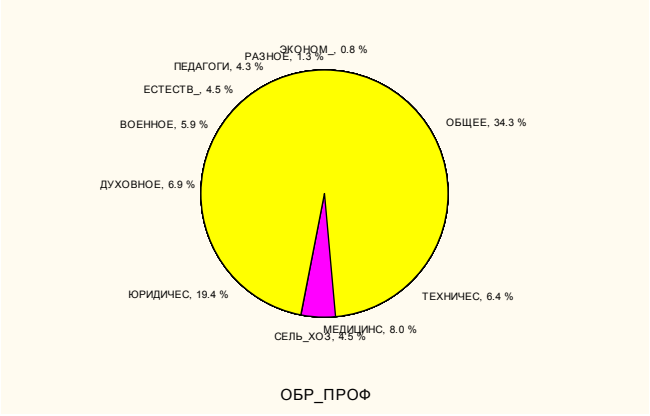


Рис. 1.19. Круговая диаграмма с параметрами, соответствующими рис. 1.18

До сих пор предполагалось, что строками в таблице исходных данных являются единичные объекты, а столбцами – конкретные значения признаков (качественных или количественных) для этих объектов. Однако в статистике исходным материалом часто служат уже сгруппированные данные, например, готовые вариационные ряды, когда строками являются интервалы значений какого-либо количественного признака или категории какого-либо качественного признака, а в столбцах стоят частоты (число единичных объектов, попадающих в каждую группу).

В первом случае для графического изображения необходимо сначала построить частотное распределение, тогда как во втором случае это распределение уже дано. Поэтому пользоваться описанными выше способами визуализации данных нельзя (иначе вы будете строить "гистограмму гистограммы"). Поэтому надо помнить, что для изображения уже готовых вариационных рядов в программе STATISTICA нельзя обращаться к командам и графическим кнопкам, которые называются **Гистограммы**. В этом случае подходят команды построения двумерных и трехмерных графиков из раздела **Графика** главного меню (**2M Графики** или **3M Последовательные графики**).

Пример 1.7. Обратимся к таблице распределения заболеваемости основными заразными болезнями по регионам Российской Империи в 1912 г. (файл Deseases.sta). В отличие от предыдущего файла в этой таблице содержатся не индивидуальные данные о том, какой болезнью заболел каждый человек, а суммарные данные, сгруппированные по названиям болезней. Таким образом, в строках этой таблицы стоят различные категории болезней, в столбцах – количество заболевших (т.е. частоты этих категорий). Различным регионам страны соответствуют разные столбцы таблицы.

	Заразные болезни в России в 1912 г. (чел.)							
	1 ЕВР_РОС	2 КАВКАЗ	3 СИБИРЬ	4 СР_АЗИЯ	5 ИТОГО	6 ПОЛЬША	7 ПО_ИМПЕР	8 НА_100_Б
ОСПА	63801	3983	4783	1156	73723	7865	81588	0,40
СКАРЛАТИ	294288	15070	18758	5898	334014	16242	350256	1,72
ДИФТЕРИТ	380993	16832	14770	8478	421073	10772	431845	7,02
КОРЬ	348763	22867	25319	6550	404499	16308	419807	2,06
КОКЛЮШ	451283	19322	31599	10826	513030	17155	530185	2,60
ГРИПП	3073041	91252	183307	55612	3403212	37070	3440282	16,87
ТИФ_РАЗН	485065	26837	27275	12137	551314	18026	569340	2,79
ДИЗЕНТЕР	345575	35696	34059	9481	424811	11309	436120	2,14
ХОЛЕРА	3131	403	247	20	3792	1317	5109	0,03
ЭПИД__ГА	252232	36030	36597	11748	336571	16837	353444	1,73
ЗАУШНИЦА	209812	17217	13189	5818	246036	6100	252136	1,24
РОЖА	166850	12145	6771	4782	190548	5512	196060	0,96
РЕВМАТИЗ	589651	80807	44750	19724	734932	17780	752712	3,69
ЦИНГА	45627	5560	17187	34773	103147	657	103804	0,51

Рис. 1.20. Исходные данные (файл Deseases.sta)

Построим графическое изображение уже готового вариационного ряда заболеваемости по Европейской России (первый столбец). Поскольку категории болезней не могут быть упорядочены, выберем круговую диаграмму для представления структуры заболеваемости.

Мы уже рассматривали построение круговой диаграммы для признака "профиль образования" в таблице Duma, поэтому обратим внимание на отличия в использовании этого типа графика в данном случае.

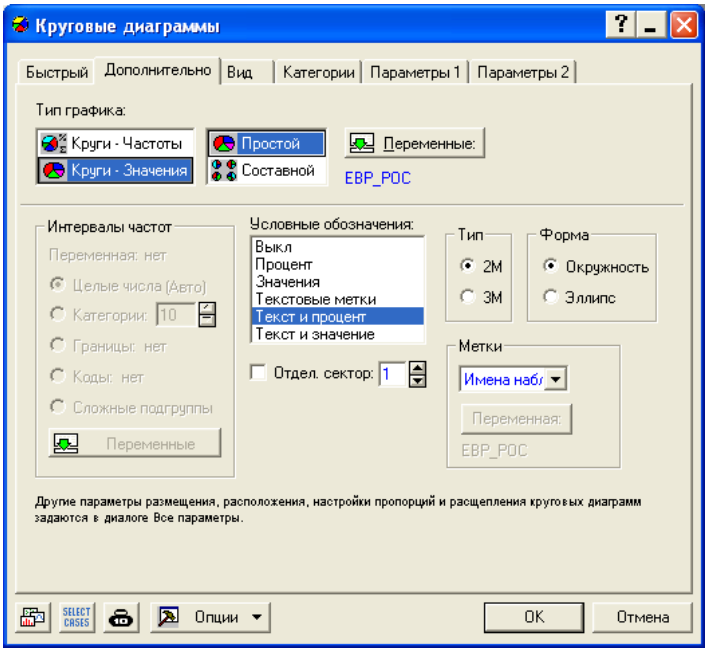


Рис. 1.21. Диалоговое окно построения круговой диаграммы для сгруппированных данных

Как и раньше, выберем команду **2M Графики | Круговые диаграммы** в пункте **Графика** основного меню программы. На вкладке **Дополнительно** в блоке **Тип графика** надо обязательно надо указать **Значения** (исходные значения), тогда программа не станет вторично группировать данные. В блоке **Метки** надо выбрать **Имена наблюдений** (т.е. *объектов*, в нашем случае – имена категорий), чтобы названия категорий появились на графике. Наконец, в блоке **Условные обозначения** надо, как и раньше, выбрать **Текст и процент**, чтобы на графике были проставлены не только названия категорий, но и относительные частоты их встречаемости (см. рис. 1.21).

После нажатия графической кнопки **ОК** получится круговая диаграмма, представленная на рис. 1.22.

Для сравнения структуры заболеваемости в различных регионах можно на одном двумерном (или даже трехмерном, объемном) графике построить сразу несколько распределений. На рис. 1.23 показан такой линейный график для Сибири и Средней Азии, полученный при выполнении последовательности действий **Графика | 2М Графики | Последовательные/Наложенные графики** и выбором двух упомянутых регионов как **Переменных**. Анализ такого графика позволяет сделать вывод о существенном совпадении структуры заболеваемости в двух регионах страны.

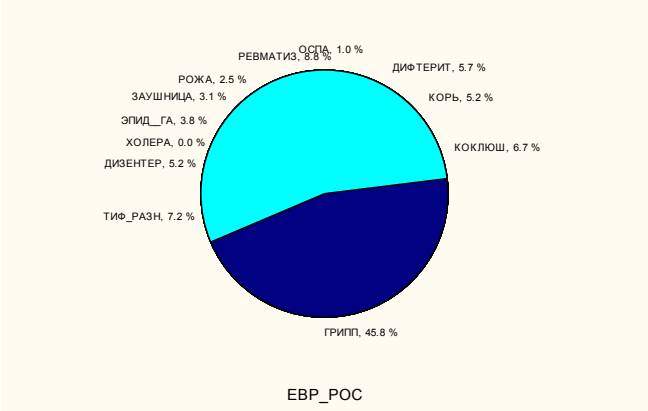


Рис. 1.22. Круговая диаграмма, соответствующая параметрам рис. 1.21

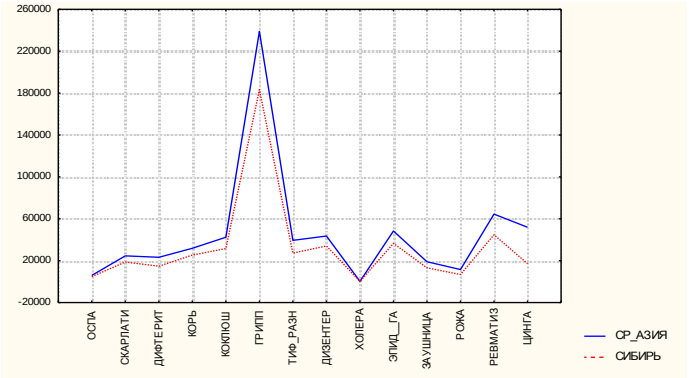


Рис. 1.23. Линейный график для двух вариационных рядов

1.4. КАТЕГОРИЗОВАННЫЕ РАСПРЕДЕЛЕНИЯ [1]

Дескриптивная статистика позволяет также строить распределения значений одной переменной в зависимости от значений другой. Чаще всего категоризованные распределения строят для того, чтобы выявить взаимосвязи между количественным и качественным признаками, точнее, влияние категории качественного признака на характер распределения количественного. Результаты при этом представляются либо в виде категоризованных гистограмм, либо в виде категоризованных средних.

Пример 1.8. Вернемся к таблице Duma.sta. Эта таблица кроме количественного признака "возраст" содержит номинальный признак "партия", отражающий фракционную принадлежность депутата. Построим распределение депутатов по возрасту отдельно для фракций кадетов и трудовиков.

Прежде всего, вернемся к диалоговому окну дескриптивной статистики, воспользуемся графической кнопкой **Переменные** и выберем тот признак, для которого будут строиться распределения – "возраст". Далее перейдем на вкладку **Категоризованные графики** (см. рис. 1.24).

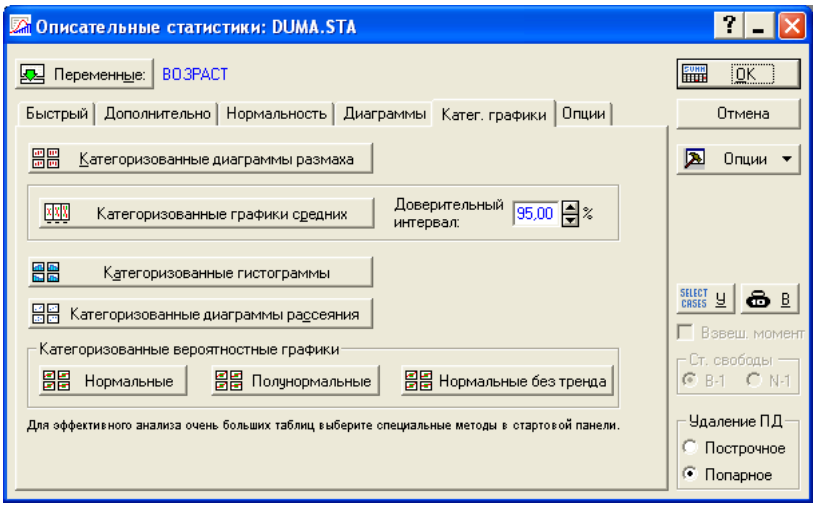


*Рис. 1.24. Вкладка **Категоризованные графики** диалогового окна дескриптивной статистики*

[1] Этот раздел имеет отношение также и к материалу главы 3, поскольку он касается зависимости между признаками. Исходя из этого, к нему полезно вернуться при изучении методов анализа взаимосвязей.

Нажав на кнопку **Категоризованные гистограммы**, можно выбрать группирующие признаки (их быть несколько, но мы ограничимся более простым случаем – см. рис. 1.25). В нашем случае это будет признак "партия". Нажав кнопку **OK**, мы окажемся в диалоговом окне (см. рис. 1.26), в котором необходимо указать, для каких значений группирующей переменной мы хотим строить распределение.

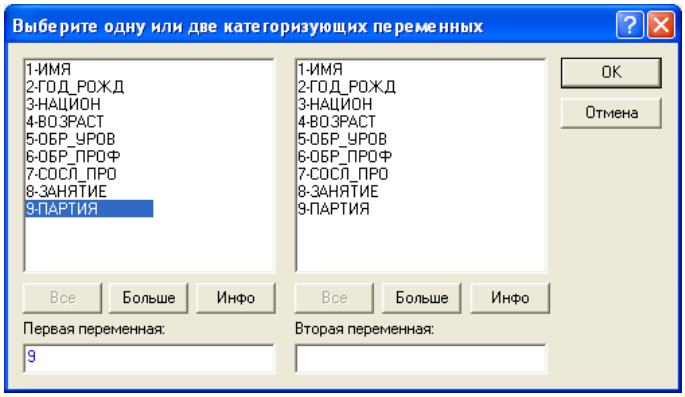


Рис. 1.25. Окно выбора группирующей переменной

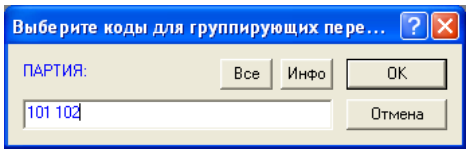


Рис. 1.26. Диалоговое окно выбора групп для построения категоризованных распределений

В нашем распоряжении две графические кнопки – **Все** (*все категории*) и **Инфо** (*просмотр списка категорий*). Если вы не помните числовых кодов категорий, нажмите **Инфо** и просмотрите список кодов; затем впишите нужные, разделяя их пробелами. Задав коды, соответствующие значениям "кадет" и "трудовик", мы получим одновременно две гистограммы, приведенные на рис. 1.27.

Видно, что для двух разных категорий параметры распределения несколько различаются. К примеру, мода возраста для категории "трудовик" находится на интервале от 30 до 35 лет, в то время как для "кадетов" – от 35 до 40 лет. Видно также, что во фракции трудовиков больше молодых депутатов, тогда как во фракции кадетов – больше пожилых.

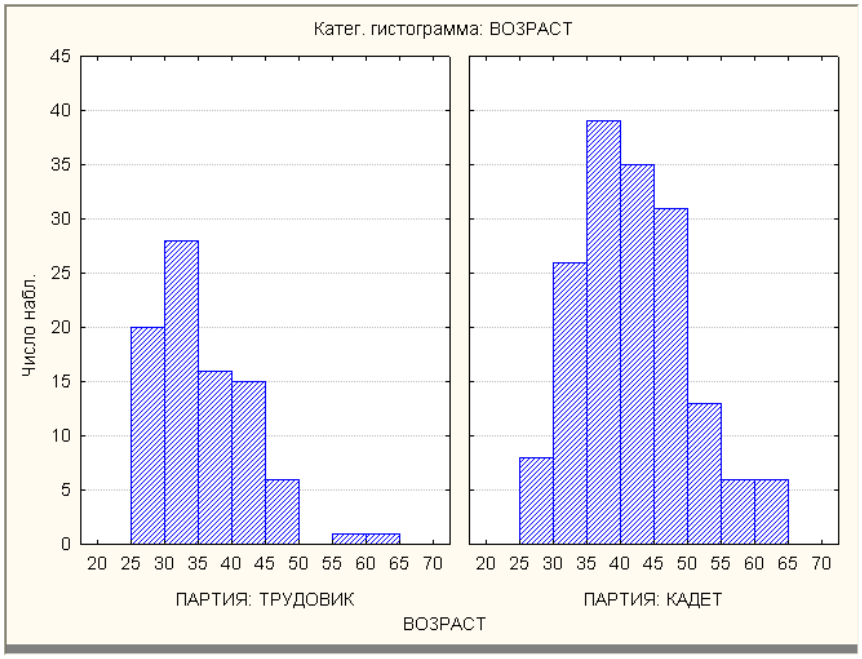


Рис. 1.27. Распределение по возрасту в зависимости от фракционной принадлежности

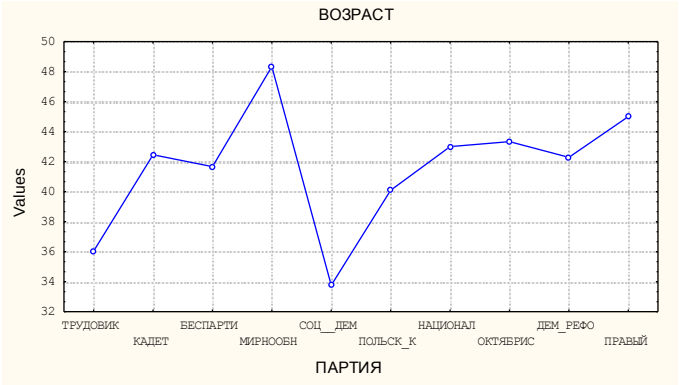


Рис. 1.28. Зависимость среднего возраста от категории группировочного признака "партия"

Если же на вкладке **Категоризованные графики** в диалоговом окне дескриптивной статистики воспользоваться другой графической кнопкой –

Категоризованные графики средних, то можно получить графическое представление зависимости среднего возраста от фракционной принадлежности по всем значениям признака "партия" (см. рис. 1.28). Для того чтобы получить результат, необходимо выполнить ту же последовательность действий, что и для построения категоризованных гистограмм, только в диалоговом окне выбора категорий для признака "партия" надо указать – **Все**.

ВОПРОСЫ

1. Типы признаков.
2. Что такое количественный признак? Непрерывные и дискретные признаки.
3. Что называется вариационным рядом?
4. Что такое относительная частота?
5. Графическая интерпретация вариационного ряда.
6. Что такое гистограмма?
7. Меры среднего уровня.
8. Меры разброса.
9. В чем сходство и различие между σ и V?
10. В каких единицах измеряется коэффициент вариации?
11. Как можно сравнить два вариационных ряда?
12. Что такое категоризованное распределение?

ЗАДАНИЯ

1. Используя файл Industry.sta, построить распределение предприятий по**:**

а) числу рабочих;

б) мощности двигателей;

в) объему производства.

В какую группу попадает наибольшая доля предприятий? Объяснить значения кумулятивных частот.

2. Используя файл Industry.sta, построить распределение предприятий по:

а) числу рабочих в целом по всей совокупности;

б) числу рабочих в металлообрабатывающей отрасли промышленности;

в) объему производства по всей совокупности;

г) объему производства в металлообрабатывающей отрасли промышленности.

3. Построить гистограмму ряда:

возраст	число людей
до 20	40
20-40	60

40-80 70

4. Найти медиану ряда: 25, 20, 27, 32, 21, 17, 22, 28.
5. Найти $\bar{x}$, σ и V для ряда: 2, 3, 4, 5, 6.
6. Некий коллектив людей разбит на 3 группы, составляющие, соответственно, 1/4, 5/8 и 1/8 части от численности всего коллектива. Средний возраст в первой группе – 20 лет, во второй – 23 года и в третьей – 29 лет. Найти средний возраст для всего коллектива.
7. Первая группа, состоящая из 14 человек, имеет средний стаж работы 10 лет, а вторая группа, состоящая из 36 человек, имеет средний стаж 15 лет. Определить средний стаж объединенной группы из 50 человек.

ГЛАВА 2

ВЫБОРОЧНЫЙ МЕТОД

Перед тем как обратиться к материалу двух следующих глав, рассмотрим нормальное распределение, знакомство с которым необходимо для понимания смысла выборочного метода и метода статистической проверки гипотез.

2.1. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Среди всех вероятностных распределений есть такие, которые особенно часто используются на практике, они хорошо изучены. Особое место среди них занимает т.н. нормальное распределение.

Понятие нормального распределения играет в статистике важную роль, поскольку для корректного использования многих статистических методов необходимо, чтобы признаки подчинялись закону нормального распределения. Что же такое нормальное распределение?

Если взглянуть на гистограмму распределения любого признака, которую строит программа STATISTICA (см., например, рис. 1.16 главы 1), можно увидеть на фоне собственно гистограммы плавную симметричную колоколообразную кривую. Это и есть теоретическая кривая нормального распределения, которая показывает, что у большинства объектов значения признака близки к среднему, а значения, сильно отклоняющиеся от среднего как в одну, так и в другую сторону, встречаются тем реже, чем дальше они от среднего.

Многие эмпирические распределения реальных величин действительно близки к нормальным. Считается, что величина распределяется нормально, если на характер распределения влияют много факторов, причем ни один из них не является определяющим. Например, такие признаки, как возраст или рост людей в достаточно больших совокупностях распределены нормально, а зарплата или доход демонстрируют сильно "скошенное" влево распределение (т.е. пик такого распределения находится не посередине, а смещен в сторону меньших значений признака).

Нормальное распределение можно изобразить графически в виде симметричной одновершинной кривой, напоминающей по форме колокол. Высота (ордината) каждой точки этой кривой показывает, как часто встречается соответствующее значение. Эти ординаты обобщают введенное ранее понятие частоты вариационного ряда. Если эмпирическое распределение признака по своему характеру близко к нормальному, то форма гистограм-

мы напоминает (конечно, в огрубленном виде) форму нормальной кривой и "стремится" к этой нормальной кривой, если увеличивать число интервалов, одновременно уменьшая их величину.

Форма нормальной кривой и положение ее на оси абсцисс полностью определяются двумя параметрами: средним арифметическим значением и средним квадратическим отклонением. Вершина кривой соответствует среднему арифметическому значению, т.е. наиболее часто встречаются значения, близкие к среднему, а по мере удаления от него частота падает.

Геометрически вероятность значений, меньших данного, изображается площадью под кривой распределения слева от этого значения. Площадь под всей кривой равна 1, что соответствует полной достоверности, т.е. вероятности того, что признак вообще принимает какое-то (любое) значение.

Ввиду своей важности для практических приложений функция нормального распределения табулирована, т.е. в соответствующих таблицах многочисленных учебников каждому значению признака ставится в соответствие определенная вероятность. Разумеется, разные признаки могут быть несравнимы между собой, так как измеряются в разных единицах и соответственно имеют разный масштаб и диапазон значений. Однако если вместо исходных значений признаков использовать их отклонения от среднего, деленные на стандартное отклонение, то все разномасштабные признаки, распределенные по нормальному закону, приводятся к стандартному виду (с нулевым средним значением и стандартным отклонением, равным единице). Именно стандартизованные значения и приводятся во всех статистических таблицах, в частности и в пакете STATISTICA.

Пример 2.1. Вычислим вероятность того, что отклонение значения нормально распределенного признака от своего среднего арифметического более чем в три раза превысит стандартное отклонение этого признака. Для того чтобы сделать это, обратимся к разделу **Вероятностный калькулятор** модуля **Основные статистики и таблицы** и в списке **Распределение** выберем Z (Нормальное), а также (поскольку нас интересуют отклонения от среднего в обе стороны) включим флажок **Двусторонняя** (см. рис. 2.1).

В поле Z зададим стандартизованное отклонение (в данном случае 3) и щелкнем по графической кнопке **Вычислить**, чтобы вычислить вероятность p того, что стандартизованное отклонение признака, т.е. $x-\bar{x}$ по абсолютной величине превзойдет 3σ, т.е. $|x-\bar{x}|>3\sigma$, или вероятность того, что значение x попадет в диапазон от $\bar{x}-3\sigma$ до $\bar{x}+3\sigma$. В поле p получим ответ: 0,9973, или 99,73%. Следовательно, вероятность того, что значение признака отклонится от среднего в любую сторону на столь значительную величину, очень мала и составляет менее 0,3%. Величину 99,7% можно считать практически совпадающей с 100%. Следовательно, в интервале [$\bar{x}-3\sigma$,

$\bar{x}$ +3σ] находятся практически все значения признака (в статистике это называется правилом "трех сигм").

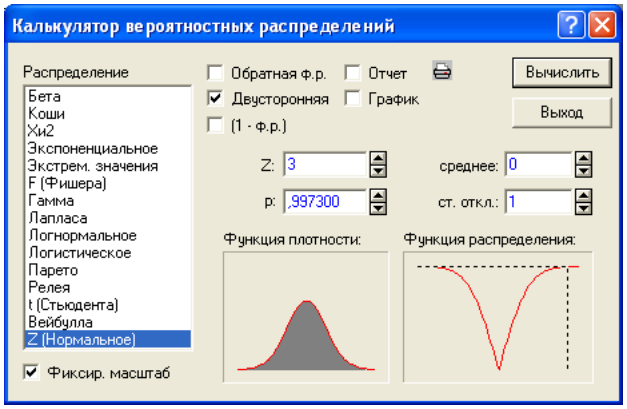


Рис. 2.1. Вычисление вероятности стандартизованного отклонения

Здесь речь шла об отклонениях от среднего в обе стороны, т.е. о симметричных интервалах, о чем говорит флажок **Двусторонняя**. Если же нас интересует отклонения от среднего в одном направлении, например, надо вычислить вероятность того, что значение *x* признака отклоняется вправо от среднего арифметического на величину, большую утроенного среднего квадратического отклонения σ ($x - \bar{x} > 3\sigma$), тогда при вычислении вероятности *p* надо убрать флажок **Двусторонняя**.

С другой стороны, можно не только вычислять вероятность стандартизованных отклонений признака, но и наоборот – по определенной вероятности вычислять эти отклонения. Вычислением отклонений через вероятности занимается раздел математической статистики, который называется статистическим оцениванием (он рассматривается в следующих разделах данной главы).

Например, можно определить величину стандартизованного отклонения, вероятность превзойти которую (по абсолютной величине) равна всего 5%.

В разделе **Вероятностный калькулятор** зададим вероятность *p* (0,05) и включим флажок (**1 – ф.р.**) (это вероятность не превзойти искомую величину). Щелчок по графической кнопке Подсчет позволяет по значению вероятности вычислить отклонение (см. рис. 2.2).

Получим, что с вероятностью 5% значение *x* отличается от своего среднего больше чем на *1,96* величины среднего квадратического отклонения *σ*. Этот вывод можно сформулировать и по-другому: с вероятностью 95% зна-

чение x отличается от своего среднего не больше чем на $1{,}96\sigma$, т.е. в интервал $[\bar{x} - 1.96\sigma, \bar{x} + 1.96\sigma]$ попадет 95% всех значений признака.

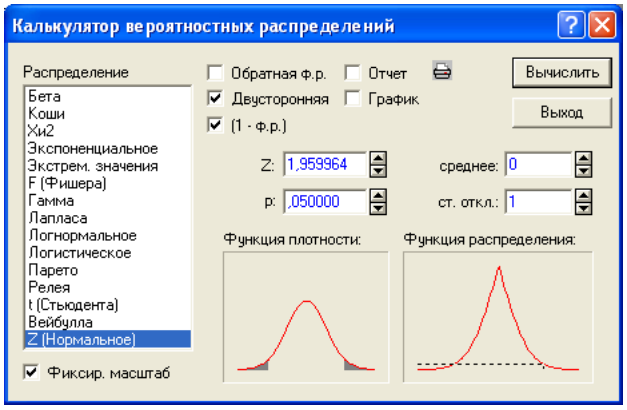


Рис. 2.2. Вычисление стандартизованного отклонения по заданной вероятности

Вычисление вероятностей для нормированных отклонений различных статистических показателей относится к разделу математической статистики, посвященному статистической проверке гипотез (он рассматривается в гл. 3).

2.2. ОСНОВНЫЕ ПОНЯТИЯ ВЫБОРОЧНОГО МЕТОДА

Множество всех единиц совокупности, подлежащей статистическому изучению, носит название **генеральной совокупности**. Многие задачи статистического анализа связаны с описанием больших совокупностей объектов. Зачастую на практике по тем или иным причинам невозможно рассмотреть все элементы таких совокупностей. В этом случае ограничиваются изучением лишь некоторой части генеральной совокупности. Эта часть называется выборочной совокупностью или **выборкой**. Полученные при ее изучении результаты стремятся распространить на всю генеральную совокупность. Разумеется, для этого выборка должна быть не любой произвольной частью генеральной совокупности, а такой ее частью, которая достаточно правильно отражает основные параметры этой совокупности. Таким образом, выборка, результаты изучения которой можно обобщить на всю совокупность, должна быть, как говорят, **репрезентативной** (представительной).

Каким образом можно добиться репрезентативности выборки, т.е. того, чтобы она правильно отражала основные свойства, присущие генеральной совокупности? Ответ кажется, на первый взгляд, довольно парадоксальным:

выборка должна быть случайной. Что это значит? Случайность никоим образом не отождествляется здесь со стихийностью или произвольностью отбора – напротив, случайность означает то, что все объекты генеральной совокупности должны иметь равные шансы попасть в выборку

На практике выборочный метод и соответствующая терминология знакомы многим по результатам социологических опросов, данным изучения общественного мнения, которые публикуются в прессе или предлагаются в аналитических телепередачах. Здесь налицо ситуация, когда сплошное изучение генеральной совокупности практически невозможно, и исследователи формируют выборки, пользуясь специальными методиками.

Наиболее простым является случайный отбор, например, при помощи обычной жеребьевки (для небольших совокупностей) или с использованием таблиц случайных чисел. Для более обширных, но достаточно однородных совокупностей используется механический отбор (применявшийся еще в земской статистике). Для неоднородных совокупностей с определенной структурой чаще применяется типический отбор. Существуют и другие методы, в том числе – комбинации разных способов отбора на нескольких этапах построения выборочной совокупности [1]

Не касаясь здесь проблемы т.н. "естественных выборок", подчеркнем, что знание основных способов формирования репрезентативной выборки может помочь историку не только проводить отбор объектов для выборочного изучения, если объем генеральной совокупности слишком велик (например, при анализе первичных данных переписей), но и правильно оценивать данные источников, которые уже представляют собой результаты выборочного изучения.

2.3. ОШИБКИ ВЫБОРКИ

Никакая, даже самым тщательным образом сформированная выборка, не может дать точного знания о генеральной совокупности. Таким образом, в выборочных результатах всегда присутствуют ошибки. Эти ошибки можно разделить на два класса: случайные и систематические. К первым относятся случайные отклонения выборочных характеристик от генеральных, обусловленные самой природой выборочного метода. Величина случайной ошибки поддается вычислению (оценке). Систематические ошибки, наоборот, не носят случайного характера; они связаны с отклонением структуры выборки от реальной структуры генеральной совокупности. Систематические ошибки появляются тогда, когда нарушается основное правило слу-

[1] Количественные методы в исторических исследованиях / Под ред. И.Д. Ковальченко. М., 1984. С. 104–108.

чайного отбора – обеспечение для всех объектов равных шансов попасть в выборку. Ошибки этого рода статистика не может оценивать.

Основными источниками систематических ошибок являются: а) неадекватность сформированной выборки задачам исследования; б) незнание характера распределения в генеральной совокупности и, как следствие, нарушение в выборке структуры генеральной совокупности; в) сознательный отбор наиболее удобных и выигрышных элементов генеральной совокупности.

Рассмотрим пример. Данные промышленных переписей 1900 и 1908 гг. по предприятиям Закавказья (файл Industry.sta) дают среднее число рабочих на предприятии, равное 77 чел. (по всей генеральной совокупности из 1060 предприятий). Случайный отбор 5% объектов (53 предприятия) дает среднее число рабочих, равное 81 чел. [1] Ошибка выборки, очевидно связанная с тем, что не все 1060 предприятий попали в выборку, равна разности между этими средними – генеральным ($\overline{X}_{г.с.}$) и выборочным ($\overline{x}$). Если сформировать другую выборку того же объема из нашей генеральной совокупности, она даст другую величину ошибки и т.д. Оказывается, что все эти выборочные средние при достаточно больших выборках распределены **нормально** вокруг генеральной средней при достаточно большом числе повторений выборки из генеральной совокупности одного и того числа объектов.

При этом неизбежный разброс выборочных средних вокруг генеральной средней (т.е. стандартное отклонение выборочных средних) называется **стандартной ошибкой выборки** μ, которая выражается формулой $\mu = \sigma / \sqrt{n}$ (σ – среднее квадратическое отклонение [2], n – объем выборки) [3].

---

[1] Этой выборке соответствует файл Sample.sta, которым и можно воспользоваться для изучения параметров выборочной совокупности. Файл Sample.sta получен из файла Industry.sta в результате механической выборки – отбиралось 5% предприятий, т.е. каждое двадцатое, начиная с первого (1-е, 21-е, 41-е и т.д.). Методы, с помощью которых можно построить выборку, будут рассмотрены в одном из следующих разделов данной главы.

[2] Вообще говоря, в этой формуле должна стоять величина среднего квадратического отклонения в генеральной совокупности, но на практике ее заменяют аналогичной выборочной характеристикой.

[3] Точнее, данная формула дает стандартную ошибку выборки в случае т.н. повторного отбора, когда каждый из отобранных объектов возвращается в генеральную совокупность и, следовательно, может быть отобран не один раз (именно в этом случае шансы для всех объектов попасть в выборку остаются равными и постоянными в течение всего отбора). Для т.н. метода бесповторного отбора, когда число объектов в генеральной совокупности уменьшается с каждым отобранным объектом, величина μ будет несколько меньше.

Это значит, что с определенной долей уверенности можно говорить, что большинство выборочных средних должно находиться в интервале $\overline{X}_{г.с.} \pm \mu$. Видно, что стандартная ошибка выборки тем меньше, чем меньше величина σ (которая характеризует разброс значений признака) и чем больше объем выборки *n*.

В нашем примере величина σ в генеральной совокупности известна точно и равна 187 чел.; из приведенной выше формулы легко подсчитать, что $\mu \approx 26$. Это значит, что большинство выборочных средних должно находиться в интервале 77±26 (т.е. от 52 до 103). Для нашей выборки число 81 действительно попало в этот интервал.

Однако надо пояснить, что значит "большинство". Для нормального распределения (а распределение выборочных средних как раз и является нормальным) известно, какая часть совокупности попадает в любой интервал вокруг среднего значения. В частности, две трети (приблизительно 67%) всех выборочных средних попадут в интервал $\overline{X}_{г.с.} \pm \mu$; 95% – в интервал $\overline{X}_{г.с.} \pm 2\mu$; 99,7% – в интервал $\overline{X}_{г.с.} \pm 3\mu$.

2.4. ТОЧНОСТЬ И НАДЕЖНОСТЬ ВЫБОРОЧНОГО МЕТОДА. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ

На практике проблема заключается, однако, в том, что характеристики генеральной совокупности нам не известны, а выборка делается именно с целью их оценки. Поэтому вместо величины $\overline{X}_{г.с.}$ известна $\bar{x}$ для выборки, и σ тоже считается по выборке. Значит, если мы будем делать выборки одного и того же объема *n* из генеральной совокупности, то в 68% случаев на интервале $\bar{x} \pm \mu$ будет находится значение $\overline{X}_{г.с.}$ (оно же в 95% случаев будет находится на интервале $\bar{x} \pm 2\mu$ и в 99,7% случаев – на интервале $\bar{x} \pm 3\mu$). Поскольку реально делается только одна выборка, то формулируется это утверждение в терминах вероятности: с вероятностью 68% среднее значение признака в генеральной совокупности заключено в интервале $\bar{x} \pm \mu$ (с вероятностью 95% – в интервале $\bar{x} \pm 2\mu$ и т.д.). В этом и состоит *статистическое оценивание*: по выборочным данным с определенной долей уверенности указать интервал, в котором находится значение какого-либо параметра для всей генеральной совокупности. Видно, что степень уверенности зависит от величины интервала. В примере со средним значением интервал $\bar{x} \pm t\mu$ при $t = 1$ дает уверенность 68%, а вдвое больший интервал при $t = 2$ – уверенность 95%.

Интервал $\bar{x} \pm t\mu$ называется **доверительным интервалом**, соответствующая вероятность – **доверительной вероятностью *P*** (надежностью),

величина $\Delta = t\mu$, которая задает длину доверительного интервала, называется **предельной ошибкой выборки** или точностью оценки (коэффициент t, очевидно, показывает, во сколько раз предельная ошибка превышает среднюю ошибку выборки μ).

Легко догадаться, что точность и надежность оценки параметров генеральной совокупности по выборке находятся в обратной зависимости: чем больше точность (т.е. чем меньше предельная ошибка и, соответственно, чем уже доверительный интервал), тем меньше надежность такой оценки (степень уверенности). И наоборот – чем ниже точность оценки, тем выше ее надежность. Часто доверительный интервал строят для надежности 95%, соответственно предельная ошибка выборки обычно равна удвоенной средней ошибке μ.

Пример 2.2. Построим доверительный интервал для числа рабочих на одном предприятии, пользуясь выборочными данными из таблицы Sample.sta.

На рис. 2.3 показана вкладка **Дополнительно** диалогового окна **Описательной статистики.** Если вы хотите увидеть доверительный интервал для среднего арифметического значения, то к стандартному набору выборочных характеристик надо добавить флажок **Доверительный интервал среднего**.

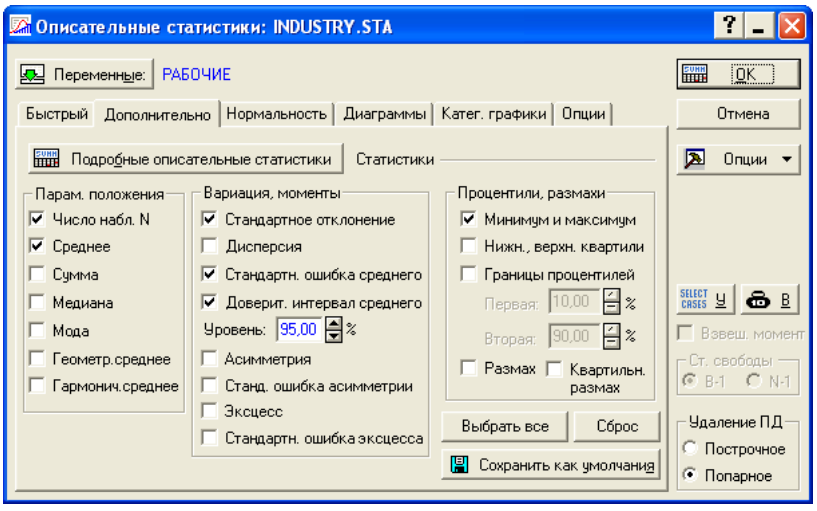


Рис. 2.3. Задание доверительных интервалов в диалоговом окне дескриптивной статистики

Если же вам интересно также увидеть и величину ошибки выборки (т.е. μ), то надо включить также флажок **Стандартная ошибка среднего** (см. рис. 2.3).

На рис. 2.4 можно видеть результат статистического оценивания среднего значения признака "число рабочих" в генеральной совокупности по нашей выборке: доверительным интервалом для среднего арифметического значения признака в генеральной совокупности является интервал [32, 131] (заметьте, что известное нам $\overline{X}_{г.с.}$, равное 77 чел., находится на этом интервале).

	Описательные статистики (Sample.sta)							
Переменная	N набл.	Среднее	Доверит. -95,000%	Доверит. +95,000%	Минимум	Максимум	Стд.откл.	**Станд. Ошибка**
РАБОЧИЕ	53	81	32	131	0,00	1115	179,3	24,6

Рис. 2.4. Результаты статистического оценивания

Попробуйте самостоятельно менять уровень доверительной вероятности (в поле **Уровень** на рис. 2.3) и вы убедитесь, что с ростом этой величины доверительный интервал расширяется (а точность, естественно, падает) и наоборот.

2.5. ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ

Напомним, что ошибка выборки μ зависит от объема выборки n: чем больше n, тем меньше μ.

На этом строится тактика формирования выборочной совокупности: если исследователь ставит перед собой задачу создания выборки, которая давала бы оценку для генеральной совокупности с определенной точностью Δ и надежностью (уверенностью) P, то количество объектов, которое надо отобрать (методом случайного отбора) для изучения определяется по формуле

$$n = t^2\sigma^2 / \Delta^2 ,$$

где t определяется по специальным таблицам нормального распределения (см. раздел 2.1 данной главы) в зависимости от P [1], σ – среднее квадратическое отклонение признака в генеральной совокупности (или в выборке). Проблема состоит в том, что пока выборка не сделана, величина σ не

[1] Заметим, что в таблицах для нормального распределения величина t обозначается как z, а вместо P (доверительной вероятности) табулированы значения 1–P (обозначаются как p).

известна, поэтому для окончательного формирования выборки приходится делать предварительную или пробную выборку для определения σ.

Пример 2.3. Определить, сколько предприятий Закавказья следует отобрать, чтобы определить среднее число рабочих на предприятии с надежностью 90% и точностью 20 рабочих (в качестве σ использовать величину 187 чел.).

Обратимся к разделу **Вероятностный калькулятор | Z (Нормальное)** модуля **Основные статистики и таблицы**. В поле *p* зададим величину 0,1 (т.к. $p = 1–P$), включим флажки **(1 – ф.р.)** и **Двусторонняя** и щелкнем по графической кнопке **Вычислить**. Результатом будет величина *t* (здесь она обозначается *Z*), равная 1,6449 (см. рис. 2.5). Подставив в формулу для *n* все необходимые значения, получим: $n = (1{,}65)^2 (187)^2 / (20)^2 = 144$.

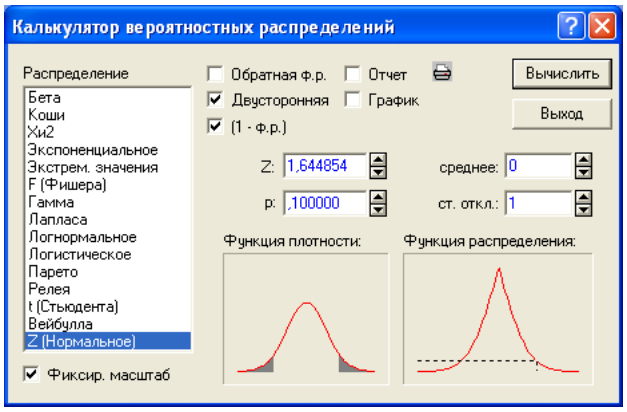


Рис. 2.5. Определение величины t для заданной доверительной вероятности

Таким образом, 144 предприятия в выборке дадут ответ на вопрос о среднем числе рабочих на одно предприятие в генеральной совокупности с точностью до 20 человек и надежностью 90%.

Напомним, что выборка – это часть генеральной совокупности, но не любая часть совокупности является выборкой со статистической точки зрения. Так, если необходимо отобрать определенную часть объектов (например, все предприятия, владельцами которых были купцы), это можно сделать, формируя подмножество объектов совокупности (*subset*). Но это подмножество, разумеется, не является выборкой, так как не соответствует данному выше определению. Как же построить "правильную" выборку? Рассмотрим возможности пакета STATISTICA реализовать выборочный

метод в "чистом виде", т.е. проводить отбор объектов, пользуясь т.н. генератором случайных чисел.

Пример 2.4. Сделаем пятипроцентную выборку, взяв в качестве генеральной совокупности файл Industry.sta (1060 объектов); результат сохраним в файле Sample1.sta.

В разделе **Данные** главного меню программы выберем возможность направленного или случайного отбора объектов – **Подмножество / Случайный выбор**. Открывается диалоговое окно, в котором надо выбрать переключатель **Простая случайная выборка** и в поле **Прибл. %** задать значение 5 (см. рис. 2.6). Щелчок по графической кнопке OK дает случайную выборку нужного объема, которую можно сохранить как новый файл в формате программы STATISTICA.

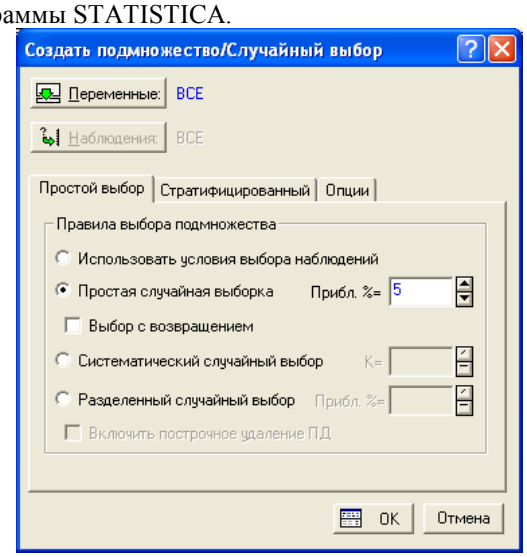


Рис. 2.6. Формирование случайной пятипроцентной выборки

2.6. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ДОЛИ КАЧЕСТВЕННОГО ПРИЗНАКА

Если выборочный метод используется для работы с неколичественными данными, то роль среднего арифметического значения в совокупности играет доля или частота q признака. Доля вычисляется как отношение числа объектов, обладающих данным признаком (n_0), к числу объектов во всей совокупности: $q = \frac{n_0}{n}$. Роль меры разброса играет величина $\sigma = \sqrt{q(1-q)}$.

В этом случае стандартная ошибка выборки μ вычисляется по формуле:

$$\mu = \sqrt{\frac{q(1-q)}{n}} .$$

Пример 2.5. Попробуем найти, пользуясь этой формулой, стандартную ошибку 5% выборки из генеральной совокупности промышленных предприятий Закавказья (уже упомянутые выше 53 предприятия из общего числа 1060, файл Sample.sta) при определении доли частных предприятий. По данным выборки из 53 предприятий оказалось 38 частных, т.е. доля частных равна 0,72 (или 72%). Стандартная ошибка для выборок объема 53 по приведенной выше формуле равна ≈ 0,06. Как известно, стандартная ошибка выборки используется для построения доверительных интервалов: так, с вероятностью 95% можно утверждать, что неизвестное значение доли частных предприятий в генеральной совокупности лежит в границах 0,72±2(0,06), т.е. от 0,60 до 0,84 или от 60 до 84%.

Пользуясь формулой для стандартной ошибки выборки при вычислении доли качественного признака, можно получить и формулу объема выборки для определения неизвестного значения доли качественного признака в генеральной совокупности с заданными точностью Δ и надежностью *P*. Эта формула имеет вид:

$$n = \frac{t^2 q(1-q)}{\Delta^2}$$

Величина *t*, как и раньше, вычисляется для каждого значения вероятности *P* по таблице нормального распределения (в частности, напомним, что для *P* = 95% величина *t*≈2).

Пример 2.6. Определим объем выборки, которая с надежностью 95% должна оценить долю частных предприятий, и точность оценки (Δ) должна быть равной 0,05 (5%).

Пользуясь приведенной выше формулой для *n*, получим $n = (2)^2(0{,}72)(1–0{,}72)/(0{,}05)^2 \approx 161$ предприятие.

ВОПРОСЫ

1. Когда в историческом исследовании возникает проблема выборки?
2. Что такое репрезентативность?
3. "Естественная" выборка
4. Случайные и систематические ошибки
5. Может ли быть абсолютно точным результат выборочного исследования?
6. Механизмы случайного отбора

7. В чем отличие бесповторного отбора от повторного?
8. Типы выборок
9. Верно ли, что выборка дает тем лучший результат, чем больше ее объем?
10. Из одной генеральной совокупности сделана 5% выборка, а из другой – 10% выборка. Какая их них более точно отражает "свою" генеральную совокупность?
11. Что такое доверительный интервал?
12. Что такое уровень доверия?
13. Верно ли, что увеличение точности результата выборочного исследования связано с уменьшением надежности?
14. Последовательность действий при использовании выборочного метода
15. Зачем нужны пробные выборки?

ЗАДАНИЯ

1. Используя файл Duma.sta, рассчитать величину доверительного интервала для среднего возраста депутатов I Государственной думы, исходя из предположения, что мы имеем сведения только для:
   а) 200 человек;
   б) 300 человек.
2. Используя файл Industry.sta, построить доверительные интервалы:
   а) для среднего числа рабочих по всем отраслям;
   б) для среднего числа рабочих по всем губерниям;
   в) для средней мощности двигателей по всем отраслям.
3. Для выборки объемом 256 студентов из общего числа студентов МГУ определен их средний возраст, равный 23 года. Построить доверительный интервал для возраста студентов МГУ с надежностью 99,7%, учитывая что $\sigma_{г.с.}$ = 6 лет.
4. Определить объем выборки, гарантируя точность 0,1 с вероятностью 99,7% из генеральной совокупности объемом 1000 единиц ($\sigma^2_{г.с.}$ = 5). Сравнить с объемом повторной выборки.
5. Выборочное обследование 900 человек показало, что 18 чел. не информированы о крупном событии в стране. С вероятностью 0,95 (95%) найти доверительный интервал доли таких лиц в стране.
6. Сколько надо отобрать хозяйств из общей совокупности в 900 хозяйств, чтобы определить долю хозяйств с применением наемного труда с точностью до 5% при уровне надежности 95%?
7. Построить доверительный интервал для оценки стажа работы в генеральной совокупности объемом 1000 человек по выборке 800 чел. Сте-

пень надежности взять равной 99,7%. По выборке получены следующие результаты: $\bar{X}$ = 15 лет; σ = 3 года.

8. Из генеральной совокупности 1000 тыс. крестьянских хозяйств отобраны 100 хозяйств. Оказалось, что 10 хозяйств используют наемный труд. С надежностью 95% определить долю таких хозяйств в генеральной совокупности.
9. Определить объем выборки для определения доли студентов среди населения с точностью до 1% и надежностью 95% (в качестве q взять величину 2%).
10. Выборку какого объема надо взять, чтобы оценить долю лиц данной профессии в генеральной совокупности по данным выборки с точностью. до 2% и надежностью 95% (в качестве величины q принять значение 11%).
11. Выборку какого объема надо сделать для определения состава студентов по полу, чтобы точность (Δ) была равна 0,02, надежность (P) была равна 99,7%? (В качестве q берется значения 0,5.)
12. Из 900 дней 100 оказались облачными в данном районе. Каковы границы для процента облачности в этом районе?
13. Требуется с вероятностью 99,7% обеспечить такой объем выборки лиц трудоспособного возраста из генеральной совокупности, чтобы отклонение доли безработных в выборке от их доли в генеральной совокупности не превышало 0,01. (Известно, что доля безработных в генеральной совокупности не превышает 0,1.)
14. При обследовании 900 лиц трудоспособного возраста определен их средний возраст – 45 лет. Для надежности 95% найти доверительный интервал, в котором содержится генеральная средняя. (В качестве σ принять значение 10 лет.)
15. Из генеральной совокупности 250 рабочих взята выборка объемом 25 человек. Их средний заработок оказался равным 900 руб. (при σ^2 = 490 руб.) С вероятностью 95% определить, в каких пределах заключена средняя зарплата в генеральной совокупности.
16. Выборку какого объема надо взять для оценки среднего возраста студентов МГУ с точностью до 1 года и надежностью 99,7%, если пробные выборки дают значение σ = 10 лет.
17. В группе из 20 человек после коллоквиума у 6 человек посещаемость не изменилась, а у остальных – повысилась. Найти доверительный интервал (P = 0,95) для вероятности того, что коллоквиум улучшает посещаемость.

ГЛАВА 3

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

Одним из важных инструментов научного исследования является построение и проверка тех или иных гипотез, касающихся объекта изучения, его состава, структуры, связей. Иногда такие гипотезы не формируются в ходе исследования явным образом, однако все же присутствуют в неявной форме.

В исторической науке так же, как и в других областях знания, исследователь либо строит и пытается обосновать некую гипотезу, либо стремится опровергнуть другую гипотезу. Так, примерами могут служить гипотеза о росте реальной зарплаты в период капиталистической индустриализации, гипотеза о снижении экономического уровня крестьянского хозяйства в России к концу периода крепостничества и т.п. Это примеры научных гипотез, которые выдвигают историки. Теория же статистической проверки гипотез суживает и формализует общенаучное понятие гипотезы, что делает возможным применение этой теории в самых различных областях знания.

Задача этой главы несколько иная, чем у других глав. Дело в том, что статистическая проверка гипотез используется во многих других разделах (например, при анализе таблиц сопряженности, в корреляционном и регрессионном анализе и др.). Поэтому здесь будут рассмотрены теоретические основы методы и несколько наиболее важных приложений (сравнение средних значений и сравнение распределений). Остальные приложения будут вводиться в текст других глав как отдельные разделы (например, проверка статистической значимости коэффициента корреляции – в главе корреляционный анализ и т.д.).

3.1. ОСНОВНЫЕ ПОНЯТИЯ

Статистическая гипотеза. Понятие статистической гипотезы гораздо уже, чем просто научной гипотезы. Статистическими гипотезами называют различного рода предположения о свойствах генеральной совокупности (совокупностей), подтверждаемые или отвергаемые методами математической статистики на основе выборочных данных.

Таким образом, ясно, что статистические гипотезы, как и выборочное исследование, связаны с необходимостью делать выводы обо всем явлении, процессе на основе имеющихся данных об его части. Поскольку при статистической проверке гипотез мы не располагаем данными обо всей гене-

ральной совокупности, всегда есть риск совершить ошибку: отклонить верную гипотезу или принять неверную.

Поэтому статистическая проверка гипотез является по своей природе вероятностной. Однако на практике чрезвычайно важно, что теория статистической проверки гипотез позволяет оценить вероятность совершить ошибку и дает, таким образом, оценку надежности получаемых выводов.

Хотя из приведенного определения видно, что не всякая гипотеза в обычном смысле является статистической, многие научные гипотезы можно свести к целому ряду статистических гипотез (например, гипотезу о росте реальной зарплаты рабочих свести к гипотезе о положительном значении коэффициента линейного тренда в динамическом ряду среднегодовых уровней зарплаты рабочих той или иной отрасли промышленности).

Итак, проверка любой статистической гипотезы начинается, естественно, с ее формулирования в соответствии с приведенным определением.

Пример 3.1. Пусть имеются данные 5%-ной случайной выборки промышленных предприятий Закавказья, всего 53 предприятия. Напомним, что создание этой выборки описано в главе 2, а файл называется Sample.sta. По этим данным установлено, что среднее число рабочих на одном предприятии равно 81 чел. Нам неизвестно среднее значение числа рабочих на одно предприятие по всей переписи, но положим, что в работах некоторого автора дается в качестве оценки этого значения величина, равная 50 чел.

Таким образом, нашей гипотезой является гипотеза о том, что среднее число рабочих на одном предприятии в начале XX в. равнялось 50 чел. Эта гипотеза может быть проверена и называется она *испытуемой или нулевой гипотезой* и обозначается в статистике H_0. Как же проверить эту гипотезу?

Статистический критерий и статистическая характеристика. Центральным понятием в теории статистической проверки гипотез является понятие статистического критерия. Статистическим критерием называется совокупность строго определенных правил, указывающих, при каких результатах выборочного исследования испытуемая гипотеза отклоняется, а при каких – считается допустимой.

Вернемся к нашему примеру и выясним смысл введенных понятий.

Как известно, выборочные средние отличаются от генеральной средней, но эти отличия характеризуются тем, что чаще встречаются выборочные средние, близкие к генеральной (см. главу 2). Таким образом, наличие выборочной средней, которая сильно отклоняется от предполагаемой генеральной, свидетельствует, скорее всего, о неверности испытуемой гипотезы. Действительно, чем дальше от 50 чел. находится выборочное значение числа рабочих, тем больше у нас оснований сомневаться в справедливости испытуемой гипотезы.

Значит, желательно, чтобы при больших отклонениях выборочной средней от значения 50 критерий отвергал испытуемую гипотезу, а при небольших отклонениях признавал ее допустимой. Остается решить, какие же отклонения считать большими, а какие – небольшими.

Вспомним, что для больших выборок известно распределение выборочных средних, т. е. вероятность появления каждого конкретного значения $\bar{x}$ (выборочного среднего), если известно значение $\bar{X}_{\text{г.с.}}$ Оказывается, распределение выборочных средних нормально, и, следовательно, чем больше значение отличается от 50, тем меньше вероятность его появления в конкретной выборке, причем эта вероятность точно известна. Значит, при получении маловероятного значения $\bar{x}$ критерий будет отклонять испытуемую гипотезу.

В основе такого вывода (и в основе всей теории проверки статистических гипотез) лежит так называемый *принцип практической невозможности*, который гласит, что маловероятное событие практически невозможно в единичном случае, каким является наша выборка. К сожалению, нельзя указать годную для всех случаев границу, такую, что событиями с вероятностью, меньшей этой границы, мы пренебрегаем, считая их невозможными. Дозволенная степень риска, который связан с пренебрежением событиями с малой вероятностью, зависит от различного рода обстоятельств и связана с практической важностью следствий, вытекающих из наступления таких событий.

Итак, в соответствии с принципом практической невозможности статистической характеристикой нулевой гипотезы H_o может служить величина отклонения выборочной средней от предполагаемого генерального значения. По таблице нормального распределения можно найти вероятность появления отклонений, превышающих по абсолютному значению данное отклонение ($\bar{x}$ –50). Для удобства пользования этой таблицей в качестве статистической характеристики берется не само отклонение $\bar{x}$ –50, а так называемое нормированное отклонение t:

$t = (\bar{x} - \bar{X}_{\text{г.с.}})/\mu$

где $\mu = \sigma / \sqrt{n}$ – стандартная ошибка выборки.

Величина t имеет нулевое среднее значение, а ее стандартное отклонение равно единице.

В данном случае известно, что μ = 26. Тогда фактическим значением статистической характеристики (t_ϕ) в нашем примере является величина (81–50)/26≈1,2.

Пользуясь таблицей нормального распределения (см. раздел 2.1 гл. 2), найдем вероятность того, что абсолютное значение *t* превзойдет величину 1,2. Эта вероятность приближенно равна 0,23 (см. рис. 3.1).

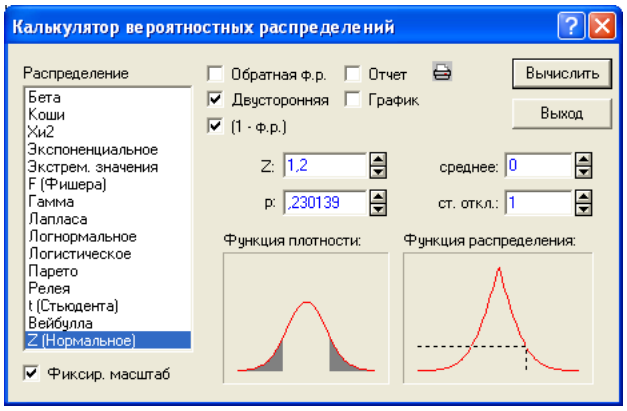


Рис. 3.1. Вероятность нормированного отклонения, превышающего 1,2

Тогда вероятность попадания t в интервал [–1,2; 1,2] приближенно равна 1–0,23 или 0,77. Напомним, как следует интерпретировать полученное значение. Вероятность 0,77 означает, что если бы мы произвели 100 выборок из нашей генеральной совокупности, то в 77 из них отклонение $\overline{X}$ от 50 не превысило бы величину 31 (разность между средним значением числа рабочих в выборке, т.е. 81 чел. и оценкой среднего числа рабочих в генеральной совокупности, т.е. 50 чел.) и в 23 из них (23 = 100 – 77) оно могло бы превзойти 31 (по абсолютной величине), разумеется, при условии, что среднее число рабочих в генеральной совокупности действительно равно 50. (Графически это распределение показано на рис. 3.1. и в более крупном масштабе – на рис. 3.2). Таким образом, площадь под нормальной кривой между значениями 19 и 81 составляет 0,77 всей площади под нормальной кривой.

Осталось выяснить, можно ли считать вероятность 0,23 настолько малой, чтобы пренебречь площадью справа от точки 81 и слева от точки 19 по сравнению с площадью под всей кривой (как известно, она равна 1). Если считать вероятность 0,23 достаточно малой, то значения выборочной средней, равные 81 или большие, являются практически невозможными, и гипотезу H_0 придется отклонить. Однако если считать, что эта вероятность не слишком мала, значение 81 не противоречит нашей гипотезе.

Тем самым очевидно, что наши выводы существенно зависят от того значения вероятности, при котором $\overline{X}$ считается практически невозможным для данной испытуемой гипотезы.

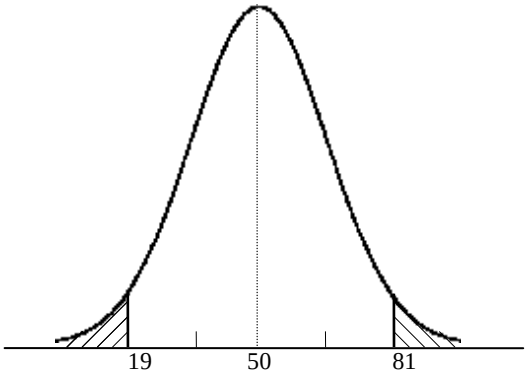


Рис. 3.2. Нормальное распределение выборочных средних при генеральной средней, равной 50 чел.

Уровень значимости и критическая область. Значение вероятности, начиная с которого событие считается практически невозможным, называется **уровнем значимости** критерия.

Уровень значимости (в программе STATISTICA он обозначается p – см. раздел 2.1 главы 2) обычно полагается равным 0,1; 0,05; 0,025 или 0,01. (Например, уровень значимости 0,01 или 1% означает, что мы считаем практически невозможными события, вероятность наступления которых не более 1%, т.е. события, которые могут произойти не более чем в 1 случае из 100.)

Рассмотрим на нашем примере, какие значения статистической характеристики t соответствуют уровню значимости 0,1. По таблице нормального распределения **Z (Нормальное)** из раздела **Вероятностный калькулятор** модуля **Основные статистики и таблицы** для уровня значимости 0,1 находим t = 1,64. Это значит, что с вероятностью 0,1 значение выборочного среднего может отклониться от предполагаемого генерального значения больше, чем на $t\mu$ = 1,64·26 = 43, т.е. практически невозможными считаются значения $\overline{X}$, большие 93·(т.е. 50+43) и меньшие 7 (т.е. 50–43).

Вспомним, что в нашем случае было получено фактическое значение среднего числа рабочих ($\overline{X}_{\text{ф}}$), равное 81, которое, следовательно, не является практически невозможным, и при таком уровне значимости (0,1) испытуемая гипотеза не отклоняется. Следует подчеркнуть, что уровень значимости выбирается исследователем в зависимости от конкретной задачи.

Каждому уровню значимости соответствует **критическое значение** статистической характеристики, которое делит все множество значений характеристики на две области: **допустимых значений** и **критическую**. Критической областью испытуемой гипотезы являются все значения статистической характеристики, вероятность появления которых меньше выбранного уровня значимости. Все остальные значения статистической характеристики образуют область допустимых значений.

Между статистической характеристикой, выбранным уровнем значимости и критической областью существует следующее соотношение: вероятность того, что статистическая характеристика попадет в критическую область, если верна испытуемая гипотеза, равна выбранному уровню значимости, т.е. критическая область содержит именно те значения статистической характеристики, которые мы считаем практически невозможными.

В нашем примере критической областью ($|t| \geq t_{кр}$) являются те значения t, вероятность появления которых меньше уровня значимости, а областью допустимых значений ($|t| < t_{кр}$) – те, вероятность появления которых больше уровня значимости.

Вновь поясним изложенное на графике (рис. 3.3). Заштрихована критическая область, соответствующая уровню значимости 0,1. Видно, что фактическое значение $t_{ф}$, которое равно 1,2, попадает в область допустимых значений, т.е. не является практически невозможным, если $\overline{X}_{г.с.}$ = 50; значит, испытуемая гипотеза не отклоняется.

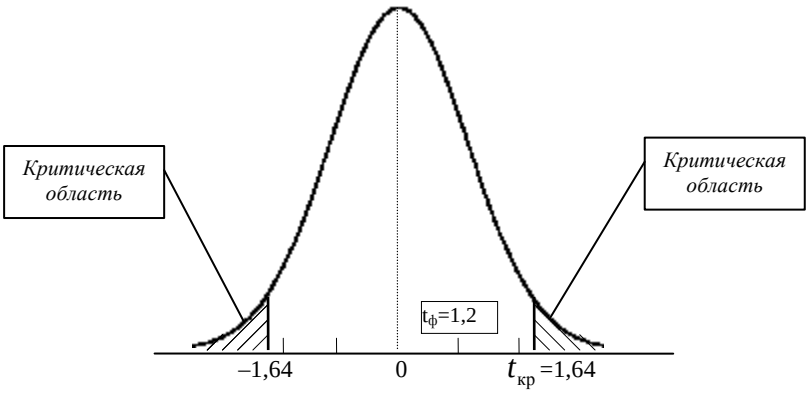


Рис. 3.3. Критическая область для уровня значимости 0,1

Из приведенного примера видно, что статистическая гипотеза проверяется в такой последовательности:

1) формулируется испытуемая гипотеза (в нашем примере $\overline{X}_{\text{г.с.}}$ = 50);
2) определяется статистическая характеристика этой гипотезы $\left(t = \frac{x - \overline{X}_{\text{г.с.}}}{\mu}\right)$;
3) рассчитывается ее фактическое значение ($t_{\text{ф}}$ = 1,2);
4) выбирается уровень значимости (0,1);
5) определяется соответствующая ему критическая область ($|t| \geq 1{,}64$).
6) по таблице соответствующего (в нашем случае нормального) распределения рассчитывается вероятность того, что значение статистической характеристики превысит ее фактическое значение $t_{\text{ф}}$. Если величина вероятности (в нашем примере она равна 0,23) меньше уровня значимости, это соответствует ситуации, когда фактическое значение статистической характеристики попадает в критическую область. В таком случае критерий отклоняет испытуемую гипотезу. Если же величина вероятности для данного критического значения окажется больше уровня значимости (а в нашем случае 0,23>0,1), это соответствует попаданию в область допустимых значений и гипотеза признается допустимой (т.е. $|t| < t_{\text{кр}}$, и гипотеза не отклоняется на уровне значимости 0,1).

Если гипотеза не отклоняется, это еще не значит, что она верна: дальнейшие исследования могут привести к отклонению гипотезы, но наш критерий не дает оснований отклонить ее.

Очевидно, попадание или непопадание точки $t_{\text{ф}}$ в критическую область зависит от размеров этой области, а они, в свою очередь, зависят от уровня значимости. Например, если увеличить уровень значимости и вместо 0,1 взять 0,32, что соответствует (***проверьте самостоятельно!***) значению $t_{\text{кр}}$ = 1, то критическая область расширится и $t_{\text{ф}}$, вероятность которого (0,23) меньше нового уровня значимости, попадет в нее, что означает отклонение испытуемой гипотезы (см. рис. 3.4). Какой же результат считать более обоснованным и не противоречат ли они друг другу?

Ошибка первого рода. Вспомним смысл уровня значимости. Если уровень значимости = 0,32, это говорит о том, что в 32 выборке из 100 могут все же получиться отклонения выборочного среднего от 50 чел., превышающие критическое значение, и **при справедливости** испытуемой гипотезы (за счет случайностей выборки). Однако при этом $t_{\text{ф}}$ попадет в критическую область, и критерий **отклонит** гипотезу, тогда как она **верна**. Та-

ким образом, попадание в критическую область не обязательно связано с отклонением действительно неверной гипотезы – в 32 случаях из 100 это означает ошибочное отклонение верной гипотезы, и, значит, уровень значимости – это риск совершить так называемую **ошибку первого рода**. Величина *P*, которая дополняет уровень значимости до 1, называется уровнем доверия или доверительной вероятностью и представляет собой вероятность правильного результата проверки гипотезы.

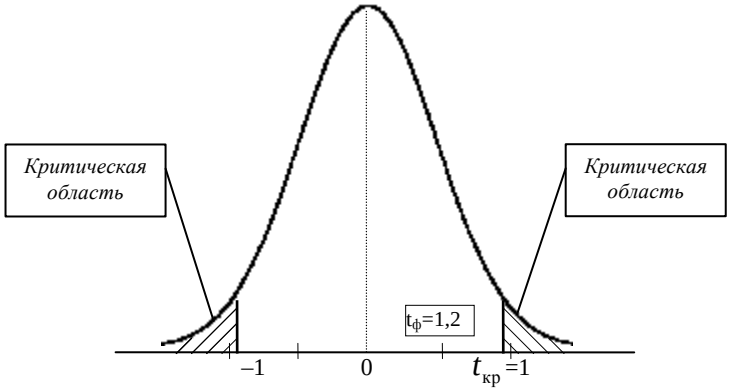


Рис. 3.4. Критическая область для уровня значимости 0,32

Итак, ошибкой первого рода называется вероятность отклонить испытуемую гипотезу, когда она верна (т.е. ошибка первого рода совпадает с уровнем значимости).

Мы убедились, что для уровня значимости 0,1 риск совершить ошибку первого рода меньше, чем для уровня значимости 0,32. На первый взгляд, надо стремиться уменьшать уровень значимости, уменьшая тем самым вероятность ошибки первого рода. Но из сравнения рис. 3.2 и рис. 3.3 видно, что с уменьшением уровня значимости критическая область сужается и, следовательно, расширяется область допустимых значений, т.е. становятся вполне допустимыми значения $\overline{X}$, даже далекие от 50, а такие значения, возможно, лучше соответствуют не нулевой, а каким-то другим гипотезам.

Следовательно, слишком широкая область допустимых значений способствует неотклонению испытуемой гипотезы, когда она неверна, т.е. все реже неотклонение будет эквивалентно принятию гипотезы. Например, если вместо гипотезы, что среднее число рабочих в генеральной совокупности равно 50, испытать гипотезу, что генеральное среднее равно 40, получим $t_ф$ = (81–40)/26 ≈ 1,6, причем снова $|t| < t_{кр}$ = 1,64 при уровне значимости 0,1. Таким образом, на уровне значимости на одном и том же уровне

значимости 0,1 и эта гипотеза тоже не отклоняется, значит, наши данные не противоречат не только первой, но и второй гипотезе (а также и многим другим, например, гипотезам, дающим значения $\overline{X}_{\text{г.с.}}$ в интервале 40–50 чел.).

Ошибка второго рода. С уменьшением уровня значимости повышается вероятность ошибочного принятия неверной гипотезы, точнее – неотклонения испытуемой гипотезы, когда на самом деле она неверна.

Вообще говоря, при испытании гипотезы H_0 можно получить четыре возможных результата:

I. Гипотеза H_0 верна:

1) она не отклоняется (правильный результат),

2) она отклоняется (ошибка первого рода).

II. Гипотеза H_0 неверна:

3) она отклоняется (правильный результат),

4) она не отклоняется (ошибка второго рода).

Таким образом, когда нулевая гипотеза не отклоняется требуется проверить, не соответствуют ли данные выборки другой, конкурирующей с H_0, гипотезе, т.е. требуется выдвинуть *альтернативную* гипотезу. Альтернативная гипотеза может быть сформулирована по-разному. Как правило, выявление незначительных различий между гипотезами не имеет практического значения, поэтому в качестве альтернативной гипотезы обычно берут гипотезу, значимо отличающуюся от испытуемой, повышая тем самым *мощность* критерия.

Односторонняя и двусторонняя проверка. В заключение первого раздела необходимо подчеркнуть, что при изложении основных понятий и этапов проверки гипотезы мы пользовались так называемой **двусторонней проверкой** – нас одинаково интересовали как положительные, так и отрицательные отклонения $\overline{x}$ от $\overline{X}_{\text{г.с.}}$. Именно поэтому критическая область у нас состояла из двух частей. Однако возможны ситуации, когда нас интересуют только положительные или только отрицательные отклонения, (например, если среднее число рабочих на предприятии не может быть меньше 50 чел.). В таких случаях вместо двусторонней проводят **одностороннюю проверку**. При этом критическая область состоит из одного участка и соответственно имеется только одно критическое значение, несколько меньшее по абсолютной величине, чем $t_{\text{кр}}$ для двусторонней проверки (при одном и том же уровне значимости).

Значение $t_{\text{кр}}$ для односторонней проверки находится с помощью таблицы нормального распределения в уже знакомом нам разделе **Вероят-**

ностный калькулятор модуля **Основные статистики и таблицы** (только снимается флажок **Двусторонняя**). Например, для уровня значимости 0,1 критическим значением в случае односторонней проверки является $t_{кр}$ = 1,28. Однако если нет оснований считать направление отклонений вполне определенным, следует использовать, как более строгий, двусторонний критерий.

3.2. КРИТЕРИИ ДЛЯ СРЕДНИХ

3.2.1.Критерий для сравнения групповых средних

При изучении выборки можно сравнивать средние значения какого-либо параметра для разных групп внутри одной совокупности объектов. Как правило эти средние сравнивают с целью проверки гипотезы о том, что изучаемые группы объектов не различаются по данному параметру, а реальные расхождения в значениях средних объясняются просто случайностями выборок.

В этом случае испытуемую гипотезу можно сформулировать следующим образом: разные группы представляют собой выборки из одной и той же генеральной совокупности, т.е. различие между их средними случайно, поскольку генеральные средние в обоих случаях равны. В качестве статистической характеристики используется величина *t*, представляющая собой разность выборочных средних, деленную на усредненную стандартную ошибку среднего по обеим выборкам.

Фактическое значение статистической характеристики сравнивается с критическим значением, соответствующим выбранному уровню значимости. Если фактическое значение больше, чем критическое, испытуемая гипотеза отклоняется, т.е. различие между средними считается значимым (существенным).

Пример 3.2. В файле General.sta хранятся некоторые биографические данные о лицах, входивших в высший командный состав Советской армии в период Второй мировой войны. Среди переменных есть "год вступления в партию" и "социальное происхождение". Проверим зависимость вступления в партию от социального происхождения. Для этого сравним среднее значение переменной "год вступления в партию" для различных социальных групп. Обратимся к разделу **Т-критерий для независимых выборок** в модуле **Основная статистика / Таблицы.**

Обратите внимание, что в появившемся диалоговом окне необходимо указать следующее: название **группирующей** переменной (в данном примере – "социальное происхождение") и название **зависимой** от нее пере-

менной, для которой будут считаться средние значения по группам (в данном примере – "год вступления в партию"). Кроме того, надо выбрать какие-либо две интересующие нас социальные группы, чтобы указать их в "окошках" **Код для группы 1** и **Код для группы 2** диалогового окна (двойным щелчком в каждом из этих окошек можно получить список таких групп и выбрать нужные, например, "из крестьян" и "из рабочих" (см. рис. 3.5).

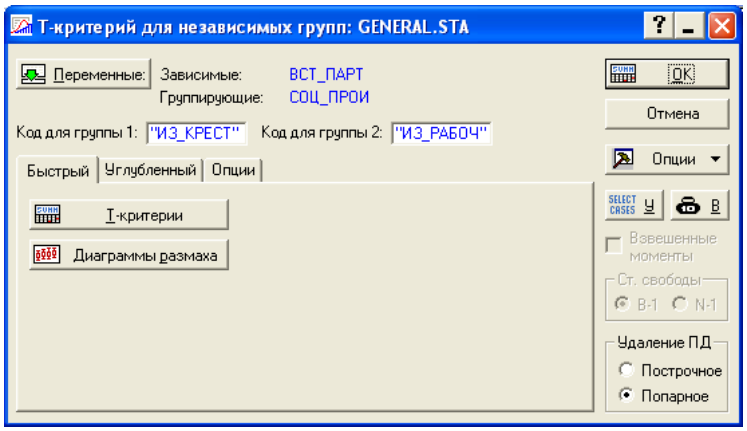


Рис. 3.5. Диалоговое окно теста на значимость различия групповых средних

Результаты выдаются в следующем виде:

Переменная	T-критерии; Группир.: СОЦ_ПРОИ (GENERAL.STA) Группа 1:ИЗ_КРЕСТ Группа 2:ИЗ_РАБОЧ Среднее ИЗ_КРЕСТ	Среднее ИЗ_РАБОЧ	t-знач.	ст.св.	p
ВСТ_ПАРТ	1923,4	1922,8	0,54	171	0,588339

Рис. 3.6. Фрагмент таблицы результатов при проведении теста на значимость различия групповых средних

Видно, что среднее значение года вступления в партию для обеих групп почти не различается, т.е. в этом отношении группы близки. Подтверждением этого вывода является значение статистической характеристики (*t-value*) и соответствующее ему значение вероятности (*p*). Величина *t* в таблице результатов невелика, а соответствующая этой величине вероятность почти достигает 0,6. Таким образом, различие между годом вступления в партию для выбранных нами социальных групп ("из рабочих" и "из крестьян") является статистически незначимым.

Специальная кнопка **Диаграммы размахат** позволяет увидеть графическую интерпретацию: показывает доверительные интервалы для оценки средних значений в генеральной совокупности (для доверительной вероятности 67% и 95%). Видно, что в случае статистически незначимых различий двух средних их доверительные интервалы пересекаются (см. рис. 3.7).

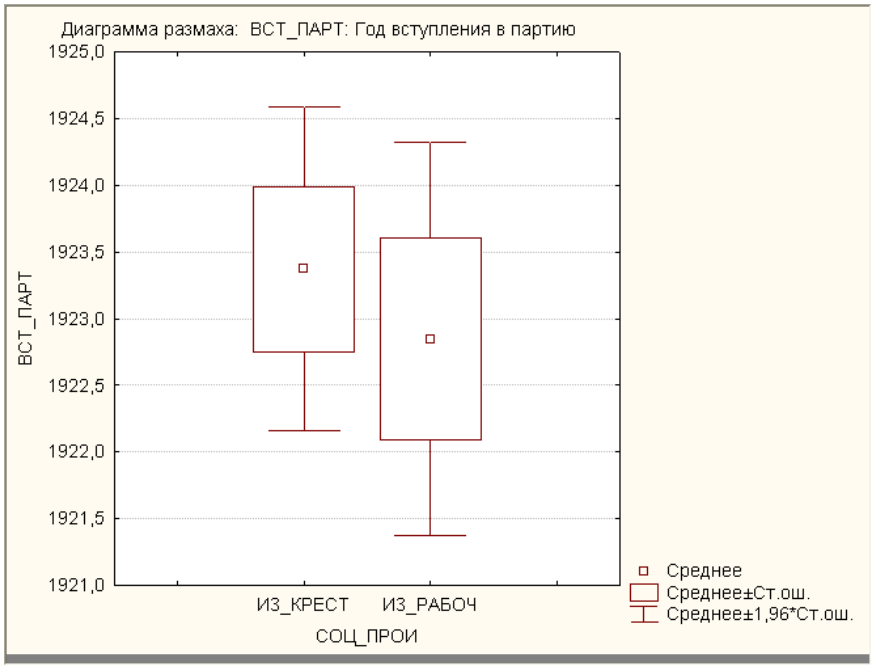


Рис. 3.7. Графическая интерпретация в случае статистически незначимых различий двух средних

Совсем другая картина получится, если сравнить две другие группы: "из рабочих" и "из служащих" (см. рис. 3.8).

	T-критерии; Группир.: СОЦ_ПРОИ (GENERAL.STA) Группа 1:ИЗ_СЛУЖА Группа 2:ИЗ_РАБОЧ				
Переменная	Среднее ИЗ_СЛУЖА	Среднее ИЗ_РАБОЧ	t-знач.	ст.св.	p
ВСТ_ПАРТ	1929,0	1922,8	4,6	148	0,00001

Рис. 3.8. Фрагмент таблицы результатов для другой пары групп

В этом случае величина *t* значительно больше, чем в предыдущем, а соответствующая ей вероятность *p* очень мала (практически равна нулю). Та-

ким образом, разница (а она составляет более шести лет) между средними значениями года вступления в партию для выходцев их рабочих и служащих является статистически значимой [1], т.е. первые раньше вступали в партию, чем вторые. Этот результат подтверждает и рис. 3.9, на котором доверительные интервалы не пересекаются.

Рис. 3.9. Графическая интерпретация в случае статистически незначимых различий двух средних

Иногда две группы, по которым проводится сравнение средних, формируются не по категориям группировочного признака (как в рассмотренном примере), а просто представляют собой две отдельные переменные в таблице исходных данных. Например, в одном столбце таблицы стоят значения года вступления в партию для тех, кто является по социальному происхождению рабочими, а в другом столбце – значения года вступления в партию для тех, кто является по социальному происхождению служащими.

В этом случае для сравнения групповых средних надо выбрать раздел **Т-критерий для независимых переменных** в модуле **Основные стати-**

[1] Заметьте, что все числа в таблице результатов на экране компьютера выделены красным цветом. Это значит, что результат проверки гипотезы является значимым на уровне 0,05 (или 5%), который в программе принят по умолчанию.

стики и таблицы и указать первую из этих переменных в поле **Первый список**, а вторую – в поле **Второй список**. Далее работа идет по той же схеме, как в рассмотренном выше примере 3.2.

3.3. КРИТЕРИИ СОГЛАСИЯ

Одним из наиболее важных разделов теории статистической проверки гипотез является проверка гипотез о законах распределения изучаемых признаков в генеральных совокупностях по выборочным данным. Это означает, что вариационный ряд, соответствующий распределению признака в выборке, отражает некий закон распределения этого признака, справедливый для всей генеральной совокупности.

Часто знание изучаемого материала или графическое изображение вариационного ряда позволяет предположить, что это неизвестное распределение является вполне определенным, например нормальным. Это предполагаемое распределение называется теоретическим, тогда как выборочное распределение называется эмпирическим. Естественно, возможно некоторое расхождение между теоретическим и эмпирическим распределениями. Для того чтобы оценить, связаны эти расхождения со случайностями выборки или же с неверным подбором теоретического закона распределения, и предназначены критерии согласия.

3.3.1. Сравнение эмпирического и теоретического распределений

Допустим, мы хотим проверить гипотезу о нормальном распределении признака в генеральной совокупности, т.е. проверяется гипотеза о том, что выборка получена из генеральной совокупности, в которой распределение является нормальным.

Чтобы сравнить имеющиеся данные с теоретическими, требуется определить те относительные частоты (доли или проценты), которые будут соответствовать каждому интервалу изменения признака в случае нормального распределения. Для этого сначала необходимо найти среднее значение и дисперсию признака по исходным данным. Пользуясь исходными (не сгруппированными) данными, вычисляют значения $\bar{x}$ и σ_x

Затем для каждого интервала изменения признака получают значение нормированного отклонения t_i:

$t_i = (x_i - \bar{x}) / \sigma_x$,

где x_i – верхняя граница интервала; i – номер интервала.

Затем по таблице нормального распределения для каждого значения t находятся вероятности нормированных (односторонних) отклонений, не превышающих t (таким же образом, как вычисляется в разделе **Вероят-**

ностный калькулятор значение *p* при одностороннем критерии). Для вычисления вероятности того, что признак *x* попадет на некий интервал значений, надо из значения вероятности для верхней границы этого интервала вычесть значение вероятности для его нижней границы. Таким образом можно получить значения вероятности попадания значений признака в каждый интервал (теоретические частоты).

Как правило, между эмпирическими и теоретическими значениями существует расхождение, однако надо измерить степень этого расхождения. Статистической характеристикой степени этого расхождения является т.н. величина X^2. Чем лучше соответствие между теоретическим и эмпирическим распределениями, тем ближе X^2 к нулю, а значит, большие значения X^2 должны способствовать отклонению гипотезы о нормальности распределения признака в генеральной совокупности.

Вычисление фактического значения статистической характеристики, выбор уровня значимости и получение по таблице распределения X^2 критического значения, соответствующего выбранному уровню значимости и числу степеней свободы *k* (которое равно для нормального распределения *n*–3, где *n* – число интервалов группировки), позволяет отклонить гипотезу о нормальности распределения в случае, если фактическое значение X^2 больше критического.

Пример 3.4. Вернемся к файлу General.sta и проверим нормальность распределения по возрасту (будем использовать для этого признак "год рождения". В программе STATISTICA обратимся к модулю **Подгонка распределений** (см. рис. 3.10.).

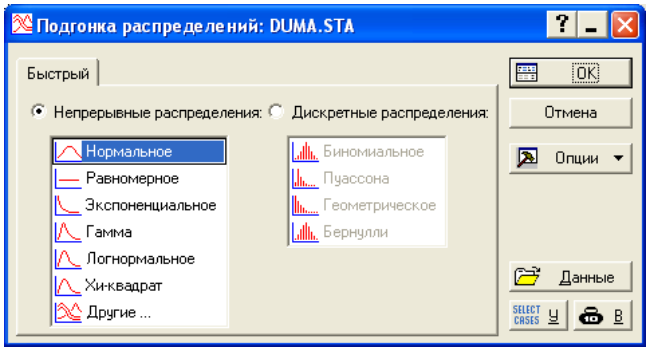


Рис. 3.10. Основное окно Подгонка распределений

В основном диалоговом окне **Подгонка распреджелений** надо выбрать нормальное распределение (в левом столбце, в списке непрерывных распределений) и нажать графическую кнопку OK. Откроется следующее диа-

логовое окно, в котором надо выбрать анализируемую переменную. Как обычно, нажатие графической кнопки **Переменные** открывает список переменных, в котором помечается нужная – в данном случае "год рождения" (см. рис. 3.11).

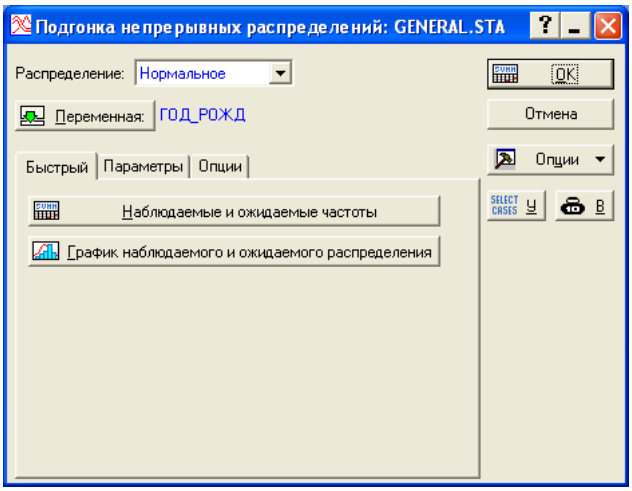


*Рис. 3.11. Вкладка **Быстрый** диалогового окна теста проверки нормальности*

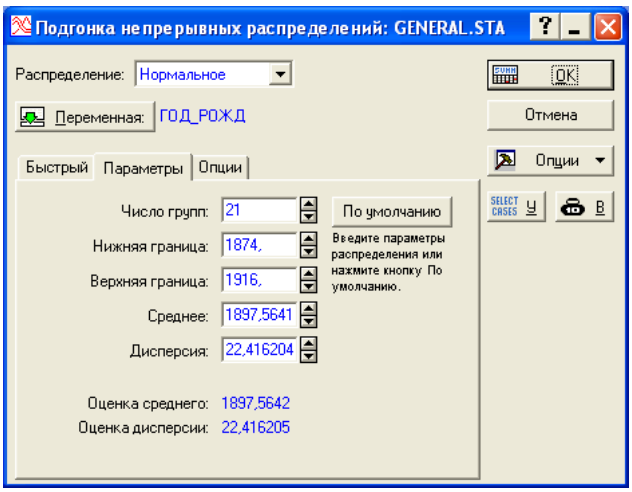


Рис. 3.12. Статистические характеристики выбранной переменной

Настройки теста проверки нормальности распределения доступны на вкладке **Параметры** (см. рис. 3.12). При этом сразу можно увидеть некото-

рые статистические характеристики выбранной переменной: число интервалов вариационного ряда, среднее значение, максимум, минимум, дисперсию и др.

Можно, однако, прямо на вкладке **Быстрый** вызвать графическое изображение двух распределений: реального распределения в виде гистограммы и теоретического, т.е. нормального распределения – в виде непрерывной кривой на фоне этой гистограммы). Щелчок по графической кнопке **График наблюдаемого и ожидаемого распределения** дает результат, показанный на рис. 3.13.

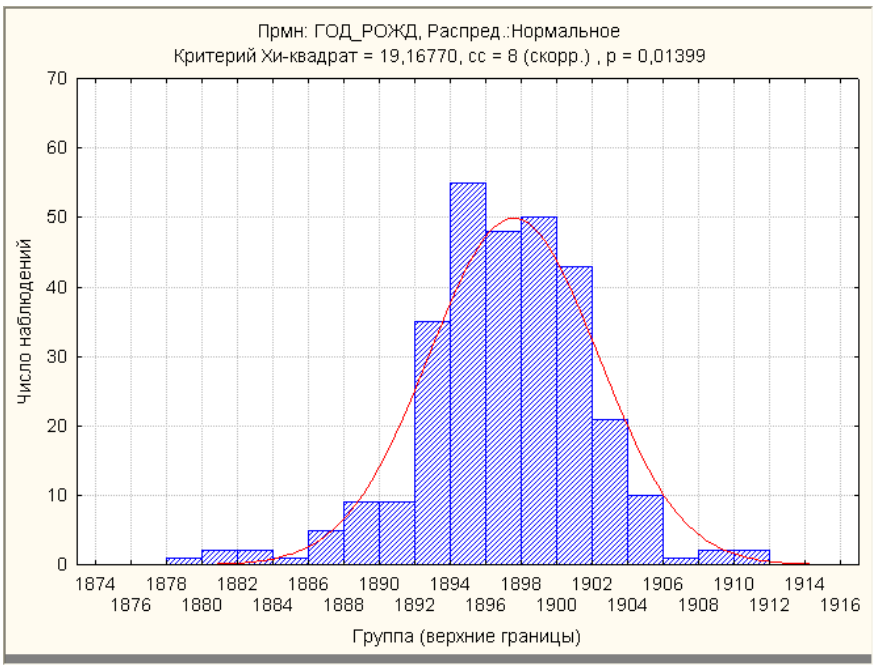


Рис. 3.13. Графическое представление соответствия реального и нормального распределения значений переменной "год рождения"

Видно, что гистограмма реального распределения не очень близка по характеру теоретической кривой нормального распределения. В заголовке этого графика приводятся значения величины критерия X^2 (19,17), числа степеней свободы (df = 8) и вероятности полученного значения X^2 (p = 0,014). Итак, тест показывает, что отклонение реального распределения от теоретического является значимым на "стандартном" уровне 5% (т.к. вероятность p меньше чем 0,05), а это означает, что гипотеза о хорошем соот-

ветствии реального распределения теоретически выбранному нормальному распределению должна быть отклонена.

3.3.2. Проверка нормальности распределения с помощью коэффициентов асимметрии и эксцесса

Укажем еще один способ проверки нормальности, который позволяет определить степень согласия между эмпирическим и нормальным теоретическим распределением. Это способ измерения коэффициентов асимметрии и эксцесса.

Коэффициент асимметрии *A* характеризует скошенность распределения в сторону больших или меньших значений признака. Для нормального распределения $A = 0$ (случай симметрии). Если $A>0$, говорят, что распределение имеет правостороннюю скошенность, если $A<0$ – левостороннюю.

Коэффициент эксцесса *E* характеризует степень островершинности распределения. Для нормального распределения $E = 0$, для островершинного – $E>0$, для плосковершинного – $E<0$.

Соответствующий метод проверки нормальности связан с сопоставлением полученных в выборке фактических значений показателей асимметрии и эксцесса со значениями, соответствующими нормальному распределению. Чем дальше значения A и E от нуля, тем хуже согласие с нормальным распределением.

Пример 3.5. Проверим гипотезу о нормальности распределения признака "год рождения" в файле General.sta с помощью коэффициентов асимметрии и эксцесса, причем ограничимся для примера группой выходцев из крестьян (поскольку для всей совокупности мы уже отклонили гипотезу о нормальности распределения). Для того чтобы выделить эту группу, надо активизировать окно таблицы исходных данных (просто "щелкнув" мышкой в любом месте этой таблицы) и выбрать в основном меню раздел **Сервис**, а в нем – команду **Условия выбора | Правка** [2]. В открывшемся диалоговом окне надо включить флажок **Задать условия выбора**, разрешающий выбирать объекты для анализа, в рамке **Включить наблюдения в анализ / графики** установить переключатель в положение **Заданные,** а в качестве **Условия** записать условие выбора объектов, для которых шестая переменная ("социальное происхождение") соответствует социальному происхождению из крестьян ("из_крест" – см. рис. 3.14). Щелчок по графической кнопке ОК завершает этап выбора объектов.

[2] Обратите внимание, что выбор объектов доступен не всегда, а лишь при условии, что предыдущий этап анализа закончен. Поэтому перед тем, как обратиться к выбору объектов, необходимо полностью завершить предыдущий диалог.

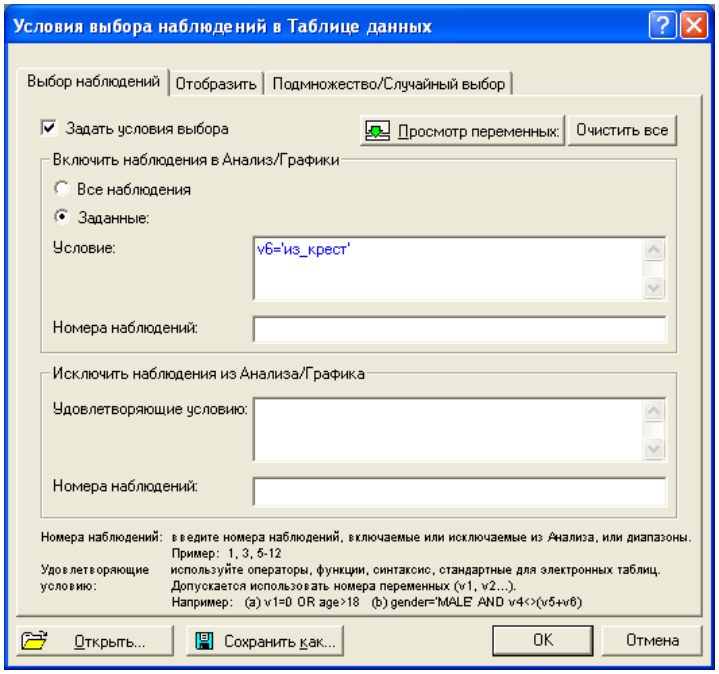


Рис. 3.14. Диалоговое окно выбора объектов

Теперь проверим гипотезу, что в генеральной совокупности $A = 0$ и $E = 0$ и что отличие их фактических значений от теоретических вызвано случайными причинами. Приближенный способ оценки близости распределения к нормальному с помощью коэффициентов A и E состоит в том, что фактические значения A и E сопоставляются с величинами их стандартных ошибок в выборке σ_A и σ_E. Выполним проверку на уровне значимости 1%: если окажется, что значения $|A|$ и $|E|$ превышают свои утроенные ошибки, гипотеза о нормальности отклоняется.

Обратимся к разделу **Описательные статистик** модуля **Основные статистики и таблицы**. На вкладке **Дополнительно** выберем вычисление коэффициентов **асимметрии** и **эксцесса** и их **стандартных ошибок**.

Полученные результаты приведены в табл. 3.1. Видно, что и коэффициент асимметрии A и коэффициент эксцесса E не превышают своих утроенных ошибок σ_A и σ_E, т.е. в данном примере распределение можно считать нормальным.

Таблица 3.1. Результаты вычисления коэффициентов асимметрии и эксцесса и их стандартных ошибок

Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
–0,696552	0,246210	0,415383	0,487732

ВОПРОСЫ

1. Что такое статистическая гипотеза?
2. Статистический критерий и статистическая характеристика
3. В чем состоит различие критической области и области допустимых значений?
4. Уровень значимости статистического критерия
5. Ошибки первого и второго рода
6. Как проверяется значимость различия средних значений?
7. Что такое критерии согласия?
8. Какими способами можно проверить нормальность распределения признака?
9. В чем смысл коэффициентов асимметрии и эксцесса?

ЗАДАНИЯ

1. По данным файла General.sta проверить с помощью коэффициентов асимметрии и эксцесса нормальность распределения признака "год рождения" для следующих групп по социальному происхождению:
   а) "из рабочих";
   б) "из служащих".
2. Проверить для файла General.sta гипотезу о том, что год присвоения звания "Дважды Герой Советского Союза" совпадает для военачальника с годом получения им высшего в своей военной карьере звания.
   Указание. Проверить гипотезу о равенстве средних значений признаков "герой_2" и "год_присвоения".
3. Для ответа на вопрос о влиянии возраста на политические взгляды депутатов I Государственной думы (файл Duma.sta) проверьте значимость различий в возрасте для:
   а) фракций трудовиков и мирнообновленцев;
   б) кадетов и партии демократических реформ.
4. Для файла Industry.sta проверить гипотезу о равенстве размеров промышленных предприятий, принадлежащих:
   а) частных владельцам и акционерным обществам;
   б) купцам первой гильдии и дворянам.

Указание. В качестве характеристики размера принять число рабочих.

5. Используя файл Industry.sta, ввести новую переменную "производительность труда" (как отношение переменных "произведено" и "рабочие"). Исследовать значимость различий средней производительности для:

а) деревообрабатывающей и строительной отраслей промышленности;

б) для хлопчатобумажной и шерстяной.

ЧАСТЬ II

СТАТИСТИЧЕСКИЙ АНАЛИЗ ВЗАИМОСВЯЗЕЙ

ГЛАВА 4

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

В изучении взаимосвязей количественных признаков можно выделить три этапа, заключающиеся в поиске ответов на вопросы: существует ли взаимосвязь между признаками; какова форма этой связи; каковы сила (теснота) и направление этой связи?

4.1. АНАЛИЗ ПАРНЫХ ВЗАИМОСВЯЗЕЙ

Самым простым случаем взаимосвязи является **парная взаимосвязь,** т.е. связь между двумя признаками. При этом предполагается, что взаимосвязь двух переменных носит, как правило, причинный характер т.е. одна из них зависит от другой. Первая (зависимая) называется в регрессионном анализе *результирующей,* вторая (независимая) – *факторной*. Следует заметить, что не всегда можно однозначно определить, какая из двух переменных является независимой, а какая – зависимой. Часто связь может рассматриваться как двунаправленная.

Пример 4.1. Рассмотрим данные по материалам Всероссийских промышленных переписей 1900 и 1908 гг. (файл Industry.sta – данные по Закавказью). В ряду сведений, содержащихся в переписи, есть количественные признаки "произведено" (объем производства в денежном исчислении, тыс. р.), "рабочие" (число занятых на предприятии рабочих) и "двигатели" (такой важный показатель технической оснащенности предприятия, как суммарная мощность установленных двигателей в л.с.). Можно поставить вопрос: зависел ли доход предприятия от числа рабочих и мощности двигателей и если зависел (что понятно и без статистического анализа), то от какого из двух признаков от зависел больше (т.е. экстенсивный или интенсивный характер носило производство)? Далее можно выяснить, как отличались эти зависимости в разных отраслях промышленности, т.е. в каких от-

раслях промышленности труд рабочих и использование машин давало больший экономический эффект (больший доход).

4.1.1. Построение диаграмм рассеяния

Простейшим (визуальным) способом выявить наличие взаимосвязи между количественными переменными является построение **диаграммы рассеяния**. Это график, на котором по горизонтальной оси (*X*) откладывается одна переменная, по вертикальной (*Y*) другая. Каждому объекту на диаграмме соответствует точка, координаты которой равняются значениям пары выбранных для анализа переменных.

Построим диаграмму рассеяния для переменных "произведено" и "рабочие". Поскольку характеристики предприятий сильно варьируют в зависимости от отрасли, ограничимся одной отраслью, например, металлообрабатывающей. Для того чтобы выделить эту группу предприятий в пакете STATISTICA следует воспользоваться *фильтром*: в меню выбрать раздел **Сервис**, а в нем – команду **Условия выбора | Правка**. Далее в открывшемся диалоговом окне (см. рис. 4.1) выполнить следующие действия: включить флажок **Задать условия выбора**, разрешающий выбирать объекты для анализа, в рамке **Включить наблюдения в анализ / графики** установить переключатель в положение **Заданные**, а в качестве **Условия** записать выражение *v3 = 131* (это означает выбор тех объектов, у которых значение признака номер 3, т.е. код отрасли, равняется 131 – металлообработка [1]). Щелчок по графической кнопке ОК завершает этап выбора объектов.

После того как задан такой фильтр, диаграмма будет содержать только точки, соответствующие металлообрабатывающим предприятиям. Для построения самой диаграммы рассеяния выберем команду **Диаграммы рассеяния** в разделе меню **Графика**. В появившемся диалоговом окне по умолчанию открывается вкладка **Быстрый** (см. рис. 4.2). Для построения простой диаграммы рассеяния, как обычно, прежде всего необходимо задать переменные, т.е. указать, какие признаки будут соответствовать осям координат графика (кнопка **Переменные**). Принято *зависимую* переменную отображать по оси *Y*, а *независимую* по оси *X*. В данной задаче естественно предполагать, что доход (как результат хозяйственной деятельности предприятия) зависит от числа рабочих (фактора производства), а не наоборот.

---

[1] Для того чтобы выяснить, какие коды присвоены программой различным значениям категоризованных признаков (таких, как код отрасли), следует в таблице исходных данных выделить столбец с интересующей нас переменной, дважды щелкнуть левой кнопкой мыши на ее названии и в появившемся окне "нажать" графическую кнопку **Текстовые метки**.

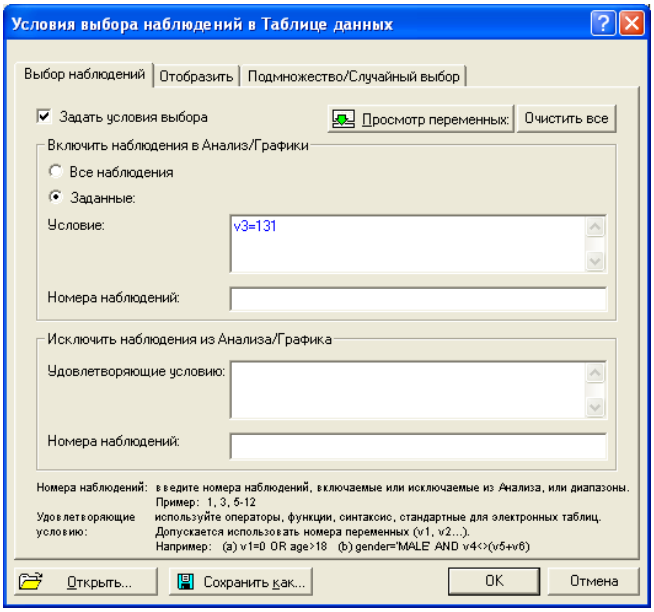


Рис. 4.1. Диалоговое окно для задания условия выбора объектов

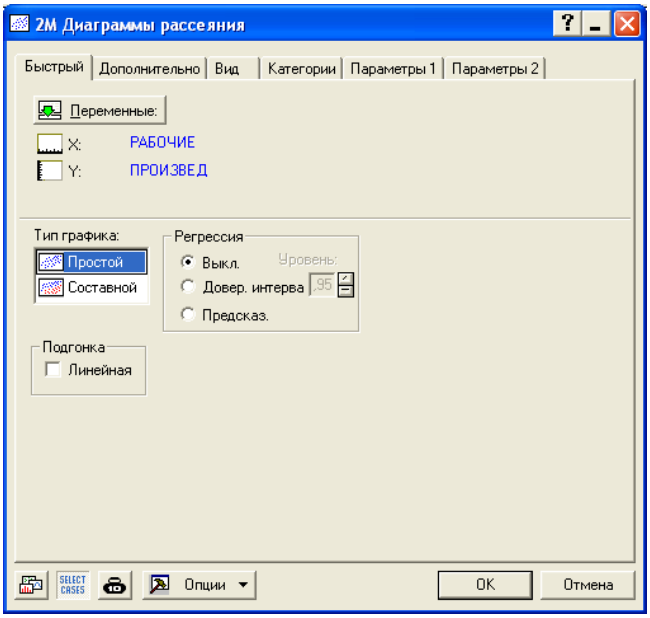


Рис. 4.2. Диалоговое окно параметров простой двумерной диаграммы рассеяния

Здесь же задаются **Тип графика**, по умолчанию – **Простой** и возможность вывода на экран графика линейной функции, которая наилучшим образом отражала бы тенденцию в расположении точек-объектов на графике (флажок Подгонка **линейная**). Подбор других (нелинейных) математических функций, а также некоторые другие параметры диаграммы рассеяния доступны на вкладке **Дополнительно**.

Завершив этап выбора объектов и щелкнув по графической кнопке **OK**, вы увидите диаграмму, представленную на рис. 4.3.

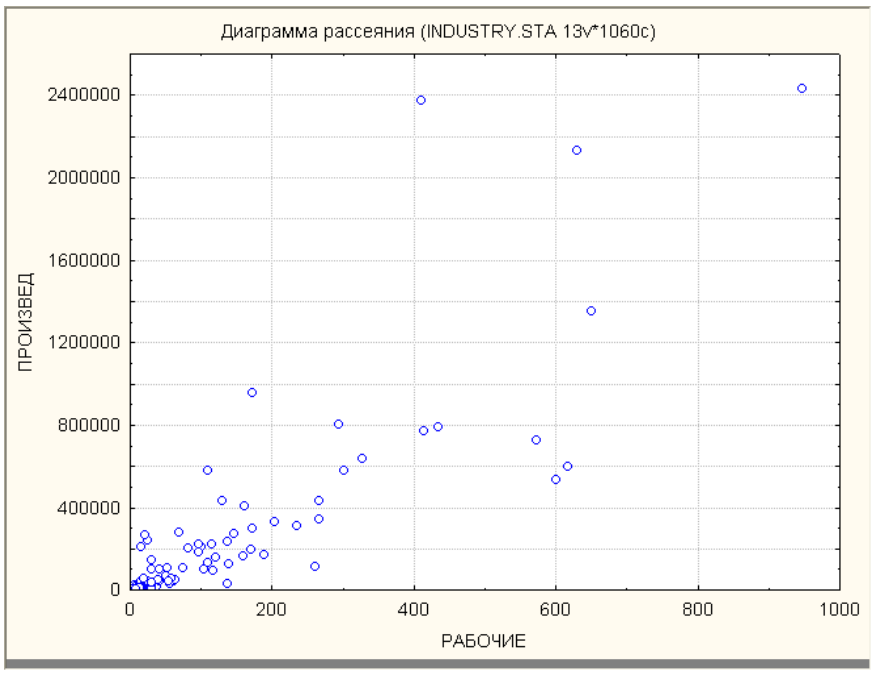


Рис. 4.3. Диаграмма рассеяния: зависимость объема производства в металлообрабатывающей промышленности от числа рабочих

Если бы существовала точная прямая зависимость между числом рабочих и доходом, т.е. если бы на каждом предприятии каждый рабочий за год производил в точности одинаковое количество продукции (в денежном выражении), то точки на диаграмме расположились бы на одной прямой. Однако в действительности эффективность труда (доход, производимый одним рабочим) различается на разных предприятиях, поэтому мы видим "облако" точек, о котором можно сказать, что оно вытянуто по диагонали от

левого нижнего угла к правому верхнему, т.е. в среднем доход растет с увеличением числа рабочих. По своему направлению такая связь называется *положительной*.

Полученный в данном случае результат можно считать тривиальным. Однако в более сложных случаях заранее ничего нельзя сказать о характере зависимости, и диаграмма рассеяния может дать необходимую информацию. Если точки оказываются хаотически рассеянными на диаграмме, т.е. их координаты представляют собой случайные пары чисел, то зависимости между переменными, видимо, не существует. Если точки образуют облако в направлении, перпендикулярном к тому, что на рис. 4.3, то между признаками существует ***отрицательная*** связь: чем больше значение одного, тем больше (в среднем) – значение другого.

Сопоставим теперь эффективность труда на предприятиях металлообрабатывающей отрасли и на предприятиях по обработке шелка. Код этой отрасли равен 127, поэтому установим фильтр v3 = 127 и построим новую диаграмму (рис. 4.4).

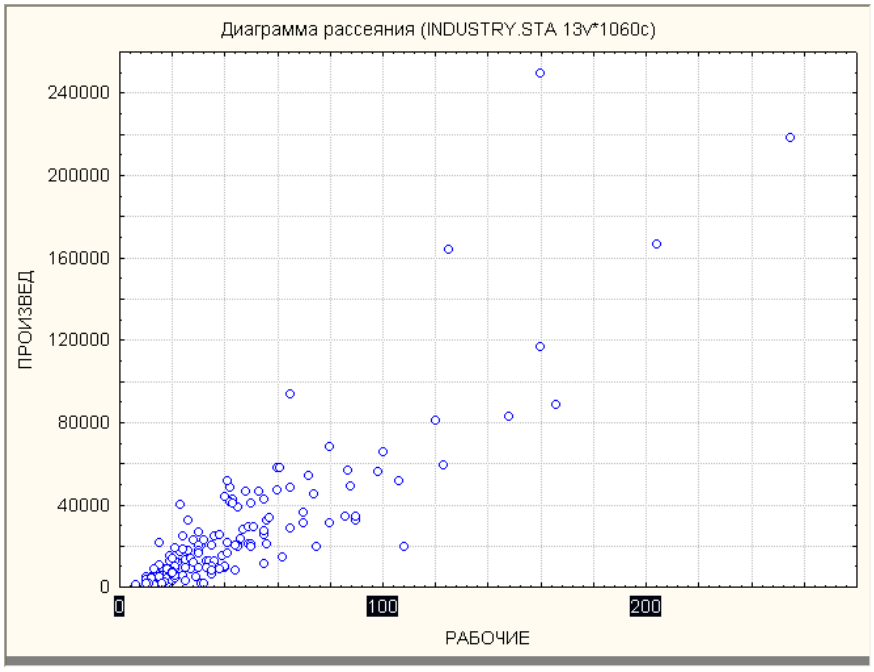


Рис. 4.4. Диаграмма рассеяния: зависимость объема производства шелковой промышленности от числа рабочих

Следует обратить внимание на то, что диаграмма для шелковой промышленности построена в ином масштабе. Это связано с тем, что программа автоматически выбирает наиболее удобный для расположения точек на отдельном графике масштаб. В шелковой промышленности предприятия меньше и по числу рабочих, и по стоимости производимой продукции, поэтому масштаб на рис. 4.4 крупнее, чем на рис. 4.3 (в десять раз по оси *Y* и в три раза по оси *X*).

Для удобства сравнения графиков можно построить *категоризованную диаграмму*, на которой одновременно будут отражены зависимости для нескольких отраслей в одинаковом масштабе. Для этого надо в разделе меню **Графика** выбрать **Категоризованные графики | Диаграммы рассеяния**.

В открывающемся диалоговом окне (рис. 4.5) надо перейти на вкладку **Дополнительно**, где задать несколько параметров:

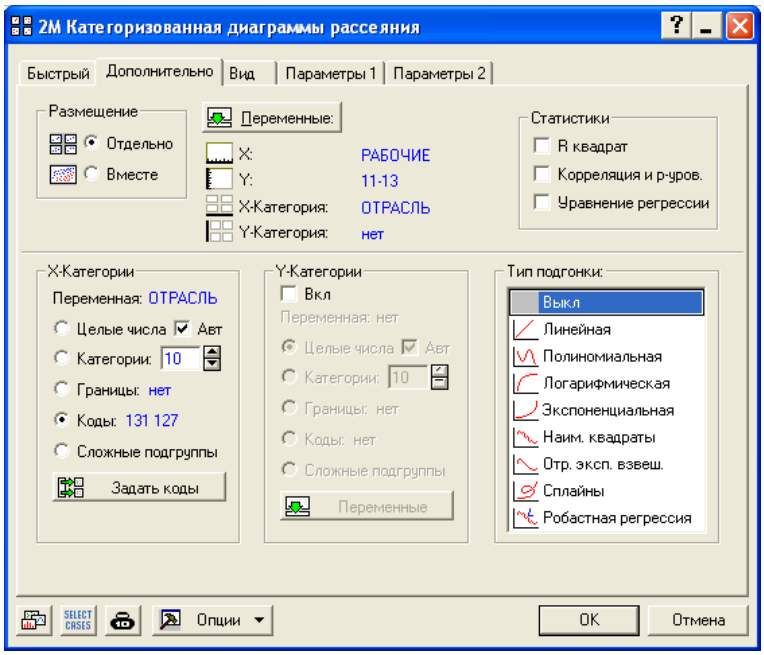


Рис. 4.5. Диалоговое окно для построения категоризованных диаграмм рассеяния. Вкладка ***Дополнительно***

- в блоке **Переменные** – указать, что по горизонтали (***X*-категория**) будут размещены диаграммы по разным категориям признака "отрасль", причем по оси *X* (**Перемен. *X***) откладываются значения признака "рабочие", а по оси *Y* (**Перемен. *Y***) – значения признака "произведено";

- в блоке ***X*-категории** – указать (с помощью переключателя **Коды** и графической кнопки **Задать коды**) конкретные коды интересующих нас отраслей (т.е. 131 и 127);
- в блоке **Тип подгонки** (подбор линии, математически выражающей зависимость признаков) пока можно оставить значение **Выкл**. (*выключено*).

Щелчок по графической кнопке **OK** даст диаграммы, показанные на рис. 4.6.

Приведенное сопоставление позволяет не только визуально оценить различия в масштабах двух отраслей промышленности, но и увидеть различия в характере зависимости между исследуемыми признаками. В принципе зависимости сходны в обеих отраслях, однако легко заметить, что облако точек на левой диаграмме – более разреженное, чем на правой.

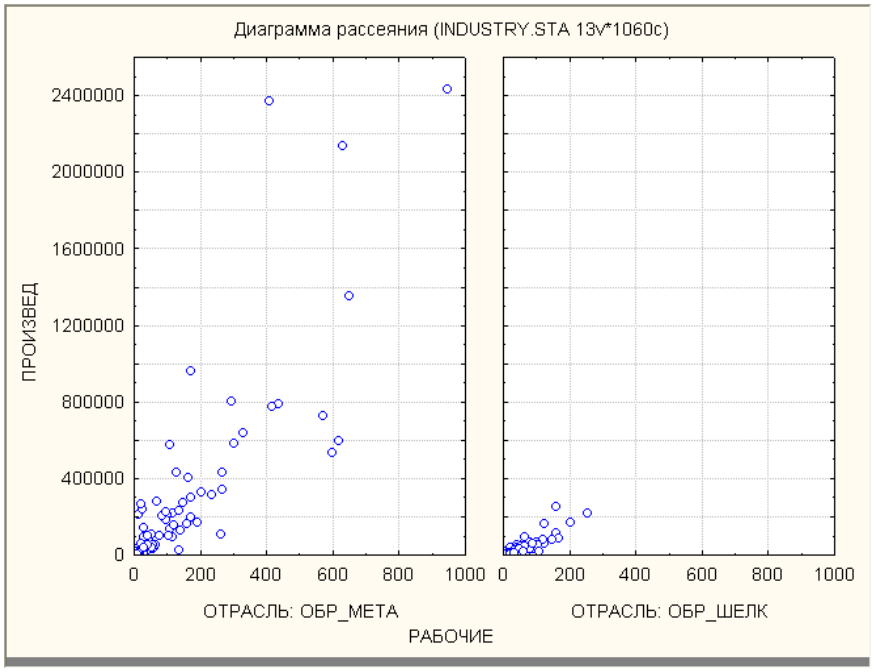


Рис. 4.6. Диаграммы рассеяния для металлообрабатывающей и шелковой отраслей промышленности в одинаковом масштабе

4.1.2. Построение уравнения линейной регрессии

Если бы между признаками существовала строгая, т.е. функциональная зависимость, все точки находились на некоторой линии, математическое

уравнение которой и представляло бы эту зависимость. Если наше облако точек напоминает очертания некоторой линии, например, прямой, то можно предполагать, что мы видим на диаграмме рассеяния именно такую по форме зависимость, однако искаженную воздействием как случайных, так и неучтенных факторов, вызывающим отклонение точек от *теоретической* кривой. Поскольку наиболее простой формой зависимости в математике является прямая, то в регрессионном анализе наиболее популярны *линейные модели*.

Попробуем провести прямую линию через каждое облако точек на рис. 4.6. Таких линий можно нарисовать множество, причем на глаз невозможно определить, какая из них лучше подходит для описания диаграммы рассеяния точек. Существует, однако, метод, который позволяет совершенно точно вычислить положение прямой линии, наилучшим образом проходящее через облако точек. Это – *метод наименьших квадратов*. Вычисляемая с его помощью линия называется *линией регрессии*. Она характеризуется тем, что сумма квадратов расстояний от точек на диаграмме до этой линии минимальна (по сравнению со всеми возможными линиями). Таким образом, линия регрессии дает наилучшее (статистически) приближенное описание линейной зависимости между двумя переменными.

Увидеть линию регрессии на экране можно, если в диалоговом окне построения диаграмм рассеяния (см. рис. 4.5) в блоке **Тип подгонки** выбрать значение **Линейная**. Щелчок по графической кнопке ОК выводит на экран диаграммы, представленные на рис. 4.7.

Как известно, прямая линия в координатах (*X,Y*) описывается уравнением

$y = a + bx$

где b характеризует наклон [1] линии и показывает, насколько изменяется значение зависимой переменной y при изменении независимой переменной x на единицу; a – координата точки пересечения линии с осью Y; она показывает, каково значение зависимой переменной, когда независимая равняется нулю.

Таким образом, уравнение линейной регрессии $y = a + bx$ показывает зависимость результативного признака y от факторного признака x. Коэффициент при x называется коэффициентом регрессии.

В верхней части рис. 4.7 приводятся уравнения регрессии. Первое из них (для металлообрабатывающих предприятий) имеет вид: $y = –19140 + 2143x$, где y – произведенная продукция, а x – число рабочих. Смысл этого уравнения состоит в том, что на металлообрабатывающих предприятиях увеличение числа рабочих на 1 приводит в среднем к увеличению объема

[1] Точнее, равняется тангенсу угла наклона.

годового производства примерно на 2000 руб. (2143 – коэффициент регрессии).

Второе уравнение (для предприятий по обработке шелка) имеет вид: $y = -8934 + 805x$, где y и x – те же переменные, что в первом уравнении. Смысл второго уравнения: для предприятий по обработке шелка увеличение числа рабочих на 1 приводит в среднем к увеличению объема годового производства примерно на 800 руб. (805 – коэффициент регрессии). Сравнение коэффициентов регрессии показывает, что в металлообрабатывающей отрасли эффективность труда выше.

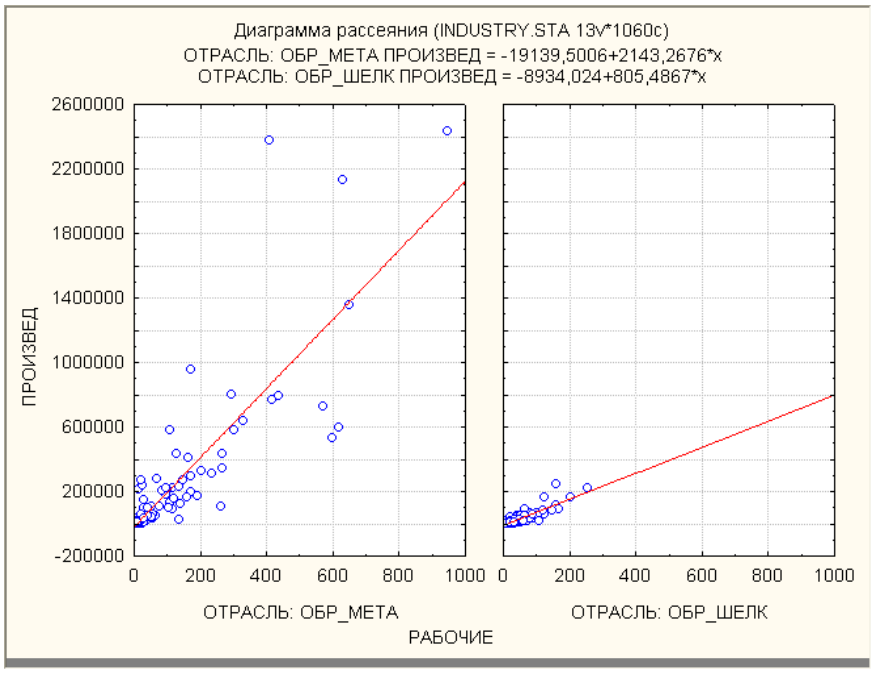


Рис. 4.7. Диаграммы рассеяния для металлообрабатывающей и шелковой отраслей промышленности с линиями регрессии

Однако полученные результаты нельзя интерпретировать таким образом: "каждый рабочий в металлообрабатывающей отрасли производит в год в среднем продукции на сумму около двух тысяч рублей". Дело в том, что в уравнениях присутствуют константы, приближенно равные –19000 и –88000, соответственно. Смысл их должен состоять в том, что в этих отраслях предприятия с нулевым числом рабочих (т.е. без рабочих) производят

продукцию на указанные суммы. Это, разумеется, противоречит здравому смыслу, хотя таков наиболее точный в математическом смысле ответ.

Причина такого несоответствия рассматриваемых моделей реальности состоит в том, что труд не является единственным производственным фактором, а промышленное предприятие не может нанимать рабочих в количестве ниже некоторого определяемого технологией минимума. Таким образом, данное уравнение регрессии не имеет смысла использовать для значений переменной "число рабочих", близких к нулю. Напротив, значения свободного члена в уравнениях регрессии для зависимости дохода от энерговооруженности могут иметь реальный смысл, т.к. предприятия ручного труда (суммарная мощность двигателей равна нулю) вполне могут производить продукцию.

4.1.3 Коэффициент корреляции

Какова бы ни была конфигурация облака точек на диаграмме рассеяния – лежат ли они в точности на одной прямой, или разбросаны хаотически – любая статистическая программа всегда сможет построить уравнение регрессии. Однако в первом случае оно будет весьма *достоверным*, а во втором – нет. Более того, может оказаться, что через два облака, различающихся степенью близости к линейной конфигурации, будут проведены одинаковые линии регрессии. Коэффициенты уравнения будут одинаковы, но тем не менее зависимость между двумя переменными будет иметь различный характер. Иначе говоря, коэффициенты уравнения регрессии не дают исчерпывающий ответ на вопрос о степени (тесноте) силы связи пары переменных.

Важной мерой, дополняющей уравнение регрессии, является **коэффициент корреляции**. Он показывает, насколько тесно две переменные связаны между собой. Коэффициент корреляции вычисляется по формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

где x_i и y_i – значения переменных у i-го объекта,

$\bar{x}$ и $\bar{y}$ – средние арифметические значения переменных,

n – объем совокупности.

Коэффициент корреляции r принимает значения в диапазоне от –1 до +1. Если $r = 1$, то между двумя переменными существует функциональная положительная линейная связь, т.е. на диаграмме рассеяния соответствую-

щие точки лежат на одной прямой с положительным наклоном. Если $r = -1$**,** то между двумя переменными существует функциональная отрицательная зависимость. Если $r = 0$**,** то рассматриваемые переменные **линейно независимы**, т.е. на диаграмме рассеяния облако точек "вытянуто по горизонтали"**.**

Следует указать на важность понятия о линейной зависимости, использовавшегося в этом разделе: уравнение регрессии и коэффициент корреляции целесообразно вычислять лишь в том случае, когда зависимость между переменными может хотя бы приближенно считаться линейной. В противном случае результаты могут быть совершенно неверными, в частности, коэффициент корреляции может оказаться близким к нулю при наличии сильной взаимосвязи. В особенности это характерно для случаев, когда зависимость имеет явно нелинейный характер (например, зависимость между переменными приблизительно описывается синусоидой или параболой). Во многих случаях эту проблему можно обойти, преобразовав исходные переменные. Однако, чтобы догадаться о необходимости подобного преобразования, т.е. для того чтобы узнать, что данные могут содержать сложные формы зависимости, их желательно "увидеть". Именно поэтому исследование взаимосвязей между количественными переменными обычно должно включать просмотр диаграмм рассеяния. Как правило, статистические пакеты позволяют соединять эту работу с исследованием параметров регрессионной зависимости. Анализ корреляций является здесь необходимым дополнением, однако, может использоваться и самостоятельно, и комбинироваться с двумя другими задачами в произвольной последовательности.

Коэффициенты корреляции можно вычислять и без предварительного построения линии регрессии. В этом случае вопрос об интерпретации признаков как результативных и факторных, т.е. зависимых и независимых, не ставится, а корреляции понимается как согласованность или синхронность одновременного изменения значений признаков при переходе от объекта к объекту.

Если объекты характеризуются целым набором количественных признаков, можно сразу построить т.н. матрицу корреляции, т.е. квадратную таблицу, число строк и столбцов которой равно числу признаков, а на пересечении каждых строки и столбца стоит коэффициент корреляции соответствующей пары признаков. Для этого можно воспользоваться разделом **Парные и частные корреляции** модуля **Основные статистики и таблицы.**

Пример 4.2. Построим матрицу корреляции между переменными "произведено", "рабочие" и "двигатели" для всей совокупности предприятий в файле Industry.

В верхней части диалогового окна (см. рис. 4.8), которое по умолчанию открывается на вкладке **Quick**, есть две графические кнопки: **Квадратная матрица** и **Прямоугольная матрица**.

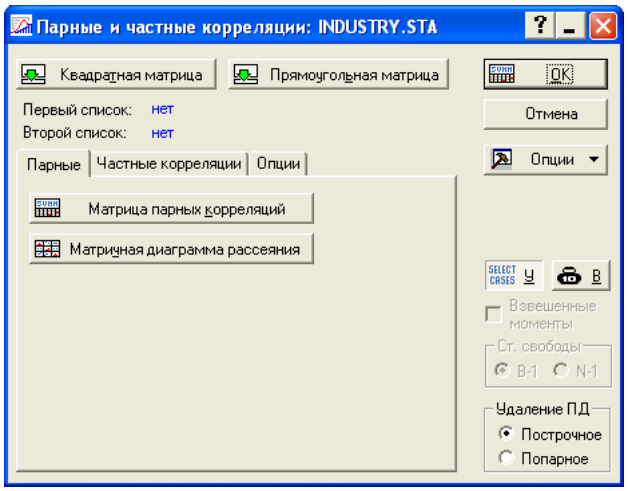


Рис. 4.8. Диалоговое окно для построения матрицы корреляции

Выбор кнопки **Квадратная матрица** означает построение стандартной квадратной матрицы корреляций для заданных переменных. Для случая, когда мы хотим получить корреляции между двумя разными группами переменных, выбирается кнопка **Прямоугольная матрица**. Например, если мы хотим получить матрицу корреляций одной переменной со всеми остальными, тогда необходимо указать два списка переменных: в одном выбрать одну переменную, в другом – все остальные. Построенная таким образом матрица уже не будет квадратной.

В нашем примере выберем кнопку **Квадратная матрица** и зададим три перечисленные выше признака. Щелкнув по графической кнопке **Матрица парных корреляций** в центре (или графической кнопке **OK** в правом верхнем углу) диалогового окна**,** мы получим корреляционную матрицу 3x3 (рис. 4.9).

Каждая клетка таблицы содержит коэффициент корреляции между переменными, обозначенными в заголовках столбцов и строк. Видно, что корреляция для признаков "произведено" – "рабочие" выше, чем для признаков "произведено" – "двигатели". Иначе говоря, число рабочих оказывает б□□ольшее влияние на производство, дает более предсказуемый результат. Двигатели же используются менее систематически, результаты их при-

менения на предприятиях при равных условиях отличаются высокой вариативностью.

	Корреляции (INDUSTRY.STA) Отмеченные корреляции значимы на уровне p <,050 N=1060 (Построчное удаление ПД)		
Переменная	**РАБОЧИЕ**	ПРОИЗВЕД	ДВИГАТЕЛ
РАБОЧИЕ	1,00	0,52	0,43
ПРОИЗВЕД	0,52	1,00	0,30
ДВИГАТЕЛ	0,43	0,30	1,00

Рис. 4.9. Матрица корреляции для трех переменных

4.1.4. Проверка гипотезы о значимости коэффициента корреляции

Если коэффициент корреляции вычислен на основе выборочных данных, то не исключено, что его ненулевое значение является не отражением действительной связи между признаками, а просто получено в результате специфики данной выборки (тогда как в генеральной совокупности он равен нулю). Возникает типичная задача статистической проверки гипотез (см. главу 3): строится т.н. статистическая характеристика, зависящая от коэффициента корреляции, распределение которой известно. Гипотеза о независимости признаков отклоняется, если вероятность значений статистической характеристики, не меньших ее фактического значения, достаточно мала, например, не превышает 5%. В этом случае коэффициент корреляции называется *значимым*.

Статистической характеристикой для проверки значимости корреляции служит отношение самого коэффициента к его утроенной ошибке, вычисляемое по формуле:

$$t = r\sqrt{n-2} / \sqrt{1-r^2},$$

где n – объем выборки. Эта величина табулирована, т.е. известны вероятности всех ее значений [1]. Чем больше значение t, тем меньше его вероятность, т.е. вероятность того, что данная или большая величина корреляции может быть получена в выборке из генеральной совокупности, в которой корреляция равна нулю. Если эта вероятность окажется меньше выбранного уровня значимости, гипотеза о некоррелированности признаков отклоняется, а связь признается значимой.

По умолчанию в пакете STATISTICA принят уровень значимости, равный 0,05. Обратите внимание, что в матрице корреляции на рис. 4.9 значимые на этом уровне коэффициенты выделены на экране красным цветом.

[1] Значения t приводятся, как правило, в приложениях к учебникам по статистике.

Все коэффициенты в нашей матрице значимы, даже имеющие небольшую величину.

	Корреляции (INDUSTRY.STA) Отмеченные корреляции значимы на уровне p <,050 N=1060 (Построчное удаление ПД)		
Переменная	**РАБОЧИЕ**	ПРОИЗВЕД	ДВИГАТЕЛ
РАБОЧИЕ	1,0000	,5159	,4332
	p= ---	p=0,00	p=0,00
ПРОИЗВЕД	,5159	1,0000	,2991
	p=0,00	p= ---	p=0,00
ДВИГАТЕЛ	,4332	,2991	1,0000
	p=0,00	p=0,00	p= ---

Рис. 4.10. Матрица корреляции для трех переменных с проверкой значимости на уровне 0,05

Итак, коэффициент корреляции не является сам по себе "большим" или "маленьким"; и стоит его принимать во внимание или нет – зависит не только от его величины, но и от уровня значимости. Относительно большой коэффициент может оказаться незначимым при малом объеме данных, а малый коэффициент может указывать хотя и на слабую, но достоверную связь, пренебрегать которой не следует.

Если в диалоговом окне на рис. 4.8 на вкладке **Опции** выбрать возможность просмотра вероятностей, соответствующих корреляционным коэффициентам (**Отображать p-уровень и N**), то можно увидеть значения вероятностей *p* и убедиться, что они не превышают уровня значимости 5%, о чем и говорится в заголовке таблицы на рис. 4.10.

4.1.5. Коэффициент детерминации

Сам по себе коэффициент корреляции не имеет содержательной интерпретации. Однако его квадрат, называемый *коэффициентом детерминации* (R^2), имеет простой смысл – это показатель того, насколько изменения зависимого признака (в процентах) объясняются изменениями независимого. Более точно, это доля дисперсии независимого признака, объясняемая влиянием зависимого.

Если две переменные функционально линейно зависимы (точки на диаграмме рассеяния лежат на одной прямой), то можно сказать, что изменение переменной *y* полностью объясняется изменением переменной *x,* а это как раз тот случай, когда коэффициент детерминации равен единице (при этом коэффициент корреляции может быть равен как 1, так и –1). Если две переменные линейно независимы (метод наименьших квадратов дает горизонтальную прямую), то переменная *y* своими вариациями никоим образом "не обязана" переменной *x*– в этом случае коэффициент детерминации равен нулю. В промежуточных случаях коэффициент детерминации указыва-

ет, какая часть изменений переменной *y* объясняется изменением переменной *x* (иногда удобно представлять эту величину в процентах).

Пример 4.3. На рис. 4.9 была приведена матрица корреляции трех переменных – "произведено", "рабочие" и "двигатели" – для всех отраслей промышленности. Коэффициент корреляции между объемом производства и числом рабочих $r_1 = 0.52$, а между объемом производства и мощностью двигателей $r_2 = 0.30$. Возведя r_1 и r_2 в квадрат, мы получим $R_1^2 = 0.27$ и $R_2^2 = 0.09$. Таким образом, в целом по предприятиям Закавказья доход предприятий почти на треть определяется числом рабочих и лишь на девять процентов – энерговооруженностью.

Необходимо отметить, что, вычисляя влияние нескольких факторов на одну переменную, в общем случае некорректно складывать полученные коэффициенты. В нашем примере нельзя непосредственно вычислить суммарное влияние переменных "рабочие" и "двигатели" на доходность, поскольку между ними тоже существует зависимость – корреляция сравнительно невысокая (0,43), но значимая. О том, как анализировать зависимости, в которых участвуют более двух переменных, говорится в следующем разделе.

4.2. МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

До сих пор мы исследовали парные зависимости. Однако чаще всего на зависимую переменную действуют сразу несколько факторов, среди которых трудно выделить единственный или главный Так, в приведенном выше примере доход предприятия зависит *одновременно* от двух факторов производства – числа рабочих и энерговооруженности. Причем оба этих фактора сами не являются независимыми друг от друга: для обслуживания большего числа машин требуется больше рабочих. Поэтому совокупная зависимость дохода от рабочих и мощности двигателей не есть простая сумма двух парных зависимостей; она находится более сложным методом, который носит название *множественной регрессии*.

4.2.1. Визуализация множественной зависимости в пространстве трех переменных

В случае трех признаков возможна визуализация данных, аналогичная диаграмме рассеяния: построение трехмерных графиков поверхностей.

Пример 4.4. Вернемся к файлу Industry и построим такой график для предприятий металлообрабатывающей отрасли промышленности.

Для выбора предприятий надо активизировать таблицу исходных данных, затем в разделе меню **Сервис** выбрать команду **Условия выбора | Правка**, в открывшемся диалоговом окне включить флажок **Задать усло-**

вия выбора, а в блоке **Включить наблюдения в анализ / графики** – переключатель **Заданные** и задать **Условие**: код признака "отрасль" равен 131 (v3 = 131), что соответствует металлообработке (см. рис. 4.11).

Затем в разделе меню **Графика** перейдем к группе трехмерных графиков – **3M XYZ Графики**. Здесь можно выбирать разные виды трехмерных графиков, например, как в случае с одной независимой переменной, выбрать диаграмму рассеяния (**Диаграммы рассеяния**), на которой объекты изображаются точками. Рассмотрим такой вид трехмерного графика, как представление исходных данных в виде некоторой поверхности (**Графики поверхностей**). Если выбрать построение поверхности, откроется диалоговое окно (см. рис. 4.12) [1], где, как обычно, используя кнопку **Переменные**, необходимо указать, в пространстве каких переменных мы будем рассматривать объекты (в нашем случае – предприятия).

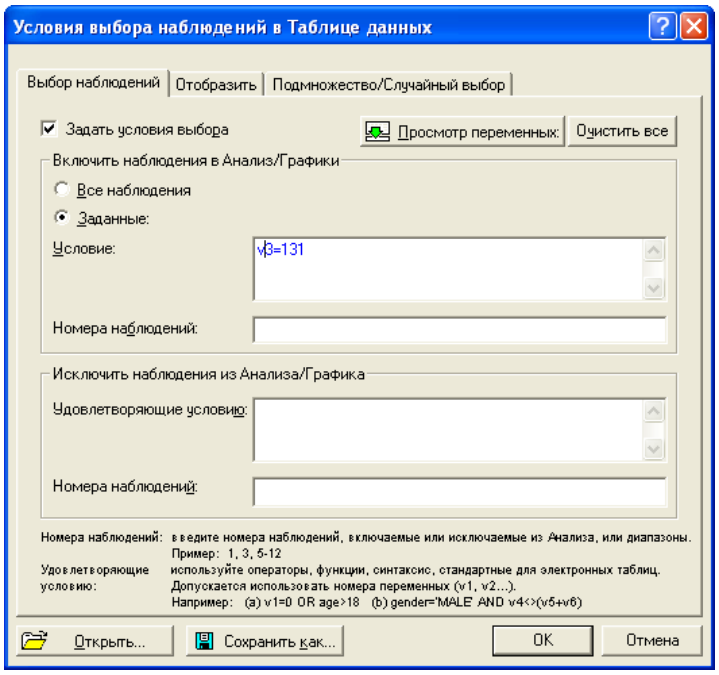


Рис. 4.11. Выбор объектов для анализа

[1] Это же диалоговое окно открывается, если в разделе меню **Графика** сразу выбрать **Графики поверхностей**.

В блоке **Подгонка** надо выбрать один из методов построения поверхности. В этом блоке перечислены основные способы построения (приближения) поверхности. Во многих случаях наиболее подходящим является сглаживание данных по принципу наименьших квадратов, при котором – по аналогии с построением линии регрессии на плоскости – среднеквадратическое отклонение поверхности от "облака" точек-объектов будет минимальным. Полученная поверхность по смыслу близка регрессионной модели, однако имеет более сложный (нелинейный) вид. В данном случае для подсчета расстояний от точек-объектов до поверхности выбран метод **Наименьшие квадраты**.

Наконец, на диаграмме можно отобразить не только поверхность, но и точки, соответствующие исходным значениям. Для этого включим флажок **Показать точки данных на поверхности.**

Задав все необходимые параметры, получим диаграмму, показанную на рис. 4.13. Как на физической карте мира, изменение цвета отражает высоту поверхности, т.е. вертикальную координату. В нашем примере вертикальная координата соответствует доходу, получаемому предприятием в зависимости от числа рабочих и мощности двигателей.

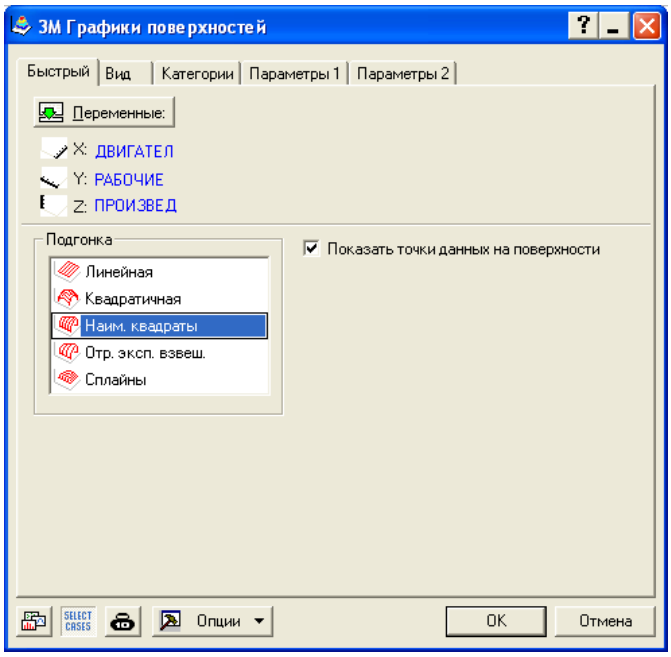


Рис. 4.12. Диалоговое окно для построения трехмерных поверхностей

Рассмотрение такой поверхности может оказаться полезным для выявления существенной нелинейности в данных, которая может поставить под сомнение применимость линейной регрессии.

Из графика видно, что предприятия металлообрабатывающей промышленности можно разделить на две группы, разделенные "пропастью" – изгибом, отражающим невысокий уровень доходности. К первой группе (слева) относятся предприятия, увеличивающие свой доход за счет увеличения числа рабочих, ко второй (справа) – за счет увеличения мощности двигателей.

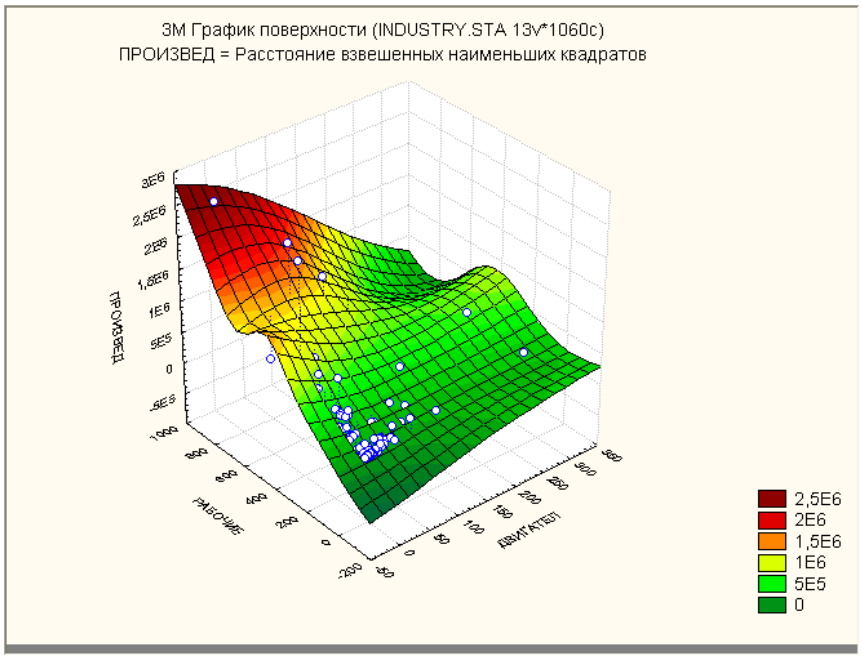


Рис. 4.13. Объекты (предприятия металлообрабатывающей промышленности) в пространстве трех переменных

4.2.2. Уравнение множественной регрессии

Уравнение множественной регрессии аналогично парной, но включает больше одной независимой переменной:

$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k$,

где x_1, x_2, . . . – независимые переменные, от которых в той или иной степени зависит исследуемая (результирующая) переменная y;

b_1, b_2 . . . – коэффициенты при соответствующих переменных (*коэффициенты регрессии*), показывающие, насколько в среднем изменится значение результирующей переменной при изменении отдельной независимой переменной на единицу (и фиксированных значениях остальных переменных).

Уравнение множественной регрессии задает *регрессионную модель*, объясняющую поведение зависимой переменной. Следует, однако, еще раз подчеркнуть, что никакая регрессионная модель не в состоянии указать, какая переменная является зависимой (следствием), а какие – независимыми (причинами). Решение такого рода всегда выносится исследователем, исходя из знания исследуемой предметной области.

Построим уравнение множественной регрессии для нашего примера. Для этого обратимся к модулю **Множественная регрессия** программы STATISTICA (рис. 4.14).

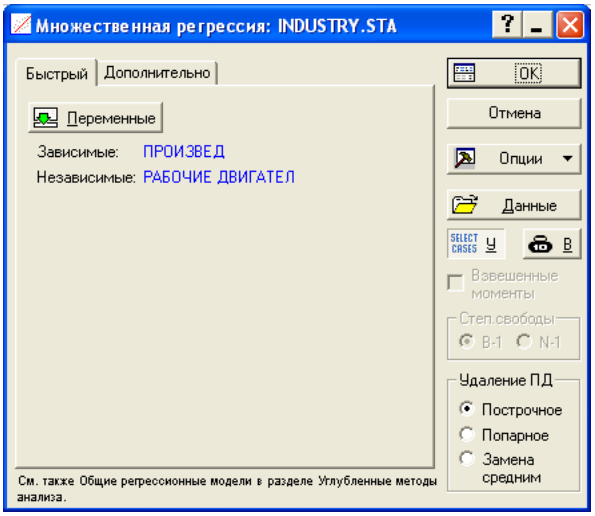


Рис. 4.14. Диалоговое окно для построения множественной регрессии

Сначала, как и всегда, с помощью графической кнопки **Переменные** укажем программе, что переменная "произведено" является **зависимой**, а переменные "рабочие" и "двигатели" являются **независимыми**. Нажав графическую кнопку **OK**, перейдем к диалоговому окну **Результаты множественной регрессии**, которое по умолчанию открывается на вкладке **Быстрый** (рис. 4.15).

Это окно разделено на три части. Верхняя часть содержит следующие наиболее важные результаты:

- коэффициент множественной корреляции (*Множест. R*) – аналог коэффициента парной корреляции, описанного выше, представляет собой меру совокупной зависимости результирующего переменной от всех независимых (факторных);
- коэффициент детерминации (R^2) – то же по смыслу, что и в случае парной регрессии, т.е. доля дисперсии результирующей переменной, объясняемая влиянием всех независимых переменных;
- уровень значимости регрессионной модели (см. следующий раздел).

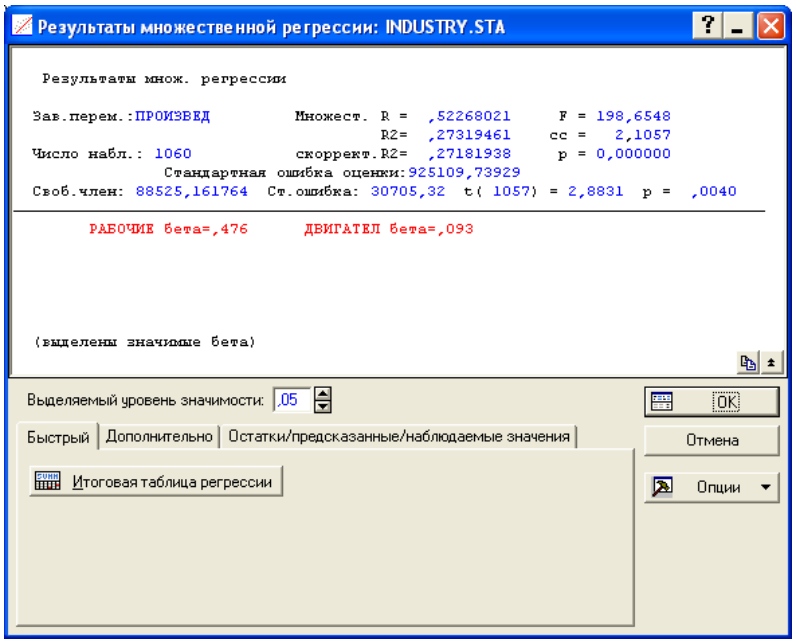


Рис. 4.15. Диалоговое окно просмотра результатов множественной регрессии (вкладка Quick)

В средней части окна результатов показаны "нормированные" коэффициенты регрессии (здесь они называются *бета*) [1] для всех указанных неза-

[1] "Нормированные" или стандартизованные коэффициенты регрессии получаются при умножении обычных коэффициентов на дробь, в числителе которой стоит стандартное отклонение соответствующего фактора, а в знаменателе – стандартное отклонение результативного признака. Переход к нормированным коэффициентам регрессии объясняется тем, что они позволяют более корректно сравнивать силу

висимых переменных, причем цветом выделяются *значимые* коэффициенты. Можно видеть, что значимыми являются оба факторных признака, включенные нами в модель. [1]

В нижней части окна находятся ярлыки вкладок и графическая кнопка **Итоговая таблица регрессии**. Для просмотра коэффициентов регрессии можно щелкнуть либо этой кнопке, либо по кнопке ОК. На экране появится таблица, показанная на рис. 4.16.

Эта таблица, кроме стандартизованных, содержит обычные коэффициенты регрессии (столбец *B*). Например, значение коэффициента регрессии 2752, соответствующего фактору "рабочие", означает, что увеличение числа рабочих на 1 в среднем приводит к повышению объема производства на 2752 руб. при условии, что мощность двигателей не изменяется. Коэффициент регрессии при втором факторе означает, что увеличение суммарной мощности двигателей на 1 л.с. приводит в среднем к увеличению объема производства на 93 руб. при условии, что число рабочих не изменяется.

Итоги регрессии для зависимой переменной: ПРОИЗВЕД (
R= ,52268021 R2= ,27319461 Скорректир. R2= ,27181938
F(2,1057)=198,65 p<0,0000 Станд. ошибка оценки: 9251E2

N=1060	**БЕТА**	Стд.Ош. БЕТА	B	Стд.Ош. B	t(1057)	p-уров.
Св.член			88525,16	30705,32	2,88306	0,004018
РАБОЧИЕ	0,475577	0,029094	2752,36	168,38	16,34643	0,000000
ДВИГАТЕЛ	0,093092	0,029094	93,32	29,17	3,19974	0,001416

Рис. 4.16. Таблица регрессионных коэффициентов

Важным свойством регрессионного уравнения является статистический прогноз – возможность вычислять значения зависимой переменной (вкладка **Остатки/предсказанные/наблюдаемые значения** в диалоговом окне

влияния разных факторов на результат. Увидеть же обычные коэффициенты регрессии можно, нажав в окне результатов на графическую кнопку **Итоговая таблица регрессии**.

[1] Регрессионная модель, рассматриваемая в нашем примере является полной регрессией, когда в результат включаются все независимые переменные. Существуют и другие методы, среди которых наиболее часто используется метод пошагового включения независимых переменных (факторов) в регрессионное уравнение: на первом шаге включается самый значимый фактор (имеющий самый высокий коэффициент детерминации с результирующей переменной), затем к нему добавляется тот из оставшихся факторов, который вместе с первым имеет наиболее высокое значение R^2 с результирующей переменной и т.д. Этот метод дает возможность увидеть "роль" каждого фактора в объяснении результата. Аналогично, метод последовательного исключения факторов, начиная со всех факторов, постепенно исключает из уравнения наименее значимые. Выбор этих моделей доступен на вкладке **Дополнительно** диалогового окна множественной регрессии.

результатов регрессионного анализа) для любых комбинаций значений факторных переменных (в частности, это хороший способ восстановления отсутствующих данных). Например, если подставить в уравнение регрессии число рабочих, равное 200, и мощность двигателей, равную 20, программа вычислит теоретическое значение дохода на таком предприятии: 640864 руб.

К прогнозу с использованием уравнения регрессии нужно относиться осторожно: строго говоря, его нельзя использовать вне диапазона значений признаков, по которому построена регрессионная модель. Если значения факторного признака выходят за пределы соответствующего диапазона значений, то нельзя исключить, что в новом интервале действуют другие закономерности и вид зависимости между данным фактором и результативным признаком может оказаться совсем иным.

4.2.3. Проверка значимости в регрессионном анализе

Рассмотрим некоторые результаты регрессионного анализа, связанные с проверкой гипотез.

Вернемся к рис. 4.15. Уравнению регрессии соответствует множественный коэффициент корреляции, равный 0,52. Рядом с ним на экране – две пока непонятных нам строчки: F = 198 и *сс* = 2 и 1057. Эти результаты относятся к *проверке гипотезы о линейной связи результативного и факторных признаков*, отвечающей на вопросы: существует ли связь?; значимо ли уравнение регрессии, используемое для отображения предполагаемой связи? На эти вопросы отвечает *статистический критерий значимости регрессии*.

В основе критерия значимости лежит идея разложения дисперсии результативного признака на две составляющие: *факторную* и *остаточную* дисперсии, т.е. объясненную регрессией часть дисперсии и часть, оставшуюся необъясненной в рамках регрессионной модели. Чем лучше объясняет поведение результативного признака регрессионная модель, тем выше доля факторной составляющей и ниже доля остаточной в общей дисперсии этого признака (кстати, значение множественного коэффициента детерминации как раз и равно отношению факторной, т.е. объясненной, дисперсии результативного признака к его общей дисперсии). Поэтому мерой значимости регрессии служит отношение F факторной дисперсии к остаточной.

Эта величина (при некоторых предположениях, которые мы здесь не будем рассматривать) распределена по т.н. закону Фишера F с m и $n–(m+1)$ *степенями свободы* (*сс*). Величина m – это число факторов, а n – число объектов (в нашем случае m = 2 и $n–(m+1)$ = 1057). Распределение Фишера, так же, как и нормальное распределение, хорошо изучено и табулировано, и

для каждого значения *F* можно найти соответствующую вероятность. Если значение этой вероятности (*p*) окажется меньше принятого уровня значимости, гипотеза об отсутствии линейной связи между результативным признаком и факторными отклоняется и регрессия признается значимой, что и произошло в приведенном примере, где вероятность полученного значения *F* (198) практически равна нулю.

Кроме проверки значимости регрессии в целом, STATISTICA дает возможность проверки гипотезы об отсутствии связи между результативным и каждым из факторных признаков. Такая гипотеза означает, что ненулевые значения регрессионных коэффициентов обусловлены лишь случайностями выборки, а в генеральной совокупности все коэффициенты этого уравнения равны нулю. Статистический критерий для проверки гипотезы об отсутствии влияния факторных признаков на результативный основан на подсчете стандартных ошибок коэффициентов регрессии и сравнении коэффициентов со своими ошибками. Гипотеза об отсутствии влияния факторного признака (о его незначимости) отклоняется в случае, если соответствующий коэффициент регрессии значительно превышает свою стандартную ошибку, и вероятность, соответствующая этой величине, меньше принятого уровня значимости

Статистическая характеристика этого критерия – уже знакомая нам величина *t*, распределенная по закону Стьюдента. Поэтому в таблице, приведенной на рис. 4.15, вы видите значения ошибок, вычисленные для них значения *t* и соответствующую каждому вероятность (4-й, 5-й и 6-й столбцы). Видно, что для уровня значимости 5% все коэффициенты регрессии оказались значимыми, т.е. существенными.

Проверка значимости коэффициентов регрессии важна еще и потому, что коэффициенты регрессии, в отличие от коэффициентов корреляции, не имеют максимальных и минимальных значений, и их величины зависят от единиц измерения соответствующих признаков. Значит, сама по себе величина коэффициента регрессии никак не определяет силу влияния фактора на результат. Например, существенным может оказаться и небольшой коэффициент регрессии, если этот коэффициент значимый. Если же коэффициент незначимый, то независимо от его величины следует считать, что соответствующий фактор не оказывает реального влияния на результативный признак.

4.2.4. Корреляции в модели множественной регрессии

Подход, при котором изменения одного фактора рассматриваются при условии, что другой фактор "заморожен", неявно основан на предположении о независимости факторов. Это не всегда выполняется. Факторные пе-

ременные рассматриваются как независимые по отношению к результирующей переменной, но, вообще говоря, не являются таковыми по отношению друг к другу. Явление коррелированности факторов друг с другом называется *мультиколлинеарностью.* При наличии мультиколлинеарности надо избегать включения в регрессионную модель сильно взаимосвязанных признаков.

Представить эффект мультиколлинеарности можно, сравнивая обычные коэффициенты корреляции между результирующим признаком и каждым из факторов (рис. 4.9) со значениями стандартизованных коэффициентов регрессии (*beta*-коэффициентов – см. рис. 4.16).

Смысл *beta*-коэффициентов состоит в том, что они являются коэффициентами *частной корреляции* – меры влияния каждой отдельной независимой переменной на результирующую, "освобожденного" от перекрестных влияний других независимых переменных. То есть, частные коэффициенты корреляции показывают "чистое" влияние фактора на результативный признак при неизменных (фиксированных) значениях всех остальных факторов. Например, обычный коэффициент корреляции между признаками "произведено" и "рабочие" равен 0,52, а частный коэффициент корреляции между этими же признаками равен 0,48, т.е. полный коэффициент больше чистого, т.к. в этом случае к прямому влиянию числа рабочих на доход предприятия добавляется косвенное влияние на этот доход и второго фактора (мощность двигателей), поскольку оба фактора имеют положительную взаимосвязь друг с другом. Косвенное влияние на результативный признак второго фактора через первый завышает корреляцию между первым фактором и результативным признаком.

ВОПРОСЫ

1. Что показывает диаграмма рассеяния?
2. Смысл коэффициента регрессии.
3. Почему эмпирические точки отклоняются от теоретической линии регрессии?
4. Когда уравнение регрессии можно использовать для прогноза?
5. Что такое коэффициент детерминации?
6. В чем отличается интерпретация коэффициентов корреляции и регрессии?
7. Смысл коэффициента корреляции.
8. Найти коэффициент корреляции по следующим данным:

x	1	2	3	4
y	40	30	20	10

9. Найти коэффициент корреляции по следующим данным:

x	1	2	3	4	5	6	7

y	5	6	6	7	8	8	9

Мог ли этот коэффициент появиться в выборке из некоррелированной генеральной совокупности?

10. В каких границах заключен коэффициент корреляции?
11. Какие значения r соответствуют тесной связи?
12. Может ли значение $r = 0$ говорить об отсутствии связи?
13. Что такое частная корреляция?
14. Смысл коэффициента множественной корреляции.
15. Как проверить значимость коэффициентов корреляции и регрессии?
16. Выборочная ошибка коэффициента корреляции.
17. Как проверить линейность связи?
18. По данным о 10000 объектов оказалось, что между двумя признаками имеется отрицательная корреляция (r = –0,0796). Существенна ли эта корреляция? (Иными словами, не могла ли она возникнуть в результате случайной выборки из некоррелированной совокупности?)
19. Начертить диаграмму, показывающую, как вместе с изменением r изменяется σ_r для выборок из а)100; б)10000 объектов.
20. В выборке 60 объектов получен коэффициент корреляции, равный 0,68. Может ли это значение быть статистически незначимым?

ЗАДАНИЯ

1. По данным файла Industry найти коэффициенты детерминации для зависимости объема производства от числа рабочих и суммарной мощности двигателей в отраслях: а) металлообрабатывающей; б) пищевой.
   Объяснить, какой фактор играет большую роль в объяснении объема производства для каждой из этих отраслей.
2. По данным файла Industry построить регрессионные зависимости объема производства от мощности двигателей для двух отраслей (см. предыдущее задание). Сравнить их между собой.
3. По данным файла Industry подсчитайте коэффициенты корреляции между объемом производства и мощностью двигателей для всех отраслей. Оцените значимость полученных коэффициентов. В каких отраслях корреляция выше?
4. По данным файла Estates построить уравнение регрессии для зависимости дохода от размера имения. Найти коэффициент корреляции.
5. По данным файла Tambov найти социально-экономические факторы, которые наиболее сильно взаимосвязаны с долей голосующих за партию а) эсеров и б) кадетов (построить прямоугольную таблицу коэффициентов корреляции между двумя этими признаками и группой социально-экономических факторов).

ГЛАВА 5

АНАЛИЗ ВЗАИМОСВЯЗЕЙ КАЧЕСТВЕННЫХ ДАННЫХ

В предыдущей главе мы рассматривали методы анализа взаимосвязей для признаков, принимающих только числовые значения – количественных признаков. Теперь рассмотрим методы анализа взаимосвязей для качественных признаков, т.е. признаков, значениями которых выступают категории.

5.1. ТИПЫ КАЧЕСТВЕННЫХ ДАННЫХ

Напомним (см. введение), что качественные (или категориальные) данные делятся на два типа: ранговые и номинальные.

Ранговые данные представлены категориями, для которых так же, как в случае интервальных или дискретных для количественных данных, можно указать порядок, т.е. категории сравнимы по принципу "больше-меньше" или "лучше-хуже".

Примеры ранговых переменных:

- Оценки на экзаменах имеют явно выраженную ранговую природу и выражаются категориями типа: "отлично", "хорошо", "удовлетворительно" и т.д.
- Уровень образования может быть представлен как набор категорий: "высшее", "среднее" и т.п.

Несомненно, мы можем ввести ранговую шкалу и с ее помощью упорядочить всех людей, для которых мы знаем их уровень образования или балл на экзамене. Однако, верно ли, что оценка "хорошо" на столько же хуже, чем "отлично", насколько оценка "удовлетворительно" хуже, чем "хорошо"? Несмотря на то, что формально, в случае с оценками, можно получить разницу в баллах, вряд ли корректно измерять расстояние от "отличника" до "хорошиста" пользуясь теми же правилами, что для расстояния от Москвы до Петербурга. В случае с уровнем образования особенно отчетливо видно, что простые вычисления невозможны, поскольку не существует единого правила вычитания "среднего" уровня образования из "высшего", даже, если мы присвоим высшему образованию код "3", а среднему – код "2".

Номинальные данные еще в меньшей степени могут быть сравнимы с числами, поскольку они представлены категориями, для которых порядок

абсолютно не важен. Для них не определен никакой другой способ сравнения, кроме как на буквальное совпадение/несовпадение.

Примеры номинальных переменных:

- Национальность: англичанин, белорус, немец, русский, японец и пр.
- Род занятий: служащий, врач, военный, учитель и т.д.
- Профиль образования: гуманитарное, техническое, медицинское, юридическое и т.д.

Если в случае с уровнем образования мы еще могли сравнивать людей в терминах "лучше-хуже" или "выше-ниже", то теперь мы лишены даже этой возможности; единственный корректный способ сравнения – это говорить, что данные персоналии "все являются историками", или "все не являются юристами".

Однако своеобразие качественных данных не означает, что их нельзя анализировать с помощью математических и статистических методов.

5.2. ВЗАИМОСВЯЗЬ РАНГОВЫХ КАЧЕСТВЕННЫХ ДАННЫХ

Ряд объектов, упорядоченных в соответствии со степенью проявления некоторого свойства, называют ранжированным, каждому числу такого ряда присваивается **ранг**. Будем обозначать ранги порядковыми числами: 1,2,...,n, где n – количество объектов. Таким образом, если какой-либо объект после ранжирования занимает третье место в ряду, ему присваивается ранг 3.

Меры взаимосвязи между парой признаков, каждый из которых ранжирует изучаемую совокупность объектов, называются в статистике коэффициентами **ранговой корреляции**. Эти коэффициенты строятся на основе следующих трех свойств:

если ранжированные ряды по обоим признакам полностью совпадают (т.е. каждый объект занимает одно и то же место в обоих рядах), то коэффициент ранговой корреляции должен быть равен +1, что означает полную положительную корреляцию;

если объекты в одном ряду расположены в обратном порядке по сравнению со вторым, коэффициент равен –1, что означает полную отрицательную корреляцию;

в остальных ситуациях значения коэффициента заключены в интервале [–1, +1]; возрастание модуля коэффициента от 0 до 1 характеризует увеличение соответствия между двумя ранжированными рядами.

Указанными свойствами обладают коэффициенты ранговой корреляции Спирмена ρ и Кендалла τ.

Не приводя формул для этих коэффициентов, отметим, что коэффициент Кендалла дает более осторожную оценку корреляции, чем коэффициент Спирмена (числовое значение τ всегда меньше, чем ρ).

При ранжировании объектов нередко возникает ситуация, когда два (или больше) объектов получают одинаковые ранги (такие объекты называют *связанными*). В этом случае значение ранга связанных объектов берется равным среднему значению тех рангов, которые имели бы эти объекты, если они были различны. Например, если связанными оказались 3-й и 4-й объекты в ранжированном ряду, то каждому из них приписывается ранг 3,5; если связываются все объекты от 2-го до 6-го, то каждый получает ранг (2+3+4+5+6)/5 = 4. Если число связанных рангов невелико, то это не сильно влияет на значения коэффициентов ранговой корреляции.

Пример 5.1. Создайте в программе STATISTICA файл на основании данных об участниках революционного движения в России во второй половине XIX в., приведенных на рис. 5.1 [1]. В первой половине таблицы дается распределение участников по возрасту, а во второй – распределение по сословиям.

Категория	% участников	Ранг категории	Ранг участия
Возраст			
До 20 лет	37,00	1	2
21-25	45,20	2	1
26-30	12,50	3	3
31-35	2,80	4	4
36-40	1,40	5	5
41-45	1,00	6	6
46-50	0,07	7	7
Более 50	0,03	8	8
Сословие			
Дворяне	30,00	1	1
Духовенство	22,00	2	2
Почетные граждане	9,00	3	4
Купцы	6,00	4	5
Военные (недворяне)	5,00	5	6,5
Мещане	14,00	6	3
Крестьяне	5,00	7	6,5

Рис. 5.1. Исходные данные

После создания файла начнем анализ с верхней половины таблицы. Столбец 1 содержит данные о проценте участников, принадлежащих соот-

[1] Миронов Б.Н., Степанов З.В. Историк и математика. Л., 1975. С. 134, 136.

ветствующей возрастной группе. Довольно очевидно, что существует связь между возрастом и участием в революционном движении – процент участия выше для более молодых людей. Эта связь особенно очевидна, если обратиться к столбцам 2 и 3, в которых стоят ранги выделенных групп по возрасту и степени участия в революционном движении, соответственно. Эти столбцы по существу различаются только в первых двух позициях, т.е. связь почти максимальная, причем положительная. Измерим степень этой связи, пользуясь коэффициентами ранговой корреляции.

Обратимся к модулю **Непараметрическая статистика** программы STATISTICA. На рис. 5.2 представлено основное диалоговое окно этого модуля. В этом окне требуется выбрать процедуру подсчета ранговых корреляций – **Корции Спирмена, тау Кендалла, гамма.**

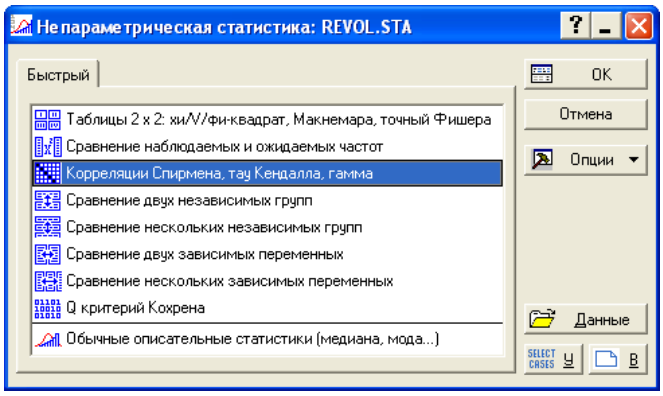


Рис. 5.2. Основное диалоговое окно модуля Nonparametric Statistics

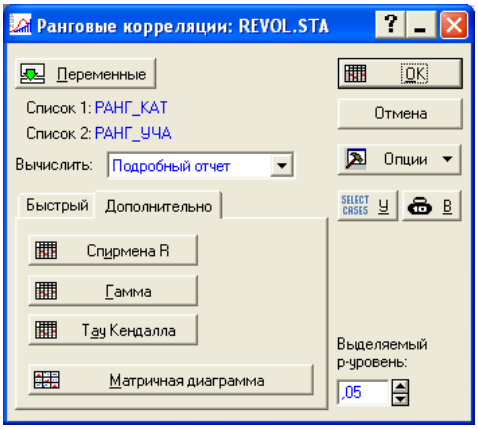


Рис. 5.3. Выбор переменных и коэффициента ранговой корреляции

Щелчок по графической кнопке **OK** открывает следующее диалоговое окно, где надо выбрать признаки для подсчета нужных корреляций и один из коэффициентов ранговой корреляции. Вкладка **Быстрый** дает возможность подсчета только коэффициента Спирмена, а вкладка **Дополнительно** предлагает на выбор несколько коэффициентов (см. рис. 5.3). Выберем, например, коэффициент Кендалла τ (**Тау Кендалла**) и получим результат, представленный на рис. 5.4.

На рис. 5.4 видно, что значение коэффициента Кендалла почти равно единице, т.е. налицо сильная связь между признаками. На этом же экране приводится величина статистической характеристики для проверки значимости коэффициента Кендалла (столбец Z), которая показывает, во сколько раз значение этого коэффициента превышает свою стандартную ошибку. В столбце p-уровень показана вероятность получения в выборке значений коэффициента Кендалла, которые по величине не ниже данного, при условии равенства нулю этого коэффициента в генеральной совокупности. Полученная вероятность очень мала и, значит, можно считать полученный коэффициент значимым, т.е. гипотеза о статистической независимости двух признаков должна быть отклонена.

	Тау корреляции Кендалла (REVOL.STA) ПД попарно удалены Отмеченные корреляции значимы на уровне p <,05				
Пара перем.	**Число набл.**	Кендалла тау	Z	p-уров.	точный p 1-стор.
РАНГ_КАТ & РАНГ_УЧА	8	0,928571	3,216666	0,001297	p < .001

Рис. 5.4. Окно результатов подсчета ранговой корреляции

Примечание. Прикладные аспекты ранговой корреляции.

У читателя может возникнуть вопрос: зачем использовать ранговые корреляции, если в данном случае оба признака (и возраст, и процент участия) являются количественными. Действительно, если представить возраст каждой группы как среднее арифметическое границ соответствующего возрастного интервала, тогда между нашими признаками можно было бы подсчитать и стандартный коэффициент корреляции *r*. Здесь есть, однако, определенные трудности, поскольку данные о возрасте в виде интервального ряда уже не являются такими точными, как исходная информация на уровне индивидуумов. Например, какое число должно быть выбрано в качестве середины интервала "до 20 лет" или интервала "старше 50 лет"?

Существуют и другие ситуации, когда при определении силы связи двух количественных признаков целесообразно вычислять коэффициенты ранговой корреляции. Так, при существенном отклонении распределения одного из них (или обоих) от нормального распределения подсчет уровня значимости выборочного коэффициента корреляции *r* становится некорректным в

то время, как ранговые коэффициенты не сопряжены с такими ограничениями.

Другая ситуация такого рода возникает, когда связь двух количественных признаков имеет нелинейный характер. Далее, зачастую значения количественного признака отличаются у разных объектов во много раз (иногда это связано с ошибками в данных). При подсчете обычного коэффициента корреляции это может сильно исказить картину взаимосвязи.

Наконец, коэффициенты ранговой корреляции могут использоваться не только для анализа взаимосвязи двух ранговых признаков, но и при определении силы связи между ранговым и количественным признаками. В этом случае значения количественного признака упорядочиваются и им приписываются соответствующие ранги. Именно так обстоит дело в нижней части таблицы исходных данных: взаимосвязь сословия (ранговый признак) и процента участников революционного движения (количественный).

5.3. ВЗАИМОСВЯЗЬ НОМИНАЛЬНЫХ КАЧЕСТВЕННЫХ ДАННЫХ

В статистическом анализе существует ряд методов, позволяющих изучать взаимосвязи номинальных признаков. Одним из них – пожалуй наиболее популярным – является метод построения *таблиц сопряженности* или кросс-табуляция.

5.3.1. Таблицы сопряженности

Таблицей сопряженности называется прямоугольная таблица, по строкам которой указываются категории одного признака (например, разные социальные группы), а по столбцам – категории другого (например, партийная принадлежность). Каждый объект совокупности попадает в какую-либо из клеток этой таблицы в соответствии с тем, в какую категорию он попадает по каждому из двух признаков. Таким образом, в клетках таблицы стоят числа, представляющие собой частоты совместной встречаемости категорий двух признаков (число людей, принадлежащих конкретной социальной группе и входящих в определенную партию). В зависимости от характера распределения этих частот внутри таблицы можно судить о том, существует ли связь между признаками. Что означает связь между социальным статусом и партийной принадлежностью? В данном случае о наличии связи свидетельствовало бы наличие определенных политических пристрастий у членов разных социальных групп. Формально говоря, эта связь понимается как более частая (или наоборот, редкая) совместная встречаемость отдельных комбинаций категорий по сравнению с ожидаемой встречаемостью – ситуацией чисто случайного попадания объектов туда (например, более высокая доля крестьян в партии трудовиков, а дворян – в партии

кадетов, чем доли этих социальных групп во всей совокупности депутатов Думы). Таким образом, связь тем сильнее, чем больше расхождения между реальными и ожидаемыми (в случае полной независимости признаков) частотами совместной встречаемости категорий этих признаков.

Пример 5.2. Используя файл Duma.sta по депутатам 1 Государственной думы, пройдем всю последовательность шагов, необходимых для построения и анализа таблиц сопряженности в программе STATISTICA.

Перейдем в раздел **Таблицы сопряженности…** модуля **Основные статистики и таблицы**. Рис. 5.5 демонстрирует основное диалоговое окно этого раздела. Здесь две похожие по составу вкладки: **Сопряженности** и **Флаги и заголовки**, на рис. 5.5 показана вторая вкладка, которая занимается формированием простых таблиц сопряженности (или таблиц "с двумя входами" – **2-вход. таблиц**).

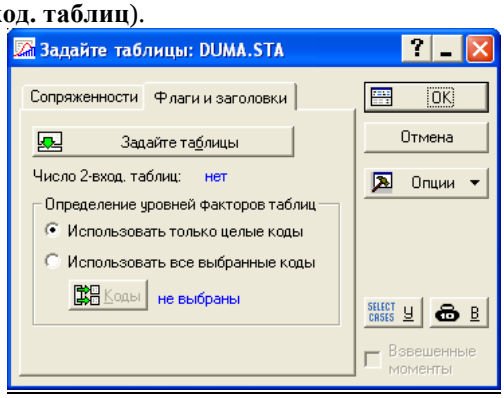


Рис. 5.5. Основное диалоговое окно для построения таблиц сопряженности

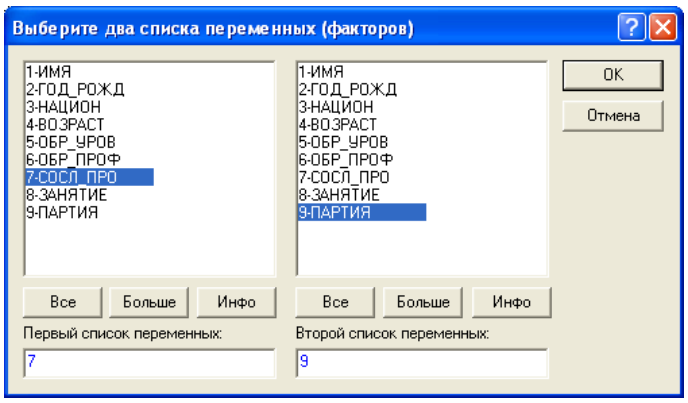


Рис. 5.6. Диалоговое окно выбора переменных для построения таблиц сопряженности

На рис. 5.5 видно, что в начале работы пока не задано ни одной такой таблицы (*Число 2-вход. Таблиц: = нет*). Графическая кнопка **Задайте таблицы** позволяет указать набор переменных, значения которых будут образовывать строки и столбцы будущей таблицы. Щелчок по графической кнопке **Задайте таблицы** открывает диалоговое окно с двумя списками всех переменных нашего файла, в которых надо сделать необходимый отбор признаков для построения кросстабуляции (см. рис. 5.6).

Выберем две переменные: "сословное происхождение" и "партия". Значения первой будут формировать строки таблицы сопряженности, а значения второй – столбцы. Щелчок по графической кнопке **OK** открывает диалоговое окно результатов кросстабуляции на экране. Блок **Таблицы** на вкладке **Опции** позволяет вычислять и выводить на экран разные виды таблиц (см. рис. 5.7):

- **Выделить частоты >** – выделять числа (частоты), большие некоторого "порогового" значения, указанного в соответствующей строке редактирования. Например, в нашем случае указано, что должны быть отмечены клетки таблицы, в которых абсолютная частота парной встречаемости больше 10.
- **Ожидаемые частоты** – одновременно с реальной таблицей сопряженности построить таблицу, содержащую ожидаемые частоты.
- **Остаточные частоты** – одновременно с реальной таблицей сопряженности построить таблицу, содержащую разности между реальными и ожидаемыми частотами.

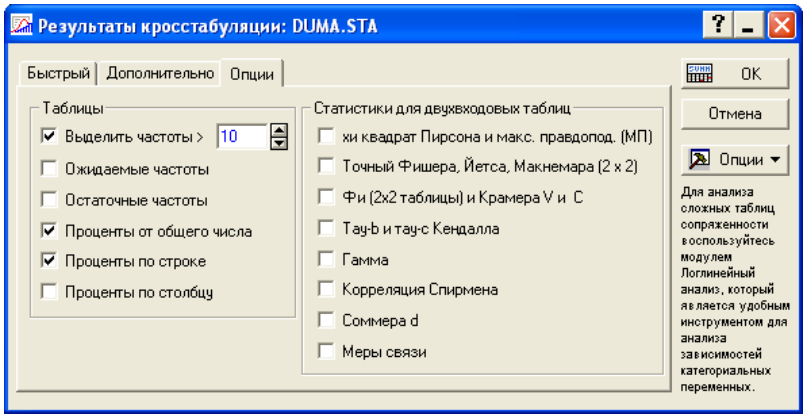


Рис. 5.7. Диалоговое окно для определения вида представления результатов кросстабуляции на экране

- **Проценты от общего числа** – кроме *абсолютных* частот парной встречаемости категорий, клетки таблицы сопряженности будут содержать *относительные* частоты в процентах к общему числу объектов.
- **Проценты по строке** – кроме *абсолютных* частот парной встречаемости категорий, клетки таблицы сопряженности будут содержать *относительные частоты* в процентах к сумме каждой строки. Например, доля трудовиков среди крестьян составляет почти 35%, а доля кадетов – менее 20% (см. табл. 5.1). Таким образом можно сравнивать распределение каждой социальной группы по различным политическим партиям.
- **Проценты по столбцу** – кроме *абсолютных* частот парной встречаемости категорий, клетки таблицы сопряженности будут содержать *относительные* частоты в процентах к сумме каждого столбца. Например, доля крестьян среди трудовиков составляет более 80%, в то время как доли всех остальных социальных групп – не более 5%. Таким образом можно сравнивать "социальный состав" различных политических фракций.

Если сделать поставить флажки **Проценты от общего числа** и **Проценты по строке** и щелкнуть по графической кнопке **OK**, мы получим результат, фрагмент которого приведен в табл. 5.1.

Таблица 5.1. Фрагмент таблицы сопряженности с абсолютными частотами, а также процентами от общей суммы и суммы каждой строки

	трудовик	кадет	беспар.	мирно-обновл.	соц.-дем.	польск. колон.	…	Totals
крестьянин	**65**	**37**	**59**	5	12	3	…	187
Row %	**34,76%**	**19,79**%	**31,55**%	2,67%	**6,42**%	1,60%	…	
Total %	**17,20**%	**9,79**%	**15,61**%	1,32%	**3,17**%	0,79%	**...**	49,47%
дворянин	3	**73**	3	**16**	3	**17**	…	133
Row %	2,26%	**54,89**%	2,26%	**12,03**%	2,26%	**12,78**%	…	
Total %	0,79%	**19,31**%	0,79%	**4,23**%	0,79%	**4,50**%	**...**	35,19%
купец	1	**12**	1	2	0	0	…	16
Row %	6,25%	**75,00**%	6,25%	12,50%	0,00%	0,00%	…	
Total %	0,26%	**3,17**%	0,26%	0,35%	0,00%	0,00%	**...**	4,23%
…	…	…	…	…	…	…	…	…
All Grps.	80	137	68	27	17	23	**...**	378
Total %	21,16%	36,24%	17,99%	7,14%	4,50%	6,08%	**...**	

Примечание. Полужирным шрифтом в таблице помечены клетки, в которых стоят частоты, большие 10 (это значение принято в пакете STATISTICA по умолчанию, но может быть изменено – см. флажок **Выделить частоты >** на рис.5.7.).

Можно получить также графическое представление таблицы сопряженности. В диалоговом окне результатов кросстабуляции на вкладках **Быстрый** и **Дополнительно** есть графическая кнопка **3М гистограмма** (*трех-*

мерная гистограмма), с помощью которой можно построить распределение парных частот встречаемости в пространстве выбранных переменных. Удобство такой гистограммы состоит в компактном отображении таблиц сопряженности больших размеров. Трехмерная гистограмма для нашего примера приведена на рис. 5.8. Она показывает, что наибольшую частоту совместной встречаемости имеют категории, расположенные в нижнем левом углу графика.

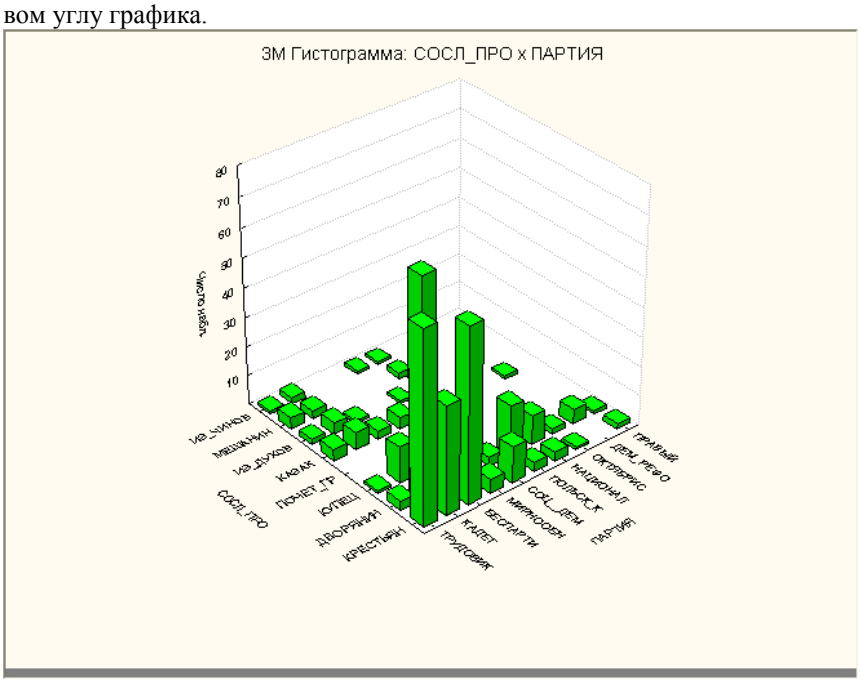


Рис. 5.8. Трехмерная гистограмма распределения частот парной встречаемости

5.3.2. Критерий значимости связи качественных признаков. (Проверка гипотезы о независимости признаков по таблице сопряженности [1])

Сравним процентные распределения по фракциям для всех депутатов и отдельных социальных групп (табл. 5.1). Например, рассмотрим фракцию

[1] В главе 3 можно прочитать общее изложение метода статистической проверки гипотез. Там же упоминается, что конкретные приложения метода будут рассматриваться в тех главах, где ставятся задачи проверки гипотез. Данная глава – один из таких случаев.

трудовиков: в нашей таблице среди всех 378 депутатов 80 человек (т.е. 21% – см. последнюю строку таблицы) принадлежат фракции трудовиков. Если бы распределение по фракциям не зависело от социального происхождения, то доля фракции трудовиков среди крестьян, дворян и т.д. была бы равна 21% численности каждой из этих групп.

Если бы это было так, то из общего числа крестьян в Думе (187 человек – см. последний столбец) 21%, или 40 человек "должны были" (по распределению частот) принадлежать фракции трудовиков. Если же взглянуть на реальное число трудовиков-крестьян, окажется, что их 65, т.е. больше, чем ожидалось. Если подсчитать, сколько дворян принадлежало бы этой партии, то окажется, что это 28 человек (т.е. 21% от общего числа дворян, общее число которых в Думе равно 133). В действительности же дворян–членов фракции трудовиков было всего трое, т.е. значительно меньше, чем ожидалось.

Такие же сравнения можно провести и для остальных фракций. Доверим эту работу программе Statistica, которая покажет, как выглядела бы таблица сопряженности при независимости признаков.

Для того, чтобы увидеть всю таблицу ожидаемых частот в случае, если два изучаемых признака независимы, надо в диалоговом окне результатов на вкладке **Опции** (рис. 5.7) включить флажок **Ожидаемые частоты**. Фрагмент таблицы ожидаемых частот приводится в табл. 5.2.

Таблица 5.2. Фрагмент таблицы ожидаемых частот

	трудовик	кадет	беспар.	мирно-обновл.	соц.-дем.	польск. колон.	…	Totals
крестьянин	39,58	67,78	33,64	13,36	8,41	11,38		187
дворянин	28,15	48,20	23,93	9,50	5,98	8,09		133
купец	3,39	5,79	2,88	1,14	0,72	0,97		16
…	…	…	…	…	…	…	…	…
All Grps	80	137	68	27	17	23		378

Проверим гипотезу о независимости фракционной принадлежности депутатов от социального происхождения, сравнивая между собой две таблицы (5.1 и 5.2). Для того чтобы понять, существенно ли они отличаются друг от друга, надо просуммировать по всем клеткам расхождения между реальными и ожидаемыми частотами (точнее – квадраты разностей между ними). Очевидно, если суммарное расхождение равно нулю, то нет оснований отвергнуть гипотезу о независимости признаков. И наоборот – чем больше суммарное расхождение, тем меньше вероятность принятия этой гипотезы.

Таким образом, статистической характеристикой гипотезы является сумма квадратов разностей реальных и ожидаемых частот по всем клеткам таблицы сопряженности. Известно, что эта величина имеет т.н. распределе-

ние X^2 (см. гл. 3) и для каждого ее значения известна вероятность того, что значения не меньше данного могут быть получены в выборке при том, что в генеральной совокупности $X^2 = 0$. Если вы обратите внимание на заголовок таблицы ожидаемых частот, то увидите значение X^2 (Хи-квадрат Пирсона), которое равно 193,71, и соответствующее ему значение вероятности (p = 0,0000), которое практически равно нулю. Это свидетельствует о том, что вероятность случайно получить в столь высокое значение X^2 при независимости признаков чрезвычайно мала. Значит, связь между признаками является значимой, а гипотеза о независимости признаков должна быть отклонена.

5.3.3. Коэффициенты взаимосвязи качественных признаков

Итак, значимая величина X^2 является свидетельством связи между двумя признаками. Как же измерить силу этой связи? Ясно, что при отсутствии связи $X^2 = 0$, и это значение является минимальным. Когда связь между признаками является максимально сильной, т.е. каждому значению (категории) одного признака в точности соответствует определенная категория другого признака, мы не можем заранее сказать, каким будет значение X^2, т.к. эта величина не имеет общего для всех таблиц максимального значения. Более того, поскольку значение X^2 зависит от числа степеней свободы, т.е. от количества строк и столбцов таблицы сопряженности, невозможно сравнивать между собой такие значения для таблиц с разным числом строк и столбцов.

Таким образом, необходимо построение коэффициента, который имел бы определенный максимум в случае максимальной связи и позволял бы сравнивать между собой разные таблицы по силе связи между признаками. Существует много коэффициентов, которые удовлетворяют этому условию. Рассмотрим один из них – **коэффициент Крамера *V*.**

Базируясь на значении критерия хи-квадрат, коэффициент Крамера позволяет измерять силу связи между двумя категоризованными переменными – измерить ее числом, принимающим значения от 0 до 1, т.е. от полного отсутствия связи до максимально сильной связи. Коэффициент позволяет сравнить зависимости разных признаков, с тем, чтобы выявить более и менее сильные связи.

Пример 5.3. Выберем для анализа другую пару качественных признаков из файла данных о депутатах 1 Государственной думы (Duma.sta): "занятие" – "профиль образования". Для вычисления коэффициента Крамера в блоке **Статистики для двухвходовых таблиц** на вкладке **Опции** окна результатов кросстабуляции (рис. 5.7) надо включить флажок с названием **Фи (2x2 таблицы) и Крамера V и C,** который дает возможность вычисления

коэффициента Крамера. Затем надо перейти на вкладку **Дополнительно** и щелкнуть по графической кнопке **Подробные двухвходовые таблицы**. Полученный результат представлен на рис. 5.9. На этом рисунке, наряду с уже знакомой величиной X^2, приводятся несколько других, в том числе и коэффициент Крамера, равный в данном случае 0,45.

	Статистики: ЗАНЯТИЕ(12) x ОБР_ПРОФ(12)		
Статистика	**Хи-квадрат**	ст.св.	p
Пирсона Хи-квадрат	818,6710	cc=121	p=0,0000
М-П Хи-квадрат	548,8777	cc=121	p=0,0000
Фи	1,485483		
Коэфф.сопряженности	,8295475		
Крамера V	,4478900		

Рис. 5.9. Взаимосвязь признаков "занятие" и "профиль образования"

Сравним полученный результат с коэффициентом взаимосвязи для пары признаков "занятие" – "сословное происхождение". Результат для этой пары представлен на рис. 5.10.

	Статистики: ЗАНЯТИЕ(12) x СОСЛ_ПРО(8)		
Статистика	**Хи-квадрат**	ст.св.	p
Пирсона Хи-квадрат	393,4427	cc=77	p=0,0000
М-П Хи-квадрат	367,1385	cc=77	p=0,0000
Фи	1,027038		
Коэфф.сопряженности	,7164748		
Крамера V	,3881837		

Рис. 5.10. Взаимосвязь признаков "занятие" и "сословное происхождение"

Оба результата показывают значимые взаимосвязи признака "занятие" как с признаком "профиль образования", так и с признаком "сословное происхождение". Однако из сравнения двух коэффициентов Крамера мы получаем, что для определения рода занятий депутатов большую роль играло не сословное происхождение, а профиль образования.

5.3.4. Бинарные признаки. Четырехклеточные таблицы

Кратко охарактеризуем частный случай номинальных признаков – бинарные признаки, число категорий у которых равно двум; иногда их называют альтернативными. К бинарным относятся признаки, которые фиксируют у каждого объекта просто наличие или отсутствие некоторого качества или относят объект в одну из двух возможных категорий ("грамотный/неграмотный", "работает/не работает", "мужчина/женщина" и т.п.).

Взаимосвязь между парой бинарных признаков представляется таблицей сопряженности, состоящей всего из четырех клеток – четырехклеточ-

ной таблицей. Например, для пары признаков "грамотность" и "занятость" четырехклеточная таблица имеет вид:

	грамотный	неграмотный	
работает	*a*	*b*	*a*+*b*
не работает	*c*	*d*	*c*+*d*
	a+*c*	*b*+*d*	*a*+*b*+*c*+*d* = *n*

a, b, c и *d* – внутренние клетки таблицы, т.е. количества объектов, попадающих в указанные категории; *a*+*b, c*+*d, a*+*c* и *b*+*d* – итоги строк и столбцов, соответственно; сумма итогов строк (столбцов) равна числу объектов *n*. Поскольку бинарные признаки – лишь частный случай номинальных, то для работы с ними могут применяться все методы, о которых шла речь выше (вычисление X^2, подсчет коэффициента Крамера и т.д.). Кроме того, для таких признаков разработаны некоторые специальные методы измерения взаимосвязей. Упомянем здесь коэффициент φ, который можно легко подсчитать, даже не пользуясь компьютером, по формуле:

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

Коэффициент φ является аналогом коэффициента корреляции для случая четырехклеточных таблиц. Он имеет знак, и его значения находятся в интервале [–1, +1]. Значение +1 коэффициент принимает в случае, когда *b* = *c* = 0; значение –1, когда *a* = *d* = 0. Если признаки статистически независимы, φ = 0.

Вычисление коэффициента φ (точнее φ^2) можно выполнить на своеобразном калькуляторе в модуле **Непараметрическая статистика**, раздел **Таблицы 2 x 2**..., в котором значения *a, b, c* и *d* непосредственно вводятся в клетки таблицы [1] и сразу выдается результат, включающий не только исходные данные, но и относительные частоты (проценты), а также ряд коэффициентов взаимосвязи, среди которых находится и коэффициент φ.

Пример 5.4. Введем в клетки таблицы значения, взятые нами из файла Edu_1897 (грамотность населения по данным переписи 1897 г.) для дворян м.п.: по столбцам – грамотные/неграмотные, по строкам – место проживания (город/уезд) (см. рис. 5.11). То есть, 422 (тыс. чел.) – число грамотных дворян м.п. в городах, 228 – число грамотных в уездах; 77 и 160 – соответственно число неграмотных в городах и уездах.

[1] К сожалению, в программе STATISTICA не предусмотрен ввод готовых таблиц сопряженности, а требуется ввод исходных данных, из которых программа сама конструирует такие таблицы. Единственным исключением являются как раз четырехклеточные таблицы.

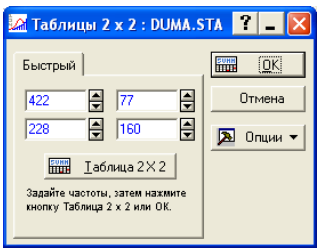


Рис. 5.11. Ввод данных четырехклеточной таблицы

Нажав графическую кнопку **OK**, получим результаты, среди которых φ^2 (*Фи-коэффициент*) со значением, равным 0,08370, т.е. абсолютная величина φ примерно равна 0,3. Это говорит о наличии определенной, хотя и не очень тесной связи между грамотностью и местом проживания. В той же таблице результатов приводится уже известный нам критерий связи X^2. Величина X^2 говорит о значимости связи, т.к. соответствующая ему величина вероятности *p* очень мала.

Завершая рассмотрение проблем анализа взаимосвязей, отметим следующие существенные положения.

1. Величины мер связи признаков различной природы не сравнимы между собой.
2. Если данные содержат признаки разной природы, то для сопоставления силы связи между любой парой признаков обычно используют меры зависимости, пригодные для номинального уровня измерения. Такой подход позволяет анализировать в комплексе все связи. При этом, однако, следует учитывать, что возникают определенные потери исходной информации, ее "огрубление". Так, для ранговых признаков теряется информация о соответствующем упорядочении объектов, а значения количественных признаков группируются в интервалы, которые при переводе на номинальный уровень измерения также оказываются неупорядоченными. Иногда такое огрубление полезно, поскольку позволяет количественные данные с грубыми ошибками трактовать как ранговые или даже номинальные. Уменьшение точности при этом компенсируется повышением надежности данных.

ВОПРОСЫ

1. Типы качественных признаков.
2. Чем качественные признаки отличаются от количественных?
3. В анкете имеются следующие пункты: фамилия, национальность, пол, возраст, образование, должность, зарплата. Указать, к каким категориям принадлежат эти признаки.
4. Привести примеры ранговых признаков.
5. Свойства коэффициентов ранговой корреляции.
6. Что такое дробные ранги?

7. Можно ли использовать коэффициенты ранговой корреляции при работе с количественными признаками?
8. Какой из двух ранговых коэффициентов, ρ или τ, дает более осторожную оценку связи?
9. Что такое номинальные признаки? Примеры.
10. Чем альтернативные признаки отличаются от неальтернативных?
11. Четырехклеточная таблица.
12. Свойства коэффициента φ.
13. Таблица сопряженности.
14. Можно ли говорить о знаке связи для неальтернативных номинальных признаков?
15. В чем смысл критерия X^2?
16. Свойства коэффициента Крамера V.
17. Какие меры связи можно использовать при изучении признаков разной природы?

ЗАДАНИЯ

1. Используя файл Revol.sta, измерить силу связи между сословием и участием в революционном движении, т.е. подсчитать коэффициенты Спирмена и Кендалла для строк 12–18.
2. Для файла Trade.sta подсчитать силу связи между населением страны и объемом внешней торговли с помощью:
   а) коэффициентов ранговой корреляции;
   б) обычных коэффициентов корреляции.
   Сравнить их значимость.
3. Используя файл Duma.sta, построить таблицы сопряженности для:
   а) уровня образования и партийной принадлежности;
   б) профиля образования и партийной принадлежности;
   в) профиля образования и сословного происхождения
   Объяснить, насколько значимо в каждом случае отклонение реальных частот парной встречаемости от ожидаемых.
4. Используя файл Duma.sta, определить, с какой переменной теснее связана партийная принадлежность: а) с уровнем образования; б) с профилем образования.
5. Используя файл Duma, построить график распределения частот парной встречаемости для переменных:
   а) уровень образования и партийная принадлежность;
   б) профиль образования и партийная принадлежность;
6. Рассчитать коэффициент четырехклеточной корреляции для признаков "грамотность" и "место проживания" по данным файла Edu_1897:

а) для лиц м.п. разных сословий (4 таблицы);

б) для лиц обоего пола разных сословий (еще 4 таблицы).

Выяснить, в каком случае связь наибольшая.

7. Даны ранги группы учащихся по их математическим (I) и музыкальным (II) способностям:

I	1	2	3	4	5	6	7	8	9	10
II	6	5	1	4	2	7	8	10	3	9

Найти значение коэффициента ранговой корреляции Кендалла (τ).

8. Дана статистика отношения к правилам уличного движения в течение месяца для мужчин и женщин:

	Мужчины	Женщины	
Нарушали	20	0	20
Соблюдали	30	50	80
	50	50	100

Сравнить значения коэффициента взаимосвязи, сделать вывод.

9. Определить, являются ли признаки А и В взаимонезависимыми или характеризуются положительной либо отрицательной взаимосвязью в следующем случае:

N = 5000; (A) = 2450; (B) = 3000; (AB) = 1600.

10. В цехе имеется 100 квалифицированных рабочих, из которых 20 перевыполняют норму. Из этих 20 рабочих 15 имеют среднее специальное образование, а из остальных 80 рабочих такое образование имеют 20 человек. Определить коэффициент связи между перевыполнением нормы и наличием среднего специального образования.

ЧАСТЬ III

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

ГЛАВА 6

МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ

При изучении массовых исторических источников исходная информация часто может быть представлена в виде набора объектов, каждый из которых характеризуется рядом признаков (показателей). В качестве объектов могут выступать хозяйства, поселения, административно-территориальные единицы и т.д., а в качестве признаков – различные показатели социально-экономической или демографической структуры изучаемых объектов.

Как показывает опыт анализа массовых источников, число объектов может достигать многих десятков и сотен; число признаков также может исчисляться десятками. Очевидно, непосредственный (визуальный) анализ матрицы данных при большом количестве объектов и признаков практически малоэффективен – можно лишь выявить отдельные особенности изучаемой структуры, извлечь иллюстративные, частные примеры. В этих условиях возникают задачи укрупнения, концентрации исходных данных, т.е. построения обобщенных характеристик множества признаков и множества объектов. Решение этих задач может осуществляться с помощью современных методов многомерного статистического анализа. При этом методы, ориентированные на анализ структуры множества признаков и выявление обобщенных факторов, известны как методы **факторного анализа**, а методы анализа структуры множества объектов образуют совокупность методов **многомерной классификации.**

6.1. КЛАСТЕРНЫЙ АНАЛИЗ

Традиционный метод построения типологии сводится обычно к группировке изучаемых объектов на основе одного (двух – трех) признаков.

Важно отметить, что традиционные приемы типологической группировки направлены на выявление качественно однородных групп объектов путем определения границ интервалов на оси одного из группообразующих признаков; эти приемы носят неформальный характер и осуществляются на основе содержательных концепций и опыта предшествующих исследований.

Современный уровень развития методов многомерного количественного анализа и компьютерные технологии позволяют осуществлять классификацию на более широкой и объективной основе – с учетом всех существенных структурно-типологических признаков и характера распределения объектов в заданной системе признаков. Такая классификация производится на основе стремления собрать в одну группу в некотором смысле схожие объекты, причем так, чтобы объекты из разных групп были по возможности несхожими. Такие методы получили название **методов многомерной классификации** (кластерного анализа, таксономии).

6.1.1. Агломеративно-иерархический метод

Будем считать, что все m признаков измерены в количественной шкале. Тогда каждый из n объектов может быть представлен точкой в m-мерном пространстве признаков. Характер распределения этих точек в рассматриваемом пространстве определяет структуру сходства и различия объектов в заданной системе показателей. О сходстве объектов можно судить по расстоянию между соответствующими точками. Содержательный смысл такого понятия сходства означает, что объекты тем более близки, похожи в рассматриваемом аспекте, чем меньше различий между значениями одноименных показателей.

Для определения близости пары точек (объектов i и j) в многомерном пространстве в случае количественных признаков используется *евклидово расстояние*, равное корню квадратному из суммы квадратов разностей значений одноименных показателей, взятых для данной пары объектов:

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} \qquad (i, j = 1,2,...,n),$$

где d_{ij} – евклидово расстояние между i-м и j-м объектами, x_{ik} – значение k-го признака для i-го объекта.

Заметим, что расстояние между объектами зависит от "масштаба" признаков: признаки, диапазон значений которых (а он зависит от единиц измерения) велик, играют большую роль при вычислении расстояния между объектами в отличие от признаков, диапазон изменения которых мал. Например, расстояния, выраженные в километрах, будут в тысячу раз меньше, чем в метрах. По этой причине данные обычно нормализуют, т.е.

все признаки приводят к стандартному виду со средним значением, равным нулю, и стандартным отклонением, равным единице. Это достигается путем вычитания из исходных значений каждого признака его среднего арифметического значения и деления полученной разности на исходное стандартное отклонение. Тем самым объекты на оси каждого признака сохраняют свое относительное положение, но "масштаб" измерения признаков становится единым. Практически во всех статистических пакетах предусмотрена возможность такого преобразования.

Подсчитав значения расстояний для всех пар объектов, получим квадратную матрицу D размером $n \times n$ (*матрицу расстояний*); эта матрица, очевидно, симметрична.

Матрица расстояний D служит основой при реализации *агломеративно-иерархического метода*, основная идея которого заключается в последовательном объединении группируемых объектов – сначала самых близких, а затем все более удаленных друг от друга. Процедура построения классификации состоит из последовательных шагов, на каждом из которых производится объединение двух ближайших групп объектов (*кластеров*) [1].

Существуют различные способы определения расстояний между кластерами (различающие методы кластерного анализа). Обычно близость двух кластеров определяется как среднее значение расстояния между всеми такими парами объектов, где один объект пары принадлежит к одному кластеру, а другой – к другому:

$$D_{pq}^2 = \sum_{i \in X_p} \sum_{j \in X_q} \frac{d_{ij}^2}{n_p n_q},$$

где D_{pq}^2 – мера близости между p-м и q-м кластерами; X_p – p-й кластер; X_q – q-й кластер; n_p, n_q – число объектов в p-м и q-м кластерах, соответственно.

На первом шаге процедуры агломеративно-иерархического метода кластерного анализа рассматривается начальная матрица расстояний между объектами, и по ней определяется минимальное расстояние; далее, наиболее близкие объекты, находящиеся друг от друга на этом расстоянии, объединяются в один кластер, в матрице вычеркиваются строка и столбец, соответствующие первому из этих объектов, а расстояния от нового кластера до всех остальных кластеров (на первом шаге – объектов) вычисляются по вышеприведенной формуле – как средние из расстояний от объектов первого кластера до всех остальных. Эти вновь вычисленные значения заносятся

[1] *Cluster* – скопление, “гроздь”, группа объектов, характеризующихся общими свойствами.

в строку и столбец матрицы расстояний, соответствующие второму объекту из первого кластера.

На втором шаге процедуры по матрице расстояний, уменьшенной на одну строку и один столбец, снова определяют минимальное расстояние и формируют новый кластер. Этот кластер может быть построен в результате объединения либо двух объектов, либо одного объекта с кластером, построенным на первом шаге. Далее, в матрице расстояний вычеркиваются одна строка и один столбец, а одна строка и один столбец пересчитываются и т.д.

Таким образом, иерархический метод кластерного анализа включает *n*–*1* аналогичных шагов. При этом после выполнения каждого шага число кластеров уменьшается на единицу, а матрица расстояний уменьшается на одну строку и один столбец. В конце этой процедуры получится один кластер, объединяющий все *n* объектов.

Результаты такой классификации часто изображают в виде **дендрограммы** (дерева иерархической структуры), содержащего *n* уровней, каждый из которых соответствует одному из шагов описанного процесса последовательного укрупнения кластеров [1].

Data: HARVEST.STA* (5v by 15c)

	Урожайность хлебов в России и других странах в 1913 г.				
	1 ПШЕНИЦА	2 РОЖЬ	3 ЯЧМЕНЬ	4 ОВЕС	5 КАРТОФЕЛ
РОССИЯ	55,000	56,000	62,000	63,000	491,000
АВСТРИЯ	89,000	92,000	107,000	94,000	602,000
ВЕНГРИЯ	88,000	82,000	92,000	91,000	470,000
ВЕЛИКОБР	149,000	0,000	127,000	117,000	1086,000
БОЛГАРИЯ	80,000	75,000	81,000	60,000	–
ГЕРМАНИЯ	157,000	127,000	148,000	146,000	1057,000
ГОЛЛАНДИ	160,000	122,000	168,000	148,000	1176,000
ИСПАНИЯ	52,000	61,000	64,000	45,000	–
РУМЫНИЯ	94,000	70,000	71,000	68,000	641,000
СЕРБИЯ	72,000	58,000	65,000	46,000	–
ФРАНЦИЯ	89,000	71,000	92,000	86,000	571,000
ШВЕЙЦАРИ	153,000	123,000	130,000	150,000	1038,000
ШВЕЦИЯ	161,000	94,000	139,000	123,000	–
КАНАДА	94,000	61,000	108,000	102,000	750,000
США	68,000	68,000	85,000	70,000	408,000

Рис. 6.1. Данные об урожайности в 15 странах мира в 1913 г. (файл Harvest.sta)

[1] Важным вопросов в кластерном анализе является выбор необходимого числа кластеров. В некоторых случаях это число может быть выбрано из априорных соображений, однако чаще оно определяется в процессе формирования кластеров, исходя из значений некоторых показателей их однородности и степени удаленности друг от друга (например, показателей внутригрупповой дисперсии или вариации).

Пример 6.1. Проведем кластерный анализ группы из 15 стран мира на основе данных об урожайности хлебов в 1913 г. (файл Harvest.sta – см. рис. 6.1).

Сначала выполним нормализацию признаков. Для того, чтобы не изменять исходные данные, прежде всего сохраним файл под другим именем, например Harvest_new.sta. Это можно сделать, обратившись к разделу **Файл** основного меню программы и выбрав команду **Сохранить как**... в меню этого раздела.

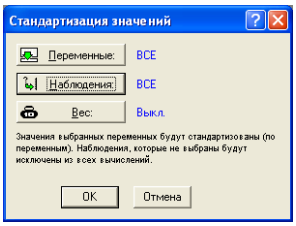


Рис. 6.2. Диалоговое окно нормализации исходных данных

Сохранив файл под новым именем, обратимся к разделу **Данные** главного меню и выберем команду **Стандартизовать.** При обращении к этой процедуре на экране появляется диалоговое окно, в котором надо указать, что преобразовываются все переменные для всех объектов (см. рис. 6.2). В результате все исходные данные будут заменены их нормализованными значениями. Не забудьте сохранить преобразованный файл (команда **Сохранить** раздела **Файл** основного меню программы STATISTICA).

Теперь перейдем к модулю **Многомерный разведочный анализ | Кластерный анализ**. В первом диалоговом окне (см. рис. 6.3) откроем соответствующий файл (знакомая вам графическая кнопка **Open Data**) с именем Harvest_new.sta, а также определим метод классификации. В программе STATISTICA предлагаются несколько методов; агломеративно-иерархическому методу в предлагаемом списке соответствует **Иерархическая классификация**. Выберем иерархическую классификацию объектов и щелкнем по графической кнопке **OK**.

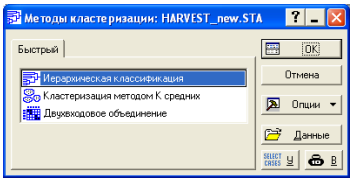


Рис. 6.3. Выбор метода кластеризации

На экране появится основное диалоговое окно агломеративно-иерархического метода с вкладками **Быстрый** и **Дополнительно**. Сразу обратимся к вкладке **Дополнительно** (см. рис. 6.4). Рассмотрим подробнее, какие параметры надо обязательно указать в этом окне, чтобы получить требуемый результат.

Прежде всего, как обычно, надо указать, какие переменные будут участвовать в анализе (графическая кнопка **Переменные** в верхнем левом

углу) – выберем первые 4 переменные (поскольку урожайность картофеля дана не для всех стран, последнюю переменную решено не включать в анализ). Очень важно правильно указать в окошке **Объекты** – строки, или наблюдения (см. рис. 6.4)**.**

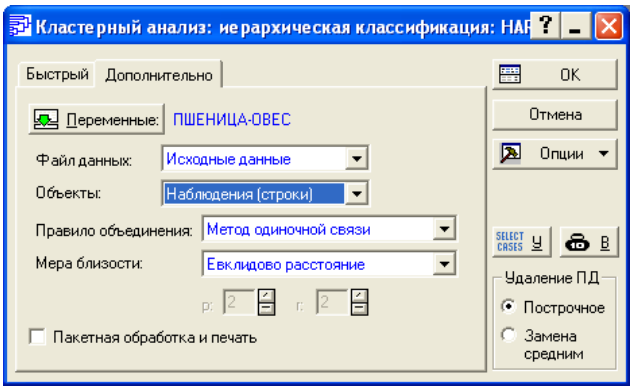


Рис. 6.4. Основное диалоговое окно агломеративно-иерархического метода. Выбор объектов классификации

По умолчанию объектами классификации являются столбцы, т.е. переменные, поэтому не забудьте поменять их на *строки*! В этом же диалоговом окне в окошке **Файл данных** можно выбрать вид представления данных: это либо **Исходные данные**, либо уже готовая **Матрица расстояний**. Так как наш файл содержит исходные данные, в этом окошке мы ничего не меняем.

В этом же диалоговом окне следует выбрать правило кластеризации, т.е. объединения групп объектов, которое можно задать в окошке **Правило объединения**. На рис. 6.5 показан список правил. Для большинства задач хорошо подходит метод объединения групп с наименьшим расстоянием между их центрами, которые определяются с учетом числа объектов в каждой группе. Это т.н. **Взвешенное попарное среднее**.

Наконец, в этом же диалоговом окне в окошке **Мера близости** надо указать правило вычисления расстояний между объектами и группами объектов. На рис. 6.6 можно видеть список различных мер для определения расстояний. Для количественных признаков обычно выбирается *Евклидово расстояние*; для качественных – абсолютное или относительное число признаков, значения которых у объектов не совпадают (*Процент несогласия*).

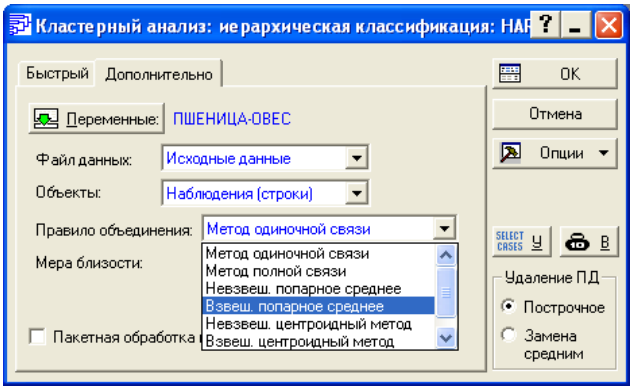


Рис. 6.5. Основное диалоговое окно агломеративно-иерархического метода. Выбор правила объединения

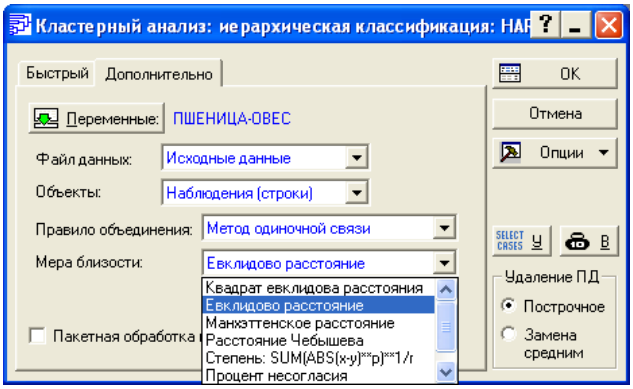


Рис. 6.6. Основное диалоговое окно агломеративно-иерархического метода. Выбор меры близости между объектами

После определения всех параметров классификации на экране появляется окно результатов иерархической классификации. В этом окне (рис. 6.7) можно выбрать возможности просмотра нескольких таблиц и графиков с результатами: **Горизонтальную** или **Вертикальную дендрограмму** – дерево иерархической классификации, **Матрицу расстояний** между объектами и др.

Обычно наибольший интерес представляет дендрограмма. Например, на рис. 6.8 приводится горизонтальная дендрограмма, иллюстрирующая классификацию для группы из 15 стран.

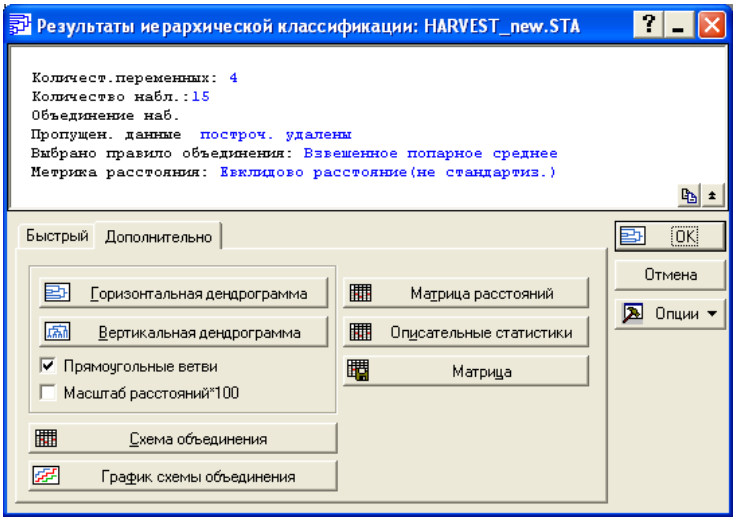


Рис. 6.7. Окно результатов кластерного анализа

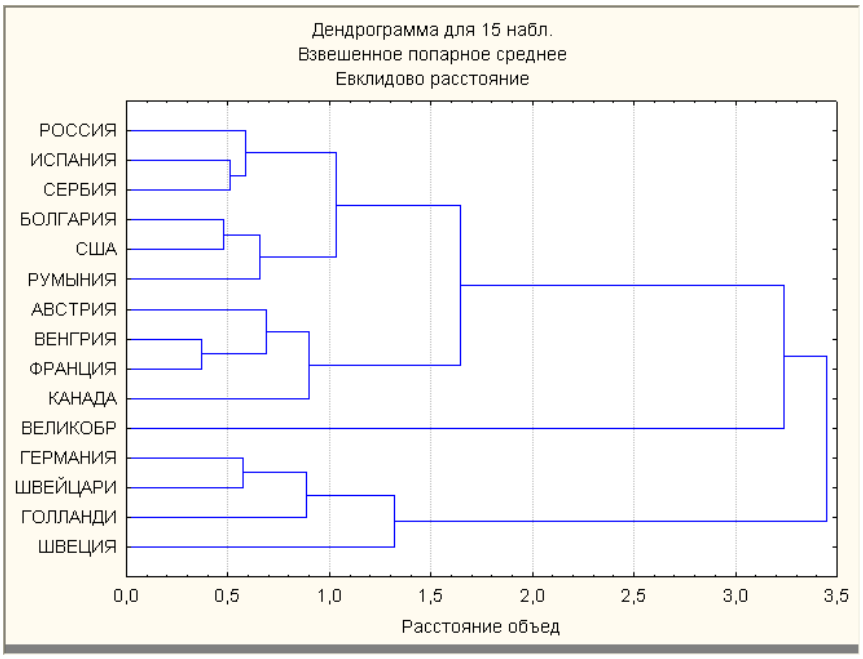


Рис. 6.8. Результаты кластер-анализа по 15 странам на основе данных об урожайности

Анализируя структуру полученной классификации, можно видеть, что объекты (страны) могут быть разделены на три группы (Германия, Швейцария, Голландия и Швеция); (Австрия, Венгрия, Франция и Канада); (Россия, Испания, Сербия, Болгария, США, Румыния); при этом можно детализировать структуру каждой группы. Так, третья группа состоит из двух подгрупп: Россия, Испания и Сербия входят в одну из них, а Болгария, США и Румыния – в другую. С другой стороны, вторая и третья группы на определенном уровне образуют общий кластер, тогда как первая группа остается достаточно далекой от этого кластера.

Может возникнуть вопрос: почему Великобритания не вошла ни в один из явно видных трех кластеров, но занимает особое положение на этой схеме? Если внимательно посмотреть на исходные данные, то причина такого результата отыщется довольно быстро: для Великобритании отсутствуют статистические данные об урожайности ржи, и поэтому в соответствующем столбце исходной таблицы стоит нулевое значение. Скорректируем параметры классификации, исключив из признаков, по которым проводится анализ, переменную "рожь". Повторим процедуру классификации и рассмотрим новый результат (рис. 6.9). На рисунке видно, что Великобритания больше не занимает особой позиции в схеме классификации, а входит в первый кластер. В остальном результаты практически не изменились.

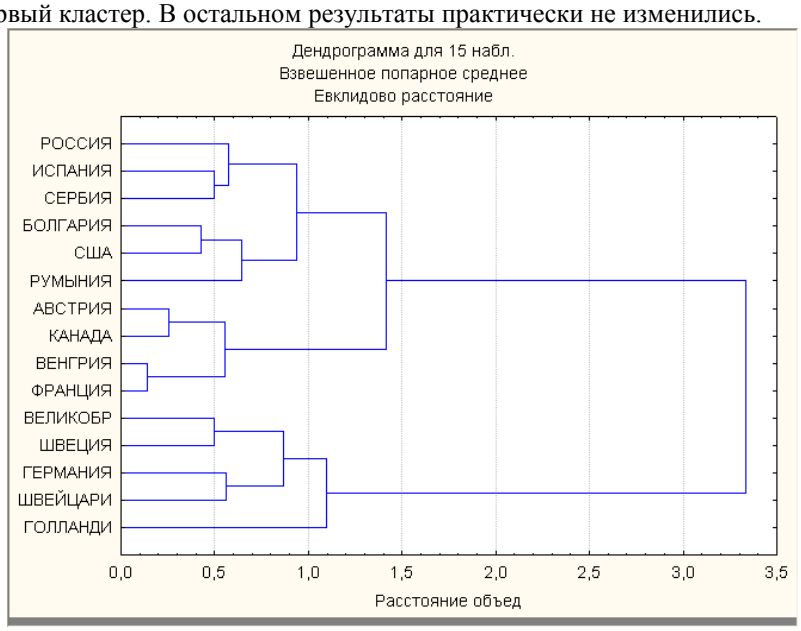


Рис. 6.9. Скорректированные результаты кластер-анализа по 15 странам на основе данных об урожайности

6.1.2. Метод *k*-средних

Другим методом кластерного анализа является т.н. *метод k-средних* [1]. В отличие от агломеративно-иерархического метода, который не требует предварительной оценки возможного числа групп объектов, этот метод основан на гипотезе о наиболее вероятном количестве классов. Задачей метода при этом является построение заданного числа кластеров, которые должны максимально отличаться друг от друга.

Процедура классификации начинается с построения заданного числа кластеров, полученных путем случайной группировки объектов. Затем следует итерационный процесс перемещения объектов между группами с целью минимизировать суммарную внутриклассовую дисперсию показателей и максимизировать межклассовую дисперсию (т.е. каждый кластер должен состоять из максимально "похожих" объектов, причем сами кластеры должны быть максимально "непохожими" друг на друга).

Результаты этого метода позволяют получить центры всех классов (а также и другие параметры дескриптивной статистики) по каждому из исходных признаков, а также увидеть графическое представление о том, насколько и по каким параметрам различаются полученные классы.

Пример 6.2. Вернемся к файлу с данными об урожайности и попробуем провести классификацию методом *k*-средних. Для предварительной оценки числа классов используем информацию, полученную при работе с агломеративно-иерархическим методом, который дает три хорошо заметных группы объектов.

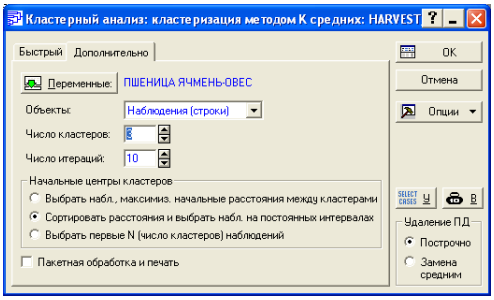


Рис. 6.10. Основное диалоговое окно метода k-средних. Выбор числа классов

На рис. 6.10 показано диалоговое окно кластеризации методом *k*-средних, вкладка **Дополнительно**. В этом окне самым существенным является поле **Число кластеров**, в котором следует указать число классов (в данном случае – 3). Объекты классификации (строки исходной таблицы) и переменные, участвующие в анализе, менять уже не надо, т.к. они были указаны при работе с агломеративно-иерархическим методом.

Щелкнув по графической кнопке **OK**, вы увидите окно результатов. Рассмотрим вкладку **Дополнительно** (рис. 6.11).

[1] Часто этот метод называют ISODATA.

Наиболее интересны в этом окне две графические кнопки: **Элементы кластеров и расстояния** и **График средних**. Первая из них позволяет получить в виде отдельной таблички перечень объектов для каждого из построенных классов. На рис. 6.12 показаны все три таблички одновременно. Если сравнить этот результат с предыдущим, видно, что получены те же самые три группы объектов.

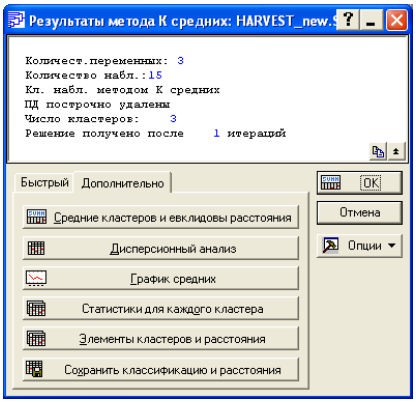


Рис. 6.11. Окно результатов метода k-средних

Теперь рассмотрим графическое представление этих кластеров (графическая кнопка **График средних**).

Видно, что по всем трем переменным, включенным в анализ (урожайность пшеницы, овса и ячменя) кластер номер 3, куда входят Великобритания, Германия, Голландия, Швейцария и Швеция, значительно превосходит остальные два кластера, т.е. урожайность в этих странах самая высокая (см. рис. 6.13).

	Элементы кластера номер 1 (HARVEST_new. и расстояния до центра кластера. Кластер содержит 4 набл.			
	АВСТРИЯ	ВЕНГРИЯ	ФРАНЦИЯ	КАНАДА
Расст.	0,126746	0,141681	0,177596	0,207720

	Элементы кластера номер 2 (HARVEST_new.STA) и расстояния до центра кластера. Кластер содержит 6 набл.					
	РОССИЯ	БОЛГАРИЯ	ИСПАНИЯ	РУМЫНИЯ	СЕРБИЯ	США
Расст.	0,280132	0,220045	0,363290	0,373999	0,231345	0,299166

	Элементы кластера номер 3 (HARVEST_new.STA) и расстояния до центра кластера. Кластер содержит 5 набл.				
	ВЕЛИКОБР	ГЕРМАНИЯ	ГОЛЛАНДИ	ШВЕЙЦАРИ	ШВЕЦИЯ
Расст.	0,425235	0,176417	0,480797	0,303605	0,238958

Рис. 6.12. Объекты, образующие три построенных программой кластера

Два других кластера ближе друг к другу (особенно по урожайности пшеницы) и все-таки по всем зерновым культурам кластер номер 1 (Австрия, Венгрия, Франция и Канада) превосходит кластер номер 2, в котором оказались страны с наиболее низкими показателями урожайности в 1913 г. (это Россия, Болгария, Испания, Румыния, Сербия и США).

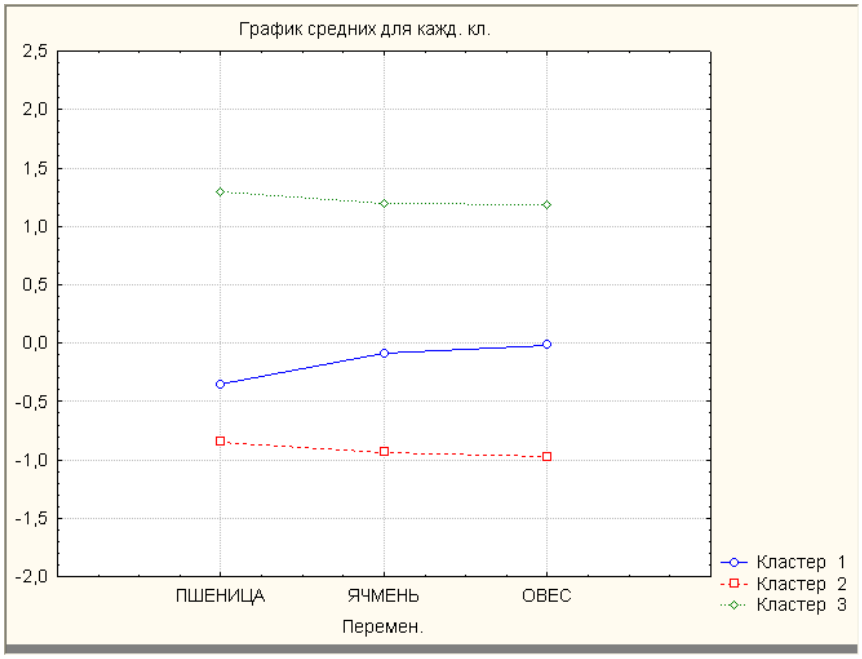


Рис. 6.13. Графическое представление центров полученных классов

6.2. ГИБКАЯ КЛАССИФИКАЦИЯ: ИСПОЛЬЗОВАНИЕ НЕЧЕТКИХ МНОЖЕСТВ

Реальный процесс развития математики в последние десятилетия показал, что импульсы для развития новых математических идей и направлений все чаще возникают под воздействием наук о человеке – биологии, гуманитарно-общественных дисциплин. Один из последних примеров такого рода – появление теории нечетких (размытых) множеств (*Fuzzy Set Theory*), предназначенной для анализа социальных ("гуманистических") систем.

Перспективность использования теории нечетких множеств (ТНМ) в исторических исследованиях связана прежде всего с созданием более адекватного, гибкого инструмента реализации историко-типологического метода. Этот метод, один из основных в процессе познания прошлого, оперирует понятийно-категориальным аппаратом, определяющим типическое в изучаемых процессах и явлениях. С другой стороны, трудности историко-типологического исследования связаны нередко с невозможностью "жесткого" определения понятия "тип", с принципиальной "размытостью" границ между типами. Это обстоятельство ограничивает использование понятий

классической теории множеств (на которой основывается практически вся математика) для формализации типов явлений или объектов в историко-типологическом исследовании. Какое место отводят ТНМ математики, связанные с разработкой этой теории? Вот слова основателя ТНМ, американского профессора Л.А. Заде:" Теория нечетких множеств – это, по сути дела, шаг на пути к сближению точности классической математики и всепроникающей неточности реального мира, к сближению, порожденному непрекращающимся человеческим стремлением к лучшему пониманию процессов познания... Нам нужна новая точка зрения, новый комплекс понятий и методов, в которых нечеткость принимается как универсальная реальность человеческого существования".

Сложность, неоднородность систем, изучаемых исторической наукой, проявляется и в том, что объекты принадлежащие к одному типу, в разной мере обладают присущими ему свойствами. Следовательно, при выделении типов (классов) объектов следует учитывать наличие **ядра** и его **периферии**. Ядро типа представляет такую группу объектов, для которых характерно концентрированное выражение всех специфических свойств типа, определяющих качественное отличие данного типа от всех иных.

Основные концепции и аппарат ТНМ достаточно подробно описаны в научной литературе. Введем лишь некоторые сведения о нечетких множествах. **Нечеткое множество** – это класс объектов, в котором нет резкой границы между теми объектами, которые входят в этот класс, и теми, которые в него не входят. Принадлежность каждого объекта к нечеткому множеству описывается с помощью величины, принимающей значения от 0 до 1. Эта величина называется **степенью принадлежности**; чем ближе она к 1, тем больше степень принадлежности объекта к данному нечеткому множеству. Если же эта величина равна 0, то объект не принадлежит данному множеству. Ядро нечеткого множества определяется как такой набор объектов, для каждого из которых степень принадлежности к данному нечеткому множеству превышает некоторое **пороговое значение** (например, 0.9) [1].

Отметим здесь отличие методов, основанных на ТНМ, от вероятностных методов классификации, определяющих вероятность отнесения объектов к "четким" классам; в этом случае неопределенность результатов классификации имеет вероятностную природу, она происходит из неполноты информации об объектах (эта неопределенность полностью снимается, если доступна вся информация об объектах). При использовании же ТНМ не-

[1] Более подробное, формализованное изложение основ ТНМ см, напр., в книге: Бородкин Л.И. Многомерный статистический анализ в исторических исследованиях. М., МГУ, 1986.

определенность связана с размытостью границ между классами. Понятие нечеткости относится к классам, в которых могут иметься различные степени принадлежности, промежуточные между полной принадлежностью (1) и непринадлежностью (0) объектов к классу. В этом случае неопределенность имеет другую природу, она не снимается и при наличии полной информации об объектах.

* * *

Наибольшую известность среди алгоритмов построения многомерной классификации, основанных на использовании концепций ТНМ, получил алгоритм FUZZY IZODATA. Цель данного алгоритма – вычисление оптимального набора весов принадлежности объектов к классам (число классов задается исследователем). Качество классификации оценивается с помощью специального критерия, минимизация которого приводит к выделению групп, состоящих из компактно расположенных точек многомерного пространства признаков. Чем ближе объект к центру соответствующего класса, тем ближе к единице значение веса его принадлежности к данному классу.

Программа FuzzyClass, разработанная в МГУ под руководством Л.И. Бородкина на основе модифицированного алгоритма FUZZY IZODATA, является оригинальным программным продуктом. Это реализованный в виде дружественного интерфейса алгоритм построения нечеткой многомерной классификации данных числовой природы и визуализации результатов. В режиме диалога с пользователем можно учесть априорную информацию о структуре "обучающей выборки" (если таковая имеется), выбрать значение "параметра размытости", провести (при необходимости) содержательную корректировку полученной типологии и выйти на построение условно-оптимальной (с учетом проведенной корректировки) структуры классов.

Пример 6.3. Вернемся к файлу данных Harvest. Для работы с программой FuzzyClass нам понадобится копия этого файла в текстовом (ASCII) формате, содержащая данные о тех же 15 странах и трех признаках, на которых проводился анализ в предыдущих разделах. Этот файл называется Harvest.txt.

Работа с программой размытой классификации идет в режиме диалога. Сначала необходимо задать имя директории и имя самого файла данных, а также имя файла, в который будут записаны результаты классификации. Затем задаются собственно параметры классификации: число объектов (15), число признаков (3), число классов (снова выберем 3 класса), а также т.н. параметр "размытости". Этот параметр определяет степень "размытости" результатов: от минимальной (параметр равен 1), когда степени принадлежности могут принимать значения либо 0, либо 1 (в этом случае резуль-

тат аналогичен результатам обычной кластеризации) – до максимальной (параметр равен 2). В нашем случае выберем значение 1,8.

Таблица 6.1. Веса принадлежности объектов классам

Объекты	Веса		
	1	2	3
Россия	**0.96**	0.04	0.00
Австрия	0.01	**0.99**	0.00
Венгрия	0.01	**0.99**	0.00
Великобритания	0.03	0.10	**0.88**
Болгария	**0.90**	0.10	0.01
Германия	0.00	0.00	**1.00**
Голландия	0.01	0.03	**0.95**
Испания	**0.95**	0.04	0.01
Румыния	0.73	0.25	0.02
Сербия	**0.98**	0.02	0.00
Франция	0.03	**0.97**	0.00
Швейцария	0.01	0.02	**0.97**
Швеция	0.00	0.01	**0.98**
Канада	0.02	**0.96**	0.01
США	0.77	0.23	0.01

Рассмотрим основные результаты работы программы. Они состоят из таблицы степеней принадлежности объектов классам (табл. 6.1) и таблицы центров классов (табл. 6.2).

Видно, что результаты в целом соответствуют полученным ранее. Выделим ядро каждого класса по порогу принадлежности, равному 0,8. В первый класс вошли Россия, Болгария, Испания, Румыния, Сербия и США. Однако в отличие от результатов "жесткой" классификации видно, что только четыре из этих шести стран входят в ядро класса, тогда как две страны (Румыния и США) относятся к его периферии, обнаруживая некоторое сходство с объектами второго класса. В определенной степени эти страны являются объектами, переходными от первого ко второму классу.

Во второй класс опять-таки вошли, как и раньше, Австрия, Венгрия, Франция и Канада, причем все они относятся к ядру этого класса. То есть второй класс является более "сплоченным", чем первый (то же самое можно сказать и о третьем классе). В целом 11 стран из 15 входят в свои классы с весом принадлежности от 0,9 до 1, две страны входят с весом от 0,8 до 0,9 и две – с весом от 0,7 до 0,8 (это уже упомянутые Румыния и США). Таким образом, типологическая картина получилась довольно четкой.

Наконец, таблица 6.2 показывает центры классов, т.е. средние урожайности пшеницы, овса и ячменя по каждому классу, подсчитанные с учетом весов принадлежности объектов классам.

Таблица 6.2. Центры классов

Класс	Урожайность		
	пшеницы	ячменя	овса
1	68,44	70,30	57,22
2	89,85	98,88	92,22
3	156,21	142,76	137,42

ВОПРОСЫ

1. Какова содержательная гипотеза, лежащая в основе кластерного анализа?
2. Рассмотрим два основных метода кластерного анализа, используемых в статистических пакетах – иерархический и метод К-средних. В чем их главное различие?
3. В каком из этих двух методов необходимо задавать число кластеров?
4. Что такое стандартизация признаков? Почему она, как правило, должна предшествовать проведению кластерного анализа?
5. Укажите два вида исходных данных для реализации кластерного анализа.
6. Что такое матрица расстояний?
7. Пусть имеющийся набор данных включает как количественные, так и качественные признаки. Можно ли использовать стандартные статистические пакеты для проведения кластерного анализа?
8. Изменятся ли результаты кластерного анализа, если удалить часть признаков из имеющегося набора?
9. Какие показатели (характеристики классов) используются в статистических пакетах для интерпретации результатов кластерного анализа?
10. Назовите основные правила соединения объектов (или кластеров) при использовании иерархического метода кластерного анализа.
11. Назовите основные меры расстояния между объектами (или кластерами) при использовании иерархического метода кластерного анализа.
12. Какую информацию дает пользователю окошко **Amalgamation rule** в меню иерархического метода кластерного анализа в пакете STATISTICA?
13. В каком случае структуру построенного дерева классификации следует считать более удачной (см. рис. а и рис. б)?

14. Какие объекты более похожи в структуре классификации: 1-й и 3-й или 4-й и 5-й (см. рис. б)?

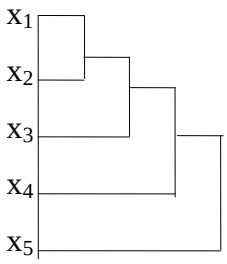


Рис. а

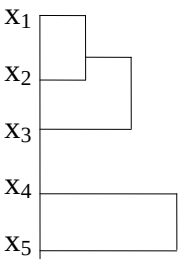


Рис. б

ЗАДАНИЯ

1. Провести кластер-анализ тех же объектов, что в приведенных выше примерах, но без предварительной стандартизации исходных данных. Попробуйте объяснить различия в результатах.
2. Классифицировать на два класса те же объекты, что в приведенном выше примере, с помощью метода размытой классификации. Дать интерпретацию ядер и периферии полученных классов. Сравнить с результатами агломеративно-иерархического метода кластер-анализа.
3. Классифицировать на два класса те же объекты, но с помощью метода *k*-средних. Сравнить с результатами агломеративно-иерархического метода.
4. Классифицировать на два и три класса с помощью метода *k*-средних данные о результатах голосования за основные политические партии на выборах 1917 г. в Учредительное собрание по избирательным округам (файл Uchred.sta). Сравнить полученные результаты.
5. Классифицировать эти же данные, пользуясь агломеративно-иерархическим методом программы STATISTICA. Выяснить, какой из методов дает лучшие результаты.
6. Используя файл Tambov.sta, классифицировать уезды Тамбовской губернии:
   а) по данным социально-экономического характера;
   б) по данным о числе голосов, поданных за различные политические партии на выборах в Учредительное собрание.

   Выяснить степень сходства полученных результатов по дендрограммам.
   Указание. Обратите внимание на необходимость стандартизации данных.

7. Постройте классификацию стран мира по данным демографической статистики (файл Demo_wor.sta). Предварительно стандартизуйте исходные данные.
8. Постройте классификацию республик СССР по данным о среднем числе членов семьи (файл Family.sta).

Указание. С помощью агломеративно-иерархического метода постройте дендрограмму, по ее структуре оцените число классов, а затем с помощью метода *k*-средних найдите характеристики этих классов. Охарактеризуйте динамику демографических процессов в каждом классе.

ГЛАВА 7

ФАКТОРНЫЙ АНАЛИЗ

7.1. ОБЩЕЕ ОПИСАНИЕ

Изучение зависимостей между признаками, описывающими сложное явление или процесс, неизменно ставит вопрос: каковы причины, обусловившие именно данную структуру связей? Без специальных методов анализа сложные зависимости такой системы признаков весьма трудно "распутать", тем более что взаимосвязи признаков (измеряемые, например, коэффициентами корреляции) могут интерпретироваться не только как зависимости одних признаков от других, но и как зависимости от неких скрытых параметров, определяющих изменение целых групп коррелированных признаков. Общая причина изменений таких признаков ведет к тому, что в силу своей согласованности они в некотором смысле дублируются. Стремление объяснить целую совокупность признаков через введение более глубинных, обобщенных характеристик явления, в основном определяющих его структуру, приводит к модели факторного анализа.

В факторном анализе предполагается, что все многообразие и структура взаимосвязей между параметрами, описывающими явление "извне" и поддающимися непосредственному измерению (признаками), обусловлены некими скрытыми, но объективно существующими причинами, так называемыми факторами, измерить которые непосредственно нельзя. Концепция факторного анализа, таким образом, обеспечивает "сжатие" информации, объясняя множество признаков через небольшое, как правило, число факторов. При этом предполагается, что эти факторы обеспечивают не просто концентрацию информации, но и являются наиболее существенными, определяющими характеристиками изучаемого явления. Кроме того, трудности, связанные с описанием сложной внутренней зависимости и интерпретацией результатов анализа, заставляют, с одной стороны, уменьшать число рассматриваемых величин, а с другой – сводить их к независимым. Факторный анализ дает основание считать его методом, позволяющим генерировать содержательные гипотезы о структуре системы признаков, о наиболее значимых, существенных признаках, о группах тесно связанных признаков. Таким образом, факторный анализ удобен не только тем, что позволяет обрабатывать одновременно большое число (десятки) признаков, но и тем, что сам может являться "источником возникновения гипотез", особенно в тех областях, где сложная структура явлений еще недостаточно изучена.

В основе факторного анализа лежит идея о том, что за сложными взаимосвязями измеренных, учтенных нами признаков стоит относительно более простая структура, отражающая наиболее существенные черты изучаемой системы, а измеренные признаки являются конкретными проявлениями скрытых общих факторов, определяющих эту структуру.

Факторный анализ позволяет выявлять общие факторы, дает ключ к их содержательному толкованию, оценивает их воздействие на отдельные показатели и на все изучаемое явление в целом, количественно выражает их значения для каждого из рассматриваемых объектов и на основании всего этого дает возможность решать целый ряд задач, возникающих при обработке массовых источников.

В общем случае поведение каждого исходного признака определяется действием на него совокупности относительно небольшого числа *общих факторов*, влияющих на все показатели и обусловливающих взаимосвязи между ними, и *характерного фактора*, воздействующего только на данный показатель. При этом различные признаки имеют собственные единицы измерения, а с другой стороны, будучи выраженными через одни и те же факторы, они должны измеряться в каких-то общих единицах. Кроме того, каждый признак, имея собственные единицы измерения, несопоставим с другими, а факторы, общие для целого ряда признаков разной размерности, вообще были бы лишены смысла. Исходя из этих соображений, в факторном анализе все величины, входящие в факторную модель, стандартизированы, т.е. являются безразмерными величинами со средним арифметическим значением 0 и средним квадратическим отклонением 1 [1].

7.1.1. Факторные нагрузки

Коэффициент взаимосвязи между некоторым признаком и общим фактором, выражающий меру влияния фактора на признак, называется **факторной нагрузкой** данного признака по данному общему фактору.

Матрица, состоящая из факторных нагрузок и имеющая число столбцов, равное числу общих факторов, и число строк, равное числу исходных признаков, называется факторным отображением или просто **факторной матрицей**. Основой для вычисления этой матрицы является матрица парных коэффициентов корреляции исходных признаков. Как известно, корреляционная матрица фиксирует степень взаимосвязи между каждой парой признаков. Аналогично факторная матрица фиксирует степень линейной

[1] В дальнейшем будем считать, что все исходные признаки стандартизированы; то же относится к общим факторам и характерным факторам. Кроме того, предполагается, что все общие и характерные факторы *попарно ортогональны*, т.е. попарно независимы.

связи каждого признака с каждым общим фактором. Таким образом, величина факторной нагрузки не превышает по модулю единицы, а знак ее говорит о положительной или отрицательной связи признака с фактором. Чем больше абсолютная величина факторной нагрузки признака по некоторому фактору, тем в большей степени этот фактор определяет данный признак. Значение факторной нагрузки по некоторому фактору, близкое к нулю, говорит о том, что этот фактор практически на данный признак не влияет.

Факторная модель дает возможность вычислять **вклады факторов** в общую дисперсию всех признаков. Суммируя квадраты факторных нагрузок для каждого фактора по всем признакам, получаем его вклад в общую дисперсию системы признаков: чем выше доля этого вклада в общей дисперсии, равной, очевидно, числу признаков, тем более значимым, существенным является данный фактор. При этом можно выявить и оптимальное в некотором смысле количество общих факторов, достаточно хорошо описывающих систему исходных признаков.

7.1.2. Факторные веса

Значение (мера проявления) фактора у отдельного объекта называется **факторным весом** объекта по данному фактору. Факторные веса позволяют ранжировать, упорядочить объекты по каждому фактору. На практике ограничиваются вычислением факторных весов по нескольким наиболее значимым факторам. Чем больше факторный вес некоторого объекта, тем больше в нем проявляется та сторона явления или та закономерность, которая отражается данным фактором. Если, например, речь идет о факторе, выражающем уровень развития в определенном аспекте, то большой факторный вес свидетельствует о высоком уровне развития данного объекта, а низкий факторный вес – о низком уровне. Факторные веса могут быть как положительными, так и отрицательными. В силу того, что факторы являются стандартизованными величинами со средним значением, равным нулю, факторные веса, близкие к нулю, говорят о средней степени проявления фактора, положительные – о том, что эта степень выше средней, отрицательные – о том. что она ниже средней.

Таблица факторных весов имеет n строк по числу объектов и k столбцов по числу общих факторов. Положение объектов на оси каждого фактора показывает, с одной стороны, тот порядок, в котором они ранжированы по этому фактору, а с другой стороны, равномерность или же неравномерность в их расположении, наличие скоплений точек, изображающих объекты, что дает возможность визуально выделять более или менее однородные группы.

Обычно факторы выделяются последовательно, в соответствии с их вкладами в суммарную дисперсию признаков: сначала находится фактор, имеющий максимальный вклад, затем его влияние устраняется, и для матрицы остаточных корреляций снова ищется фактор с максимальным вкладом и т.д. Процесс последовательного нахождения факторов прекращается, если их суммарный вклад превысит определенный, заранее выбранный порог. Существуют и другие критерии прекращения процесса поиска факторов.

7.2. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Чрезвычайно удобным в качестве метода "сжатия" информации с целью выявления обобщенных характеристик явления является *метод главных компонент* – разновидность факторного анализа [1]. Этот метод основан на допущении, что характерности всех признаков равны нулю, а число общих факторов k равно числу исходных признаков m. В этом случае переход к факторам, которые являются просто линейными комбинациями исходных признаков, означает не что иное, как переход к новой системе координат.

Главные компоненты независимы, т.е. в геометрическом плане ортогональны. Выделение первой главной компоненты по максимальному вкладу в дисперсию признаков означает, что мы находим такое направление в пространстве признаков, которому соответствует максимальная дисперсия, т.е. наибольшая дифференциация, разброс объектов. Затем находится вторая главная компонента, ортогональная первой и дающая вновь наибольшую дифференциацию объектов, не объясненную первой компонентой, и т.д. После построения всех главных компонент (число которых равно числу признаков m) остаточная дисперсия оказывается равной нулю, т.е. задача имеет точное математическое решение. Обычно суммарная дисперсия признаков раскладывается по главным компонентам таким образом, что первые несколько компонент уже объясняют почти всю эту дисперсию, а остальные почти ничего не добавляют, поэтому совсем не обязательно выделять все m компонент. На практике ограничиваются несколькими первыми, т.к. их оказывается достаточно для хорошего описания в сжатом виде всей исходной информации [2].

Критерием отбора необходимого числа главных компонент обычно служит процент объясненной дисперсии, т.е. отношение суммарного вклада уже найденных компонент к общей дисперсии исходных признаков, равной

---

[1] Некоторые авторы рассматривают метод главных компонент в качестве отдельного направления многомерного анализа.

[2] Дать более контрастную матрицу факторных нагрузок может т.н. *вращение* полученных компонент.

m. Практически, если число уже найденных главных компонент (или факторов) не больше, чем *m*/2, объясняемая ими дисперсия не менее 70%, а следующая компонента дает вклад в суммарную дисперсию не более 5%, факторная модель считается достаточно хорошей.

Пример 7.1. Рассмотрим, как работают методы факторного анализа, используя сводные данные 1987 г. об экономическом развитии нескольких европейских стран, а также США и Японии (файл Tab_1987.sta). Исходные данные приведены на рис. 7.1. Показатели 1 (национальный доход), 2 (производство электроэнергии), 3 (военные раходы), 4 (производство зерна), 5 (количество телефонов), 7 (потребление мяса) даны в расчете на душу населения, показатели 6 и 8 (количество автомобилей и телевизоров) – на 1000 чел.

Данные: TAB_1987.STA* (8v * 7c)

	Основные показатели общего уровня развития стран в 1987 г.							
	1 ДОХОД	2 ЭНЕРГИЯ	3 ВОЕН_РАС	4 ЗЕРНО	5 ТЕЛЕФОНЫ	6 АВТОМОБИ	7 ПОТР_МЯС	8 ТЕЛЕВИЗО
СССР	5975	5911	914	829	12	70	45	311
АВСТРИЯ	13421	6711	118	702	50	409	97	440
ВЕЛИКОБР	12583	2778	436	412	52	348	75	433
США	15012	11479	968	1305	92	718	112	798
ФРГ	14384	6909	360	427	66	463	97	383
ФРАНЦИЯ	13327	6823	394	984	61	438	100	389
ЯПОНИЯ	12770	5705	102	130	55	374	36	579

Рис. 7.1. Исходные данные для факторного анализа (файл Tab_1987)

Для выполнения факторного анализа в пакете STATISTICA надо обратиться к разделу **Факторный анализ** модуля **Многомерный разведочный анализ**. На рис. 7.2 показано диалоговое окно этого модуля.

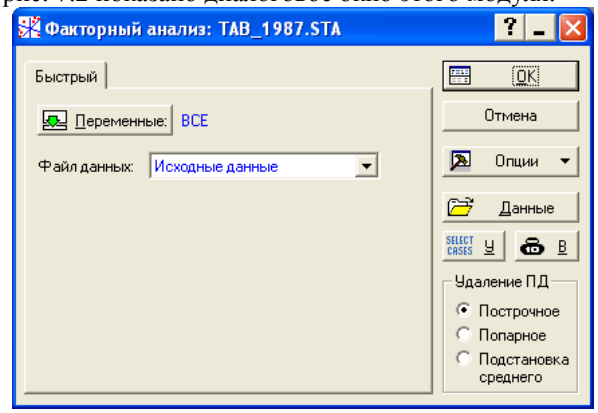


Рис. 7.2. Диалоговое окно факторного анализа

В поле **Файл данных** можно выбрать один из двух возможных способов представления данных: программа может работать либо с *Исходными данными* (исходной таблицей объекты–признаки), либо с уже готовой *Корреляционной матрицей*. В данном случае мы имеем дело с исходными данными. Как всегда, в первом диалоговом окне надо выбрать переменные (графическая кнопка **Переменные**), с которыми будет работать программа – в данном случае выбраны все переменные.

После щелчка по графической кнопке **OK** программа вычисляет корреляционную матрицу (если необходимо) и переходит к следующему диалоговому окну (рис. 7.3) – окну выбора метода факторного анализа (**Задайте метод выделения факторов**). В блоке **Метод выделения** по умолчанию предлагается метод **Главных компонент**, на котором мы и остановим свой выбор. Далее, в поле **Максимальное число факторов** надо задать максимальное число факторов (равное по умолчанию двум). Можно также изменить заданное по умолчанию минимальное **Собственное значение**, определяющее критерий окончания процесса извлечения факторов из корреляционной матрицы (но лучше оставить значение по умолчанию).

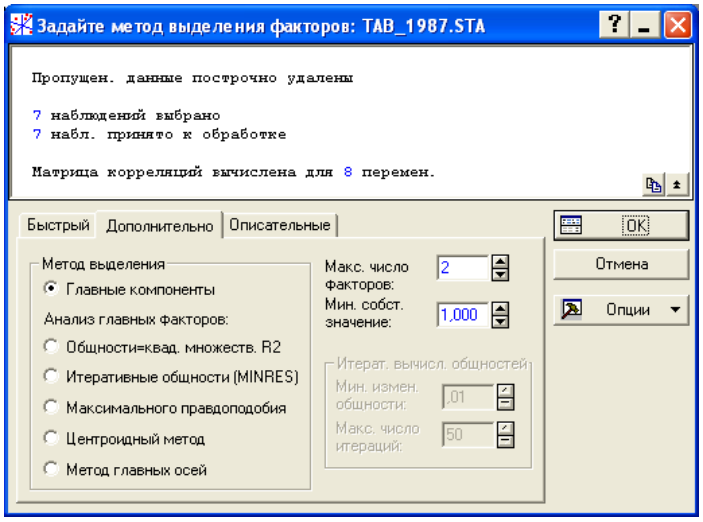


Рис. 7.3. Выбор метода факторизации (по умолчанию – метод главных компонент) и числа факторов

После выполнения вычислительных процедур программа предлагает пользователю просмотреть результаты (см. рис. 7.4). Наиболее важными из них являются: матрица **Факторных нагрузок** признаков и входящие в нее

вклады факторов в суммарную дисперсию признаков, а также матрица факторных весов объектов – **Значения факторов** [1].

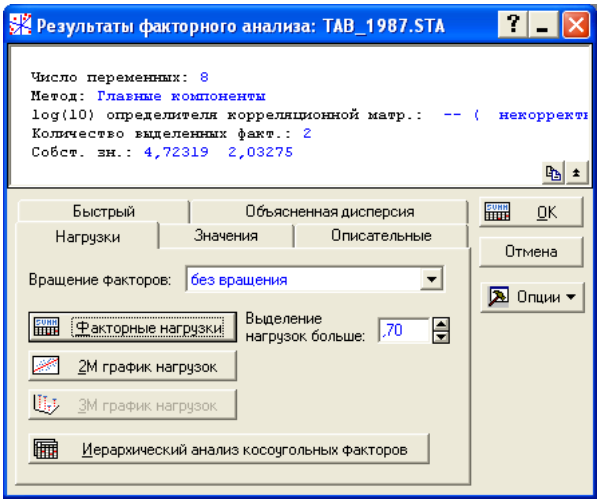


Рис. 7.4. Окно результатов факторного анализа

Начнем анализ результатов с матрицы факторных нагрузок (рис. 7.5) [2]. В столбцах этой матрицы стоят факторные нагрузки признаков. Факторы всегда упорядочены по размеру их вкладов в суммарную дисперсию признаков: самым значимым в этом смысле является первый фактор. В двух последних строках таблицы находятся значения вкладов в абсолютном и относительном выражении.

Так, число 4,55 в первом столбце строки *Общ. Дис.* означает, что первый фактор объясняет 4,55 из суммарной дисперсии, равной 8 (8 – число признаков в исходной таблице), что составляет почти 57% (строка *Доля общ.* – доля вклада). Таким образом, уже один первый фактор более, чем наполовину, описывает исходные данные. Второй фактор объясняет еще почти 25% суммарной дисперсии, а вместе с первым – более 80%.

Как можно интерпретировать полученные нами факторы? Обратите внимание на наиболее значимые факторные нагрузки (например, превыша-

[1] О факторных весах см. ниже.

[2] Матрицу нагрузок можно получить как с вращением, так и без вращения факторов. Вращением называется метод "контрастирования" факторных нагрузок, который иногда облегчает интерпретацию результатов. Наиболее часто используется вращение по методу *Varimax*. В данном случае вращение факторов не применялось.

ющие по модулю 0,75 или 0,8), именно они указывают на признаки, наиболее тесно связанные с данным фактором.

Перемен.	Фактор.нагрузки (без вращ.) (TAB_1987. Выделение: Главные компоненты (Отмечены нагрузки >,700000) Фактор 1	Фактор 2
ДОХОД	0,776713	-0,593263
ЭНЕРГИЯ	0,793882	0,402968
ВОЕН_РАС	0,206109	0,898122
ЗЕРНО	0,564466	0,730197
ТЕЛЕФОНЫ	0,920669	-0,286468
АВТОМОБИ	0,954927	-0,217189
ПОТР_МЯС	0,777162	0,010091
ТЕЛЕВИЗО	0,766660	0,009003
Общ.дис.	4,545899	1,983573
Доля общ	0,568237	0,247947

Рис. 7.5. Матрица факторных нагрузок

В нашем случае первый фактор наиболее тесно связан с такими признаками, которые позволяют в качестве рабочего названия для данного фактора принять название "уровень жизни" или "уровень потребления". Со вторым фактором наиболее тесно связан признак "военные расходы на душу", что позволяет предложить для него такое рабочее название, как "уровень милитаризации экономики". Для более детального анализа полученных факторов полезно рассмотреть, как выглядят факторные веса объектов (рис. 7.6).

Набл.	Значения факторов (TAB_1987 Вращение: без вращ. Выделение: Главные компонен Фактор 1	**Фактор 2**
СССР	-1,42619	1,75054
АВСТРИЯ	0,03923	-0,49624
ВЕЛИКОБР	-0,50950	-0,55845
США	1,90541	0,96990
ФРГ	0,18964	-0,64578
ФРАНЦИЯ	0,25608	0,05938
ЯПОНИЯ	-0,45468	-1,07935

Рис. 7.6. Факторные веса объектов

Сначала посмотрим на первый столбец матрицы факторный весов. Как известно, положительные значения весов свидетельствуют о проявлении фактора выше среднего. Это относится в первую очередь к США (факторный вес равен 1,90). Отрицательные значения свидетельствуют о проявлении фактора ниже среднего уровня. По первому фактору самое низкое значение имеет СССР (–1,42). Остальные страны находятся приблизительно на

среднем уровне, т.к. их факторные веса незначительно отклоняются от нуля. Таким образом рабочее название первого фактора ("уровень потребления") получает косвенное подтверждение.

Расположение объектов по второму фактору совсем иное. Два самых высоких значения имеют СССР и США, а самое низкое – Япония.

7.3. ФАКТОРНЫЙ АНАЛИЗ КАК СПОСОБ КЛАССИФИКАЦИИ

Факторный анализ можно использовать не только как метод сжатия информации, но и как удобный подход к классификации, когда множество исходных признаков заменяется найденными обощенными факторами.

Идея проведения классификации объектов в пространстве немногих факторов, заменяющих большое количество исходных признаков, состоит в измерении факторов на отдельных объектах и последующей визуальной или аналитической группировке объектов на оси одного фактора или в пространстве двух факторов. При типологизации по большому числу исходных признаков полученные группы бывает трудно охарактеризовать, тогда как факторы, концентрирующие информацию, служат хорошим средством построения типологии и дают удобный способ графического представления данных. Возможно и применение автоматической классификации к объектам, заданным в пространстве факторов.

Вернемся к рассмотренному выше примеру и попытаемся провести классификацию семи стран по всем показателям, включенным в исходную таблицу. Разумеется, для решения этой задачи можно обратиться к разделу **Многомерный разведочный анализ | Кластерный анализ** программы STATISTICA (см. гл. 6). В данной главе мы, однако, поставим вопрос несколько иначе: а существуют ли в пространстве признаков реальные группы (скопления) объектов? Очевидно, визуализация объектов в восьмимерном пространстве признаков нам недоступна. Попытаемся визуализировать данные, пользуясь привычным двумерным пространством, а именно, пространством двух факторов, в котором у каждого объекта всего две координаты – его факторные веса.

Это можно сделать следующим образом: сохраним таблицу факторных весов как новый файл программы STATISTICA. Для этого надо воспользоваться кнопкой **Сохранить значения** на вкладке **Значения** окна результатов факторного анализа. Появится окно с таблицей факторных весов (при желании вместе с факторными весами можно сохранить и исходные переменные, некоторые или все). Затем в разделе **Файл** основного меню программы надо выбрать команду **Сохранить как** и дать имя этому новому файлу, например, Scores.sta. Теперь можно закрыть исходный файл Tab_1987 и открыть файл Scores.sta. В этом новом файле семь объектов и

два признака (факторные веса объектов), которые называются Фактор1 и Фактор2.

Теперь воспользуемся разделом **Графика** основного меню программы и выберем вид графика: **Диаграммы рассеяния**. В соответствующем диалоговом окне (см. рис. 7.7) надо выбрать (кнопка **Переменные**) признаки, значения которых будут откладываться по горизонтальной и вертикальной осям.

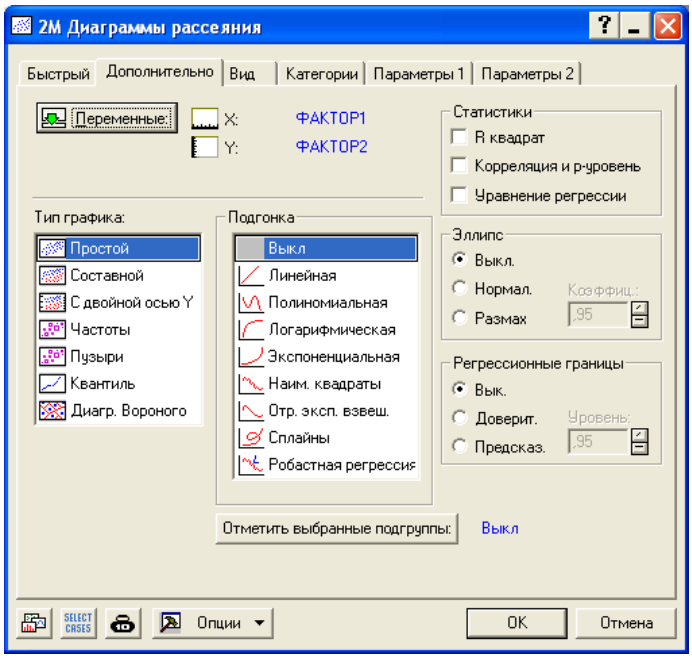


Рис. 7.7. Диалоговое окно построения диаграммы рассеяния

Обратите внимание, что в диалоговом окне на рис. 7.7 есть вкладка **Параметры 1**. Она позволяет поместить на диаграмме не просто точки, но и "подписать" имена соответствующих объектов (конечно, это имеет смысл, если объектов не слишком много). Нажатие на эту кнопку открывает диалоговое окно, представленное на рис. 7.8.

Чтобы видеть имена объектов на диаграмме, в поле **Метки наблюдений** блока **Параметры отображения** этого окна следует выбрать значение *Имена наблюдений* и щелкнуть по графической кнопке **OK**. В результате будут получена диаграмма рассеяния, которую вы видите на рис. 7.9.

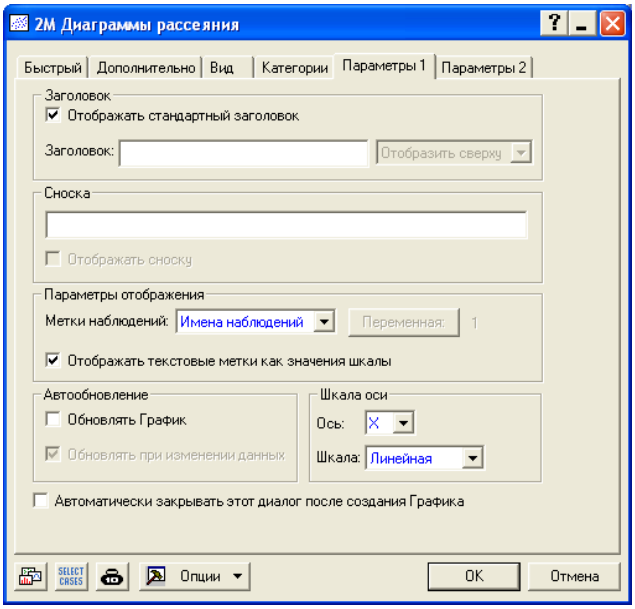


Рис. 7.8. Диалоговое окно задания параметров диаграммы

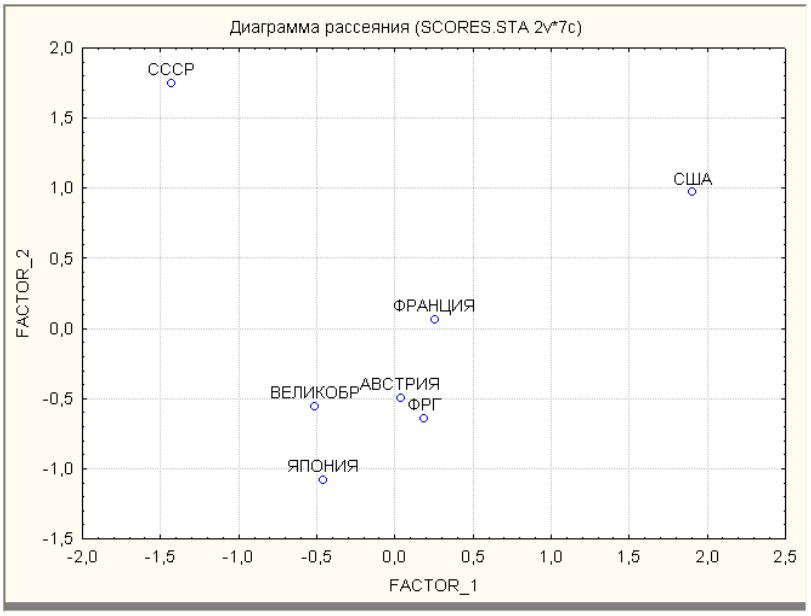


Рис. 7.9. Диаграмма рассеяния объектов в пространстве первых двух факторов

На этой диаграмме видно, что все страны, кроме СССР и США располагаются довольно тесной группой, т.е. достаточно близки друг к другу по выбранным показателям. Две точки на этой диаграмме отстоят довольно далеко от этой основной группы: СССР в верхнем левом углу (большой отрицательный вес по первому фактору и большой положительный вес по второму) и США в верхнем правом углу (большие положительные веса по обоим факторам).

Если теперь обратиться к кластерному анализу, то в результате применения метода *k*-средних будут получены результаты, хорошо согласующиеся с диаграммой рис. 7.9: при задании двух классов в один из них попадут все страны, кроме СССР, который образует отдельный класс, а при задании трех классов и СССР, и США будут занимать по отдельному классу, тогда как остальные страны по-прежнему объединяются в один общий класс.

Задание: попробуйте выполнить этот анализ самостоятельно.

ВОПРОСЫ

1. Каковы основные цели использования факторного анализа?
2. Дайте интерпретацию понятия "фактор".
3. В чем особенность метода главных компонент?
4. Что такое матрица корреляции? Какое отношение она имеет к факторному анализу?
5. Что такое факторные нагрузки?
6. Что такое факторные веса?
7. Что является показателем качества построенной факторной модели?
8. Пусть три первых фактора объясняют 40% суммарной дисперсии признаков. Надо ли увеличивать число факторов?
9. Существуют ли пределы для значений факторных нагрузок? Если да, то каковы они?
10. Существуют ли пределы для значений факторных весов? Если да, то каковы они?
11. Пусть для рассматриваемого фактора значения факторных нагрузок первых четырех признаков равны 0.78, 0.23, –0.09 и –0.87. Что это значит? Как учитываются эти значения для интерпретации данного фактора?
12. Пусть для рассматриваемого фактора значения факторных весов первых четырех объектов равны 1.78, 0.11, –0.02 и –2.87. Что это значит? Дайте интерпретацию в том случае, когда данный фактор характеризует степень промышленного развития, а объекты – губернии.
13. Что произойдет со значениями факторных весов объектов, если знаки всех факторных нагрузок изменить на противоположные?

14. Что является более предпочтительным для группировки признаков – кластерный или факторный анализ?

ЗАДАНИЯ

1. По данным об аграрном развитии губерний Европейской России на рубеже XIX–XX вв. (файл Typol.sta) провести факторный анализ, включив в него признаки с номерами 1, 3, 5, 13, 14, 16 (основные признаки для социальной аграрной типологии) [1]. Из 50 губерний Европейской России отобрать для этого анализа 24 губернии, 12 из которых относятся к помещичьему типу буржуазной аграрной эволюции (губернии с номерами 4, 5, 8, 11, 16, 17, 19, 21, 22, 23, 33, 49) и 12 – к крестьянскому типу (губернии с номерами 1, 2, 7, 10, 12, 13, 14, 27, 28, 37, 41, 47) [2]. Охарактеризовать полученные факторы (взять число факторов, равное 2) на основе таблицы факторных нагрузок.
2. По результатам предыдущего задания создать файл, в котором в качестве "координат" объектов, т.е. губерний взяты их факторные веса. Проанализировать графически распределение губерний в пространстве двух факторов (подумайте, какой из модулей пакета STATISTICA надо использовать). Сделайте вывод об эффективности использования факторного анализа для классификации объектов.
3. По данным файла Typol.sta провести факторный анализ по двум факторам и сравнить результаты для губерний: а) крестьянского типа аграрной эволюции (Южный и Юго-восточный Степной и Северо-восточный – объекты № 1, 2, 7, 10, 12, 13, 14, 27, 28, 37, 41, 47) и б) помещичьего типа (Прибалтийский и Западный – объекты № 4, 5, 8, 11, 16, 17, 19, 21, 22, 23, 33, 49). Отобрать из 19 исходных признаков шесть показателей социальной аграрной типологии с номерами 1, 3, 5, 13, 14, 16). В чем основное отличие результатов факторного анализа для двух указанных групп губерний?
4. Выявить основные факторы социально-экономического развития уездов Тамбовской губернии (файл Tambov.sta) в 1917 г. Сравнить корреляционную матрицу и графическое представление факторных весов признаков в пространстве первых двух факторов. Что можно сказать о показателях, характеризующих аграрное и индустриальное развитие губернии?

---

[1] Ковальченко И.Д., Бородкин Л.И. Аграрная типология губерний Европейской России на рубеже XIX – XX веков. (Опыт многомерного количественного анализа) // История СССР, 1979, №1. С. 92-93.

[2] Бородкин Л.И. Многомерный статистический анализ в исторических исследованиях. М., 1986. С. 93.

ЧАСТЬ IV

АНАЛИЗ ДИНАМИКИ

ГЛАВА 8

ВВЕДЕНИЕ В АНАЛИЗ ВРЕМЕННЫХ РЯДОВ [1]

Изучение процессов и явлений во времени занимает важное место в исторических исследованиях и в плане математического обеспечения этих исследований имеет свои особенности. Анализ так называемых временных рядов образует весьма непростую, обширную и разветвленную область статистики. Мы ограничимся изложением тех прикладных аспектов анализа временных рядов, которые полезны и важны при изучении динамики исторических процессов.

Временным рядом называется последовательность числовых значений, характеризующих изменение некоторого признака во времени. Отдельные значения признака, относящиеся к определенным промежуткам или моментам времени, принято называть **уровнями** или **элементами** ряда. Мы будем обозначать их x_t, где t – время.

Отдельные стороны исторических процессов отражаются с помощью временных рядов некоторых показателей. Временные ряды позволяют исследовать последовательности состояний и переходы из одних состояний к другим, определять параметры и тенденции развития явлений во времени.

8.1. ПЕРВИЧНЫЙ АНАЛИЗ ДИНАМИКИ

Рассмотрим пример из области экономической истории.

Пример 8.1. Закономерным процессом для всех стран, вставших на путь индустриального развития, является переход на более эффективные источники энергии. Для XIX – начала XX вв. это означает минерализацию топливной базы, в частности, переход на каменноугольное топливо. В этом контексте исследовательский интерес представляет развитие каменно-

[1] Для практического освоения начальной части этого раздела потребуются навыки работы в пакете Misrosoft Excel.

угольных бассейнов России конца XIX – начала XX вв. С какой скоростью, насколько интенсивно развивались различные каменноугольные бассейны России? В статистике разработана система несложных показателей для решения подобных задач.

8.1.1. Характеристики скорости и интенсивности изменения временного ряда

Основными показателями динамики являются **абсолютный прирост**, **темп роста** и **темп прироста**. Первый характеризует скорость изменения уровней ряда, второй – интенсивность изменений, третий – относительную скорость изменений процесса.

Расчет этих показателей проводится при постоянной и переменной базе, причем, за постоянную базу принимается, как правило, первый уровень временного ряда (x_1), а переменной базой всегда служит предшествующий уровень (x_{t-1}). Показатели первой группы называются ***базисными***, а второй группы – ***цепными***. Формулы основных показателей динамики представлены в табл. 8.1.

Задание 8.1. Рассчитайте абсолютные приросты, темпы роста и темпы прироста добычи каменного угля в Западной Сибири за период 1887–1913 гг. (исходные данные приведены в файле Coal.xls). Расчеты легко сделать в пакете Excel. Воспользуйтесь формулами соответствующих показателей и сравните полученные результаты с рис. 8.1. Интерпретируйте результаты. Для большей наглядности можно воспользоваться графикой пакета.

*Таблица 8.1. Формулы цепных и базисных индексов**

Абсолютные приросты		Темпы роста		Темпы прироста	
цепные	базисные	цепные	базисные	цепные	базисные
$x_t - x_{t-1}$	$x_t - x_1$	x_t/x_{t-1}	x_t/x_1	$(x_t - x_{t-1})/x_{t-1}$	$(x_t - x_1)/x_1$

*) Здесь t изменяется от 2 до n, где n – число членов ряда.

8.1.2. Средние характеристики временного ряда

Для характеристики изменения временного ряда, скорости и интенсивности этого изменения в среднем за рассматриваемый период служат такие показатели, как средний абсолютный прирост и средний темп роста.

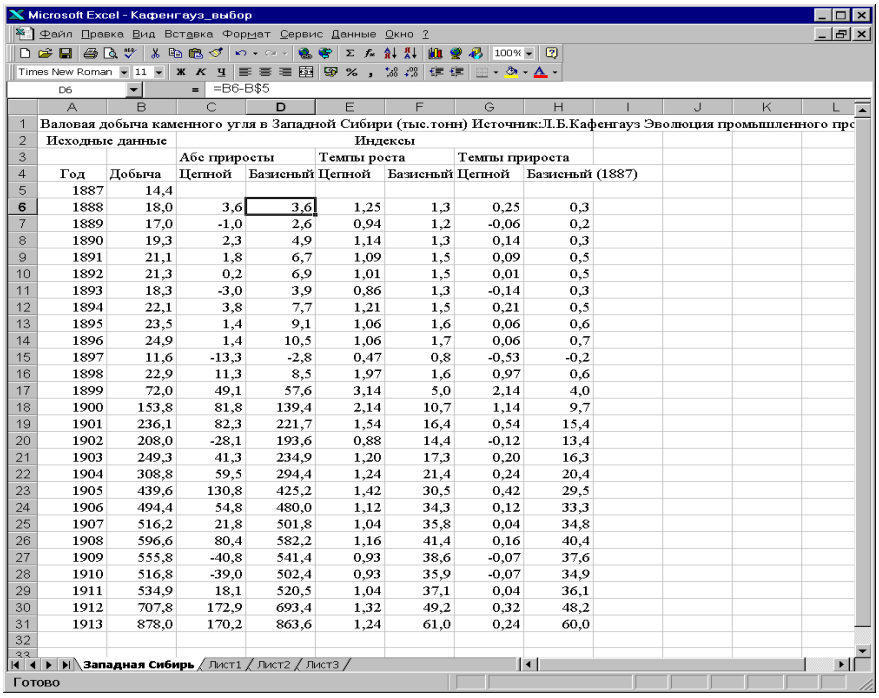
Средний абсолютный прирост показывает, насколько быстро в среднем за период изменяется конечный уровень ряда относительно первоначального, и вычисляется по формуле:

$$\Delta x = (x_n - x_1)/(n-1),$$

где x_n и x_1 – последний и первый уровни ряда соответственно.

Задание 8.2. Рассчитайте средний абсолютный прирост добычи каменного угля в Западной Сибири за период 1887–1913 гг., используя исходные данные, а также расчеты задания 8.1.

Результат: в среднем за рассматриваемый период ежегодный прирост добычи каменного угля в западносибирском угольном бассейне составлял 33,2 тысячи тонн.



Microsoft Excel - Кафенгауз_выбор

Файл Правка Вид Вставка Формат Сервис Данные Окно ?

Times New Roman 11 100%

D6 = =B6-B$5

	A	B	C	D	E	F	G	H	I	J	K	L
1	Валовая добыча каменного угля в Западной Сибири (тыс.тонн) Источник:Л.Б.Кафенгауз Эволюция промышленного про											
2	Исходные данные		Индексы									
3			Абс приросты		Темпы роста		Темпы прироста					
4	Год	Добыча	Цепной	Базисный	Цепной	Базисный	Цепной	Базисный (1887)				
5	1887	14,4										
6	1888	18,0	3,6	3,6	1,25	1,3	0,25	0,3				
7	1889	17,0	-1,0	2,6	0,94	1,2	-0,06	0,2				
8	1890	19,3	2,3	4,9	1,14	1,3	0,14	0,3				
9	1891	21,1	1,8	6,7	1,09	1,5	0,09	0,5				
10	1892	21,3	0,2	6,9	1,01	1,5	0,01	0,5				
11	1893	18,3	-3,0	3,9	0,86	1,3	-0,14	0,3				
12	1894	22,1	3,8	7,7	1,21	1,5	0,21	0,5				
13	1895	23,5	1,4	9,1	1,06	1,6	0,06	0,6				
14	1896	24,9	1,4	10,5	1,06	1,7	0,06	0,7				
15	1897	11,6	-13,3	-2,8	0,47	0,8	-0,53	-0,2				
16	1898	22,9	11,3	8,5	1,97	1,6	0,97	0,6				
17	1899	72,0	49,1	57,6	3,14	5,0	2,14	4,0				
18	1900	153,8	81,8	139,4	2,14	10,7	1,14	9,7				
19	1901	236,1	82,3	221,7	1,54	16,4	0,54	15,4				
20	1902	208,0	-28,1	193,6	0,88	14,4	-0,12	13,4				
21	1903	249,3	41,3	234,9	1,20	17,3	0,20	16,3				
22	1904	308,8	59,5	294,4	1,24	21,4	0,24	20,4				
23	1905	439,6	130,8	425,2	1,42	30,5	0,42	29,5				
24	1906	494,4	54,8	480,0	1,12	34,3	0,12	33,3				
25	1907	516,2	21,8	501,8	1,04	35,8	0,04	34,8				
26	1908	596,6	80,4	582,2	1,16	41,4	0,16	40,4				
27	1909	555,8	-40,8	541,4	0,93	38,6	-0,07	37,6				
28	1910	516,8	-39,0	502,4	0,93	35,9	-0,07	34,9				
29	1911	534,9	18,1	520,5	1,04	37,1	0,04	36,1				
30	1912	707,8	172,9	693,4	1,32	49,2	0,32	48,2				
31	1913	878,0	170,2	863,6	1,24	61,0	0,24	60,0				
32												

Западная Сибирь / Лист1 / Лист2 / Лист3

Готово

Рис. 8.1. Пакет Excel. Анализ динамики добычи каменного угля в Западной Сибири

Средний темп роста (средний коэффициент роста [1]) характеризует интенсивность изменения процесса в среднем за период, другими словами, показывает, во сколько раз изменяется уровень ряда в среднем за период, и определяется по формуле:

$$T_p = \sqrt[n-1]{\frac{X_n}{X_1}} \qquad (1)$$

где x_n и x_1 – последний и первый уровни ряда соответственно.

[1] Иногда в литературе под средним темпом роста подразумевают средний коэффициент роста, выраженный в процентах.

Пример 8.2. Пользуясь этой формулой, попробуем вычислить средние за период 1887–1913 гг. темпы роста добычи каменного угля в Западной Сибири. Заметим, что при использовании формулы среднего темпа роста возникают сложности с вычислением корня (n–1)-ой степени. Однако, если прологарифмировать обе части вышеприведенного равенства, получим:

$$Ln(T_p) = \frac{1}{n-1} \times \left(LnX_n - LnX_1 \right) \qquad (2)$$

В таком виде правая часть выражения легко вычисляется, а для окончательного расчета темпа роста достаточно выполнить обратное действие – потенцировать обе части выражения.

Воспользуемся этим алгоритмом расчета для решения поставленной задачи в пакете Excel. Поставим курсор в ячейку **F32**, где будет получен промежуточный результат, то есть логарифм темпа роста, и перепишем правую часть выражения (2) следующим образом:

=1/26*(ln(b31)–ln(b5)).

Остается вычислить темп роста. Поставив курсор в ячейку **F33,** напишем

=exp(f32)

и получим окончательный результат: в среднем за год добыча каменного угля в этом регионе увеличивалась в 1,17 раза, то есть ежегодный прирост в рассматриваемом периоде в среднем составлял 17%.

8.2. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

8.2.1. Составляющие временного ряда

Временные ряды, составленные из показателей, характеризующих социально-экономические процессы, имеют свою специфику. Прежде всего, эти показатели нередко включают более или менее выраженную, общую для достаточно длительного периода тенденцию к возрастанию или убыванию во времени (так называемый **временной тренд**). Исследование временного ряда, как правило, начинается с выявления и анализа именно этой компоненты.

Составляющей многих временных рядов являются **сезонные колебания,** которые отражают свойственную природе и человеческой деятельности повторяемость процессов во времени. Сезонные колебания, которые представляют собой последовательность почти повторяющихся циклов, чаще всего являются результатом влияния смен времен года на те или иные явления и их характеристики. Например, цены на сельскохозяйственные товары, объемы перевозок некоторых грузов, показатели производственной деятельности отраслей промышленности, связанных с переработкой сель-

скохозяйственного сырья, спрос на некоторые товары и многие другие показатели содержат сезонные колебания.

Иногда временные ряды включают растянутые на несколько лет периоды (циклы), в которых подъемы сменяются спадами, при этом длина и амплитуда этих периодов могут меняться во времени. Эту составляющую часть ряда связывают с цикличностью экономики и называют **циклическими колебаниями.**

Кроме перечисленных составляющих, характеризующих устойчивые закономерные изменения показателя, временные ряды содержат также иррегулярные, хаотические, случайные изменения или так называемую **случайную компоненту.**

Цели и стадии анализа временных рядов. Цель анализа временных рядов – выявление закономерностей в изучаемых процессах и сжатое описание характерных особенностей рядов, отражающих отдельные стороны этих процессов. Конечной целью такого анализа является построение модели, чаще всего математической, способной объяснить поведение этого ряда.

Специфические особенности временных рядов определяют последовательность их обработки или, другими словами, диктуют стадии анализа временных рядов. На практике анализ временных рядов проходит обычно следующие этапы:

- Графическое представление и визуальный анализ временного ряда. Поскольку график временного ряда дает возможность исследователю увидеть наличие основной тенденции ряда и характер этой тенденции, предположить присутствие сезонных и циклических составляющих, то анализ графического представления, как правило, определяет дальнейшее направление обработки и анализа этого ряда;
- Выделение и удаление основных составляющих временного ряда – тренда, сезонных и циклических колебаний;
- Анализ случайной компоненты (**остатков** после удаления выявленных основных составляющих) с целью проверки адекватности полученной на предыдущей стадии модели.

8.2.1.1. Временной тренд

Выявление и выделение временного тренда – важный момент анализа временного ряда. **Временной тренд** отражает существенные типические черты данного процесса и является результатом длительного воздействия на изучаемый процесс определенного комплекса факторов. Существует несколько способов, позволяющих определить основную тенденцию временного ряда, выделить временной тренд.

Выявить общую тенденцию временного ряда можно, воспользовавшись средним абсолютным приростом или средним темпом роста. Идея этого способа проста – предполагается, что ряд изменяется во времени равномерно, то есть каждый последующий уровень ряда возрастает или убывает на величину, равную среднему абсолютному приросту, или каждый последующий уровень ряда изменяется в отношении, определяемом средним темпом роста.

Выбор показателя для выравнивания зависит от вида исходного ряда. Если ряд изменяется приблизительно по закону арифметической прогрессии, то есть каждый уровень ряда отличается от соседнего на постоянную величину, то используют средний абсолютный прирост. Если нарастание или убывание ряда близко к закону геометрической прогрессии, то есть каждый уровень ряда отличается от соседнего примерно в одно и то же число раз, то применяют средний темп роста. Эта методика дает хорошие результаты только в тех редких случаях, когда ряд изменяется (возрастает или убывает) более или менее равномерно. Попробуйте применить эту технику выравнивания ряда к тем данным, с которыми вы уже работали.

Задание 8.3. Найти тренд ряда, представляющего добычу каменного угля в Западной Сибири, используя: а) средний абсолютный прирост, б) средний коэффициент роста.

Пояснения**:** а) если для нахождения тренда используют средний абсолютный прирост, то за первый уровень ряда принимается фактический уровень, второй уровень получается прибавлением к первому среднего абсолютного прироста. Для получения третьего уровня средний абсолютный прирост суммируется с предыдущим преобразованным уровнем и т.д. Наконец, для определения n-го уровня средний абсолютный прирост прибавляется к (n–1)-му преобразованному уровню. Следуя этому алгоритму, проведите расчеты в пакете Excel.

б) При использовании среднего коэффициента роста для выявления тренда за первый уровень, как и в предыдущем случае, принимается фактический уровень, но для получения второго уровня первый уровень умножается на средний коэффициент роста. Для получения третьего уровня средний коэффициент роста умножается на второй преобразованный уровень и т.д. Сделайте расчеты.

Метод скользящих средних. Механическое сглаживание или ***модель скользящего среднего*** для выравнивания временного ряда является одним из наиболее известных и давно используемых методов. Этот способ активно применяется и в настоящее время, в том числе в пакете STATISTICA. При определении средних значений случайные отклонения погашаются – на такой идее основывается метод скользящей средней. Фактические значе-

ния ряда заменяются локальными средними значениями, в которых в той или иной степени нейтрализуются случайные колебания, и выступает закономерная составляющая.

Процедура сглаживания начинается с выбора интервала сглаживания. С одной стороны, чем больше выбранный интервал, тем более плавным получается тренд. Но, с другой стороны, увеличение интервала сглаживания ведет к потере информации, так как укорачивается преобразованный ряд (действительно, если интервал сглаживания равен трем, то нельзя выполнить описанную выше процедуру сглаживания для первой и последней точки ряда, если этот интервал равен пяти – нельзя выполнить процедуру сглаживания для двух начальных и двух конечных точек и т.д.). Легко выбрать длину интервала сглаживания, если ряд имеет более или менее выраженную цикличность. В этом случае длина цикла принимается за интервал сглаживания. Подробнее об использовании модели скользящих средних для выявления и удаления сезонной составляющей будет сказано ниже.

Замечание. От величины интервала сглаживания зависит техника расчета. Если число уровней интервала сглаживания нечетное, то рассчитанное значение средней однозначно приписывается срединному уровню интервала. Затем интервал смещается на один уровень (скользит – отсюда и название метода), снова рассчитывается средняя арифметическая, которая приписывается среднему члену смещенного интервала, и т. д. Если число уровней интервала сглаживания четное, то серединой этого интервала будет промежуток между уровнями, и рассчитанное значение средней нельзя отнести ни к одному уровню ряда. Эту сложность обходят с помощью *центрирования*. Интервал сдвигается еще на один уровень, причем середина его приходится на соседний промежуток. Полусумма из средних арифметических, вычисленных для соседних интервалов, приписывается уровню, расположенному в середине объединения этих интервалов.

Метод скользящих средних весьма прост в исполнении, дает неплохие результаты в выделении тренда, но имеет и свои недостатки. Основным недостатком метода является потеря крайних (начальных и конечных) уровней ряда. Увеличение интервала сглаживания обычно приводит к лучшим результатам выравнивания, но с увеличением этого интервала растут и потери уровней ряда, что особенно чувствительно для коротких рядов. Кроме того, тренд, полученный способом скользящей средней, не имеет математического выражения, что делает значительно более сложным анализ тенденции ряда, в особенности сравнительный анализ тенденций нескольких рядов.

Задание 8.4. Сгладить ряд, представляющий добычу каменного угля в Западной Сибири, используя: а) 3-членную скользящую среднюю, б) 4-членную скользящую среднюю.

Пояснения: а) для получения первого уровня сглаженного ряда нужно рассчитать среднюю арифметическую для трех первых уровней исходного ряда и приписать это значение срединной точке интервала сглаживания, т.е. уровню, соответствующему 1891 г. Поставьте курсор в ячейку на пересечение столбца, предназначенного для получения результатов сглаживания, и строки, соответствующей 1891 г., и воспользуйтесь функцией СРЗНАЧ для расчета первой 3-членной средней арифметической. Копируя формулу для расчета остальных значений скользящей средней, не забудьте, что последнее вычисленное значение будет относиться к 1912 г.

б) для получения первого уровня сглаженного ряда нужно рассчитать среднюю арифметическую для четырех первых уровней исходного ряда и отнести это значение к срединной точке интервала сглаживания, т.е. к промежутку между 1891 и 1892 гг. Затем, сдвинув интервал сглаживания на один уровень, приписать полученную среднюю к промежутку между 1892 и 1893 гг. Чтобы получить скользящую среднюю, относящуюся к определенному году, применим центрирование, т. е. определим "центр" рядом расположенных интервалов (1891–1892 гг. и 1892–1893 гг.), а также найдем "центр" или среднее значение пары найденных скользящих средних, относящихся к этим интервалам. "Центром" интервалов является 1892 г., а среднее значение пары скользящих средних вычислим как полусумму этой пары. Таким образом, для первого уровня сглаженного ряда, который будет относиться к 1892 г., формула для расчета в Excel будет выглядеть так:

=(СРЗНАЧ(B5:B8)+СРЗНАЧ(B6:B9))/2

Копируя формулу для расчета остальных значений скользящей средней, не забудьте, что последнее вычисленное значение будет относиться к 1911 г.

В заданиях 8.3 и 8.4 была использована различная техника анализа временных рядов. На рис. 8.2 приведены графические результаты всех использованных нами вариантов выравнивания и сглаживания ряда. Самое грубое приближение ряда получено с помощью среднего абсолютного прироста. Это вполне объяснимо – поведение исходного ряда мало напоминает закон арифметической прогрессии. Лучшее приближение дает использование среднего темпа роста. Естественно, что хорошее приближение исходного ряда дают скользящие средние, но тенденция изменения ряда, выявленная таким способом, не имеет четкого, например, математического выражения. Остается также открытым вопрос о том, выявлены ли основные составляющие ряда.

Аналитическое выравнивание. Эффективным способом определения общей тенденции ряда является **аналитическое выравнивание**, то есть подбор подходящей математической модели тренда. Прежде всего, это – полиномиальная модель:

$$x_t = b_0 + b_1 \cdot t + b_2 \cdot t^2 + \ldots + b_n \cdot t^n,$$

частным случаем которой является простая линейная модель:

$$x_t = b_0 + b_1 \cdot t.$$

На практике последняя используется наиболее часто [1].

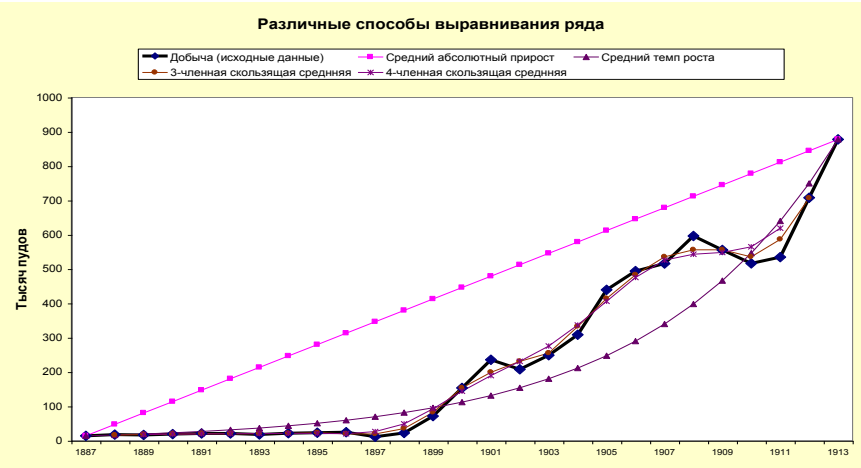


Рис. 8.2. Пакет Excel. Сравнение различных способов выравнивания ряда

Главный вопрос, требующий решения перед аналитическим выравниванием, – выбор функции для выравнивания. Тип функции может быть подсказан графиком, изображающим анализируемый временной ряд.

Пример 8.3. В файле Cotton.sta приведены данные о вывозе хлопка из Средней Азии со станций железной дороги. Выявим закономерности поведения этого ряда.

Обратимся к разделу **Временные ряды и прогнозирование** модуля **Углубленные методы анализа** пакета STATISTICA. На экране появится диалоговое окно. На рис. 8.3 изображено это окно, уже подготовленное к работе с данными, то есть открыт нужный файл (с помощью графической кнопки **Открыть**) и определены переменные для анализа (как обычно, посредством клавиши **Переменные**).

Для визуализации исходного ряда нажмем клавишу в правом верхнем углу, которая называется **ОК (преобразования, авто- и кросскорреляции, графики)**. Раскрывается диалоговое окно – **Преобразования переменных**,

[1] Среди множества других моделей, годящихся для определения тренда, следует отметить логарифмическую модель: $x_t = exp(b_0 + b_1 \cdot t)$. Эта модель хорошо описывает ряды, имеющие тенденцию сохранять постоянные темпы прироста, и нашла широкое применение для анализа временных рядов экономических показателей.

в котором, в частности, есть возможность посмотреть помеченную переменную. Для этого надо перейти на вкладку **Графики** (рис. 8.4).

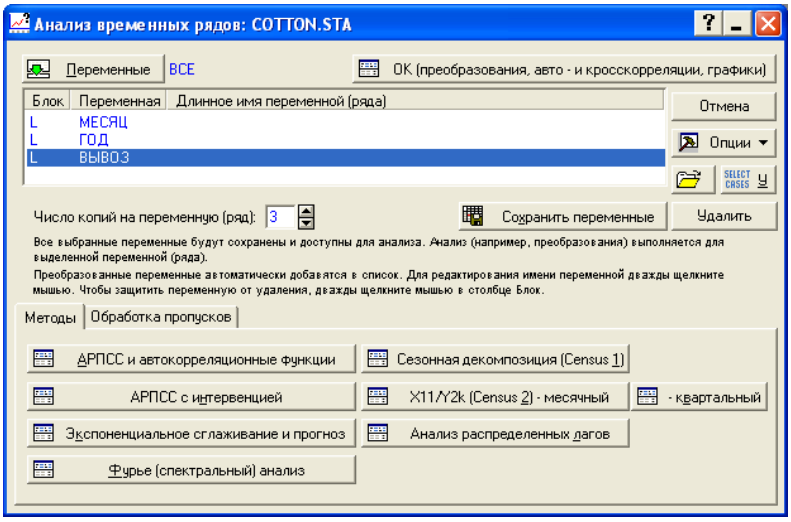


*Рис. 8.3. Основное диалоговое окно **Анализа временных рядов***

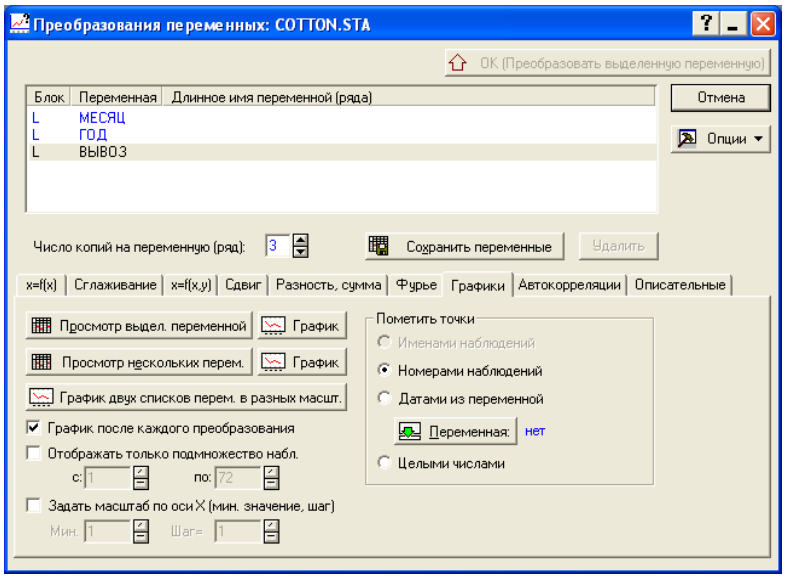


Рис. 8.4. Диалоговое окно преобразования переменных

Запросим графическое изображение ряда, щелкнув мышью на самой верхней графической кнопке **График**. На экране появится график вывоза хлопка за рассматриваемый период (рис. 8.5). На графике рис. 8.5 видно, что данные содержат сезонные колебания, а также слабый линейный тренд.

Рис. 8.5. График вывоза хлопка из Средней Азии (с сентября 1902 г. по август 1908 г.)

Удаление тренда**.** Для выделения и устранения этого тренда в диалоговом окне преобразования переменных (рис. 8.4) перейдем на вкладку **x = f(x)** (вычисление тренда) и установим переключатель **Вычесть тренд** (см. рис. 8.6). Обратите внимание на то, что параметры линейного уравнения тренда уже посчитаны программой, поскольку это предусмотрено по умолчанию (установлен флажок **Оценить a/b из данных**).

После щелчка по клавише **OK (преобразовать выделенную переменную)** можно увидеть график переменной с удаленным трендом. Затем программа возвращается в диалоговое окно **Преобразования перемеенных**, но теперь к списку переменных добавлена новая переменная – результат удаления линейного тренда.

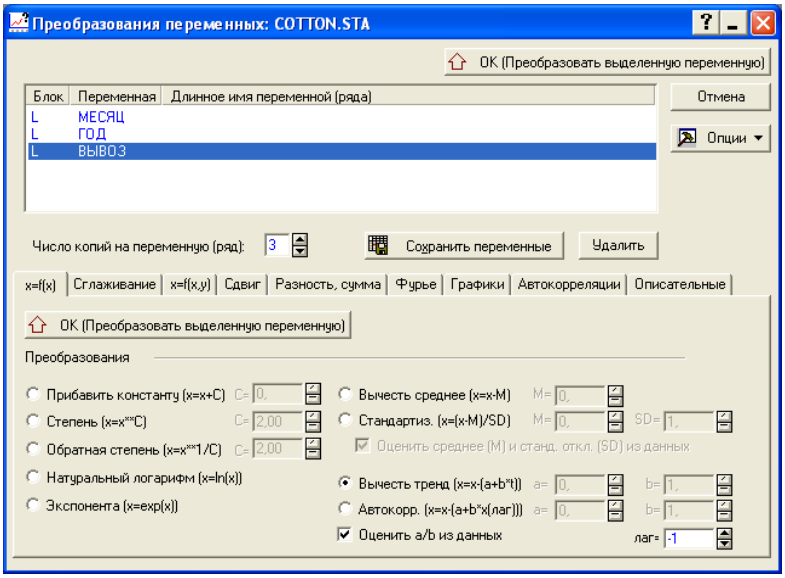


Рис. 8.6. Диалоговое окно процедур преобразования временного ряда

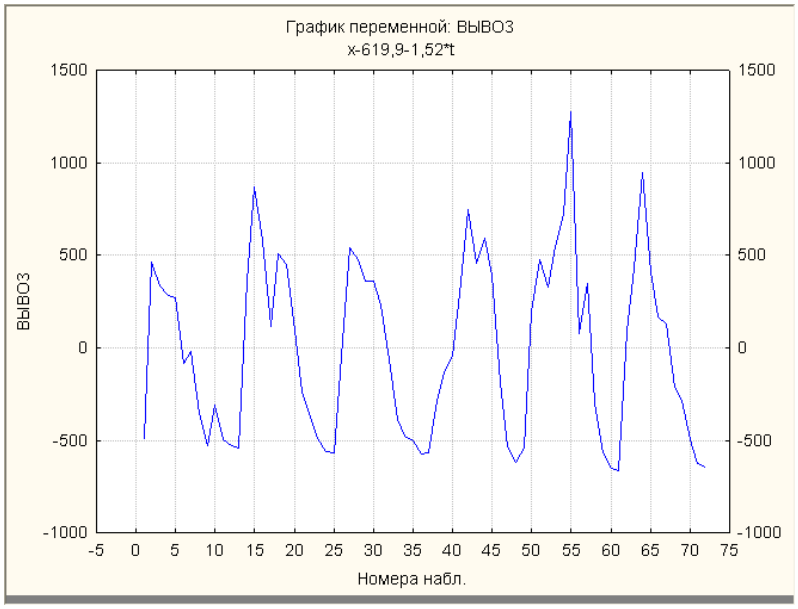


Рис. 8.7. График остатков после удаления линейного тренда

Пометив эту новую переменную, можно еще раз посмотреть ее график. Щелчок мышью по графической кнопке **График** на вкладке **Графики** – и график перед вами (рис. 8.7). В заголовке графика приведено уравнение тренда. Итак, линейный тренд, если его записать в виде уравнения, подставив вычисленные значения параметров, выглядит следующим образом:

$x = 619{,}9+1{,}524 \cdot t$, где t – время.

Это уравнение следует интерпретировать следующим образом: если не учитывать сезонную и случайную составляющие ряда, можно сказать, что в соответствии с выявленной тенденцией, в начале периода (сентябрь 1902 г.) вывоз хлопка из Средней Азии должен был составлять около 620 тысяч тонн, и в среднем за месяц его величина должна была возрастать на полторы тысячи тонны.

8.2.1.2. Анализ остатков после удаления тренда

После выявления тренда ряда следует провести анализ остатков, чтобы оценить степень адекватности полученной модели. Удаленный тренд можно считать адекватным, хорошо приближающим исходный временной ряд, если получившиеся остатки не показывают наличия определенных закономерностей (сезонных и/или циклических). Значит, если остатки не содержат периодичности, они должны быть некоррелированы, а их распределение близко к нормальному. Таким образом, необходимо прояснить два момента:

- Можно ли считать остатки некоррелированными?
- Насколько распределение остатков согласуется с нормальным распределением?

Проверка коррелированности остатков. Связь уровней внутри ряда можно обнаружить и оценить с помощью коэффициентов корреляции, измеряющих связь ряда с самим собой, т.е. с исходным рядом, но сдвинутым на несколько точек по оси времени. Эти коэффициенты называются **коэффициентами автокорреляции**. Проверка коррелированности остатков осуществляется с помощью выборочной **автокорреляционной функции** [1], которая состоит из последовательности значений коэффициентов автокорреляции, соответствующих сдвигам на 1, 2, ... k точек по оси времени. График этой функции называется **коррелограммой**. На коррелограмме кроме значений самой функции, как правило, указываются доверительные интервалы этой функции, построенные для проверки гипотезы об отсутствии ав-

[1] В англоязычной литературе используется термин *serial correlation* (сериальная корреляция).

токорреляции. Если оценки выборочной автокорреляционной функции укладываются в эти пределы, у нас нет оснований отклонять гипотезу о некоррелированности остатков.

Пример 8.4. Вернемся к файлу Cotton.sta и для выяснения коррелированности остатков ряда после удаления тренда (см. пример 8.3) вычислим оценки их автокорреляционной функции, перейдя в диалоговом окне **Преобразования переменных** (рис. 8.4) на вкладку **Автокорреляции**. Программа выведет коррелограмму (график автокорреляционной функции) [1]. Вид коррелограммы свидетельствует о наличии сезонной взаимосвязанности уровней ряда – явно просматривается сезонная волна (рис. 8.8). Поэтому следующий шаг в выявлении закономерностей поведения ряда – включение в модель сезонной компоненты.

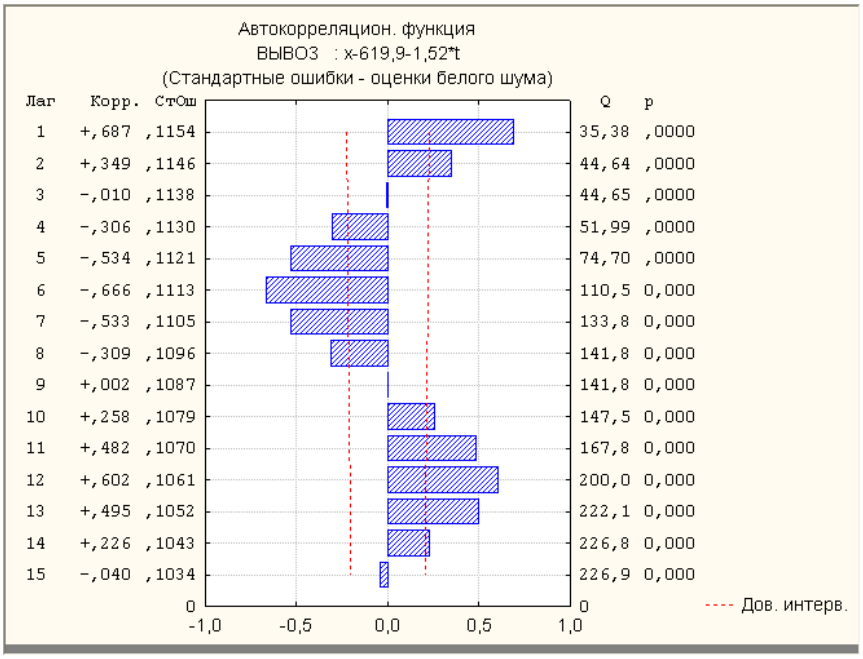


Рис. 8.8. Автокорреляционная функция остатков после удаления тренда

[1] В практических расчетах число членов автокорреляционной функции не должно превышать *n/4*, где *n* – число членов ряда. По умолчанию в этой процедуре установлено максимальное число сдвигов (лагов) автокорреляционной функции, равное 15. Для более коротких рядов этот параметр следует скорректировать.

8.2.1.3. Сезонная составляющая

Многие временные ряды экономических показателей содержат в себе сезонную волну, что необходимо учитывать при анализе этих рядов. Скажем, перевозки хлеба по железной дороге возрастают осенью и в начале зимы в связи с реализацией урожая и уменьшаются весной и летом. Если на первый план выдвигается элиминирование сезонных колебаний с целью изучения остающихся составляющих, то можно использовать простейшие приемы построения ряда, исключающие сезонную волну.

Так, можно составить ряд, уровни которого относятся к определенной фиксированной дате каждого года, или в качестве уровней брать среднегодовые показатели, которые "гасят" сезонную волну. Эти простые приемы элиминируют сезонные колебания, но приводят к значительной потере информации (например, замена помесячных данных погодовыми укорачивает ряд в 12 раз). Потери тем чувствительнее, чем короче исходные ряды.

Если предметом изучения являются сами сезонные колебания или агрегирование показателей по каким-либо причинам нецелесообразно, то ставится задача выделения и/или удаления сезонной компоненты ряда. Для выделения и удаления сезонной составляющей широко используется метод скользящих средних [1].

Сезонная компонента может быть по природе своей *аддитивной* или *мультипликативной*. Если на протяжении всего периода в отдельные месяцы года происходит увеличение или уменьшение показателя на более или менее постоянные величины, то сезонная компонента аддитивна. Если же сезонные колебания пропорциональны среднему значению процесса в рассматриваемый момент времени, то можно говорить о мультипликативной модели. Различие этих моделей можно увидеть на графике. Если амплитуда сезонных колебаний остается более или менее постоянной на протяжении всего рассматриваемого периода, даже при общей тенденции ряда к повышению или понижению, то налицо аддитивная модель сезонной компоненты. В случае мультипликативной модели амплитуда сезонных колебаний варьирует в зависимости от изменения общего тренда ряда.

Закроем график автокорреляционной функции и вернемся в диалоговое окно **Преобразования переменных**. В этом окне выберем графическую кнопку **Отмена** и вернемся в основное диалоговое окно – **Анализ временных рядов** (см. рис. 8.3). Для выявления сезонной составляющей воспользуемся классической техникой сезонной декомпозиции, известной как ме-

[1] В практических расчетах сезонный временной ряд должен быть достаточно длинным и в пять-шесть раз превосходить длину периода сезонности.

тод *Census 1*. Для перехода к соответствующей процедуре нажмем графическую кнопку **Сезонная декомпозиция (Census 1).**

В диалоговом окне (см. рис 8.9), которое откроется в результате нажатия клавиши, прежде всего необходимо выбрать анализируемую переменную (в данном случае ВЫВОЗ), а также тип модели (блок **Сезонная модель**). Судя по графику ряда, мы имеем дело с аддитивной моделью – выберем переключатель **Аддитивная**. Поскольку величина периода – число четное (12 месяцев), следует центрировать скользящие средние, т.е. поставить флажок **Центрировать скользящие средние**. После задания всех необходимых параметров, нажмем графическую кнопку **OK (выполнить сезонную декомпозицию)** и получим результаты расчетов процедуры сезонной декомпозиции – шесть новых переменных, которые перечислены ниже. Поясним смысл и способ получения каждой из этих переменных.

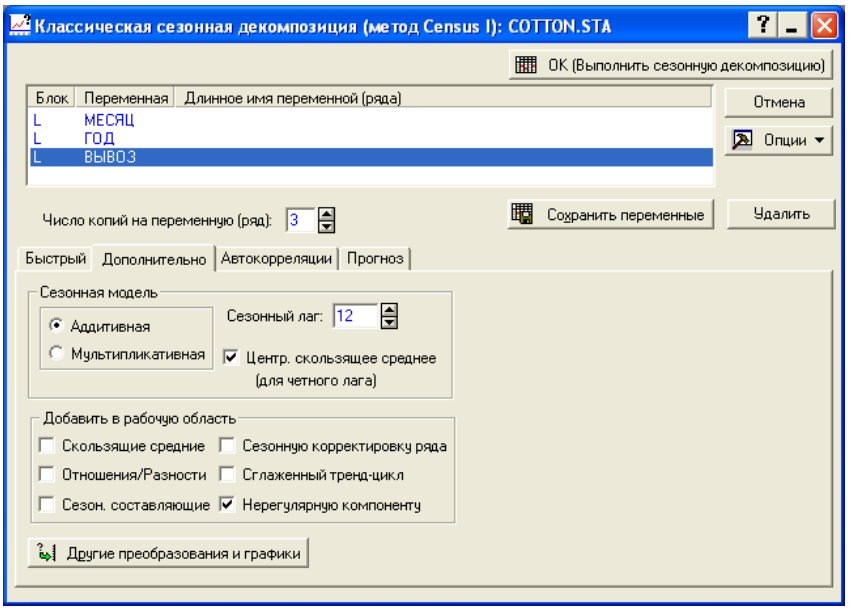


Рис. 8.9. Диалоговое окно сезонной декомпозиции (метод Census 1)

Итак, при выделении сезонной компоненты создаются шесть новых переменных:

1) *Скользящие средние*. Для выделения сезонной составляющей прежде всего рассчитываются k-членные скользящие средние, где k – длина сезонной волны (например, для помесячных данных k равно 12, т.е. числу меся-

цев в году, для квартальных данных $k = 4$). Полученный ряд можно рассматривать как ряд, свободный от сезонных колебаний.

2) *Разности*. Скользящая средняя вычитается из исходного ряда (для мультипликативной модели уровни ряда делятся на соответствующие значения скользящей средней). Вычисленные таким образом индексы показывают, на сколько уровни исходного ряда отличаются от уровней ряда, в котором сезонная волна элиминирована. Таким образом, в индексах отражаются эффекты сезонности, их можно рассматривать как результат влияния сезонных изменений ряда.

3) *Сезонные составляющие* (сезонная волна). Поскольку на поведение исходного ряда влияют всевозможные случайные причины, от них не свободны и рассчитанные отношения. Поэтому индексы, относящиеся к одному и тому же месяцу (кварталу), но к разным годам, колеблются и, тем самым, только приближенно представляют сезонные колебания. Чтобы уловить типичные черты сезонности, индексы для каждого месяца (квартала) усредняют за ряд лет и распространяют полученные оценки сезонных эффектов на весь рассматриваемый период. Результатом этого расчета и является третья новая переменная.

Чтобы увидеть, как выглядит сезонная волна, для столбца **Сезонные состаляющие** воспользуемся контекстным меню и выберем в нем команду **Графики блоковых данных | Линейный график: по столбцам**.

Полученная в результате этих действий диаграмма представлена на рис. 8.10. На оси *X* приведены порядковые номера месяцев сельскохозяйственного года (начиная с сентября). В сентябре вывоз хлопка из Средней Азии был относительно невелик, в октябре он скачком увеличивался примерно в 4 раза, ноябрь–март были периодом наиболее интенсивного вывоза, после которого наступал довольно плавный спад.

4) *Скорректированный ряд* (т.е. освобожденный от сезонной компоненты). На этом шаге из исходного ряда удаляется сезонная компонента путем вычитания сезонного фактора из всех уровней ряда (в случае мультипликативной модели – путем деления уровней исходного ряда на соответствующий сезонный фактор). Скорректированный временной ряд представляет в таблице четвертая переменная.

5) Пятая переменная представляет собой составляющую скорректированного ряда, которая объединяет тренд и возможную циклическую компоненту (*Сглаженный тренд – цикл*). Она рассчитана как взвешенная [1] 5-членная скользящая средняя ряда с удаленной сезонной компонентой.

[1] В качестве весов взяты следующие числа – 1, 2, 3, 2, 1.

6) Наконец, шестая переменная – это случайная компонента (*Нерегулярная компонента*). Она получена удалением из исходного ряда тренда, циклической (если она есть) и сезонной компонент.

Эти переменные будут выведены в таблице, они также (все или частично) могут быть добавлены в рабочую область для дальнейшего анализа. Пока ограничимся введением в рабочую область только случайной составляющей для анализа остатков.

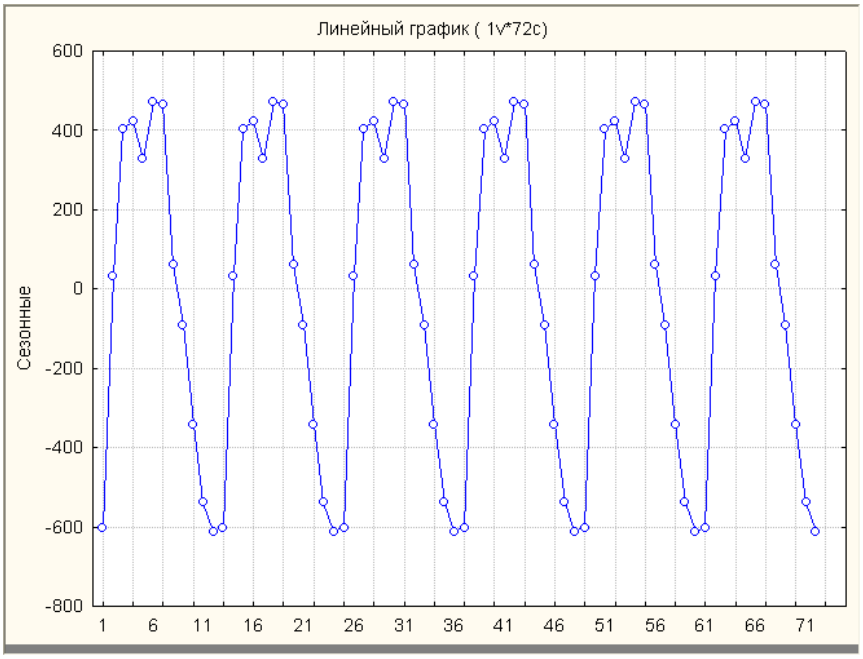


Рис. 8.10. Сезонная волна (усредненные сезонные индексы) для ряда вывоза хлопка из Средней Азии

8.2.1.4. Анализ остатков после выделения сезонной составляющей

Итак, мы выявили и выделили детерминированные составляющие ряда. Насколько хорошо они описывают анализируемый ряд? Как ведут себя полученные остатки?

Обратите внимание, что мы заранее позаботились о вводе случайной компоненты в рабочую область (это предусмотрено в блоке **Добавить в рабочую область**, где в нашем случае поставлен флажок **Нерегулярная компонента** – см. рис. 8.9). Проанализируем эту компоненту ряда.

Прежде всего, посмотрим, как она выглядит. Вернемся к вкладке **Графики** диалогового окна **Преобразования переменных** (рис. 8.4), пометим в рабочей области *нерегулярную компоненту* ряда и построим ее график, щелкнув по графической кнопке *График* (см. рис. 8.11).

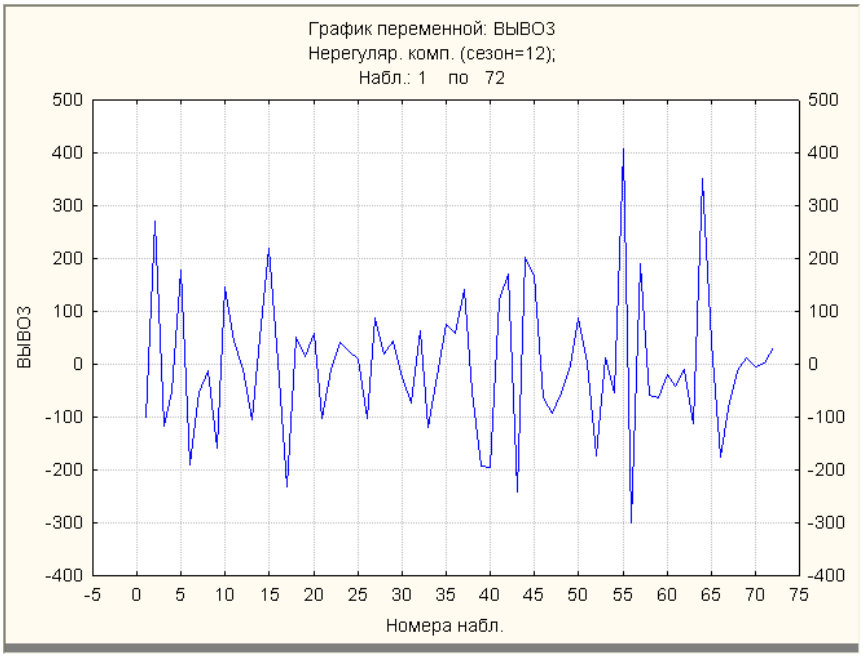


Рис. 8.11. Случайная компонента (остатки ряда) после исключения тренда, сезонной и циклической составляющих

В сравнении с исходным рядом поведение полученных остатков значительно больше напоминает поведение независимой случайной величины. Для большей определенности выводов вновь проведем проверку остатков на коррелированность.

Перейдя на вкладку **Автокорреляции**, мы получим коррелограмму, т.е. график автокорреляционной функции для остатков ряда (см. рис. 8.12). Заметим, что не все полученные оценки лежат внутри доверительного интервала для нулевых значений автокорреляционной функции (оценки для лага, равного 1, и для лага, равного 13, выходят за его пределы), однако график показывает, что удаление сезонной составляющей из ряда существенно уменьшило зависимость его членов (ср. с рис. 8.7).

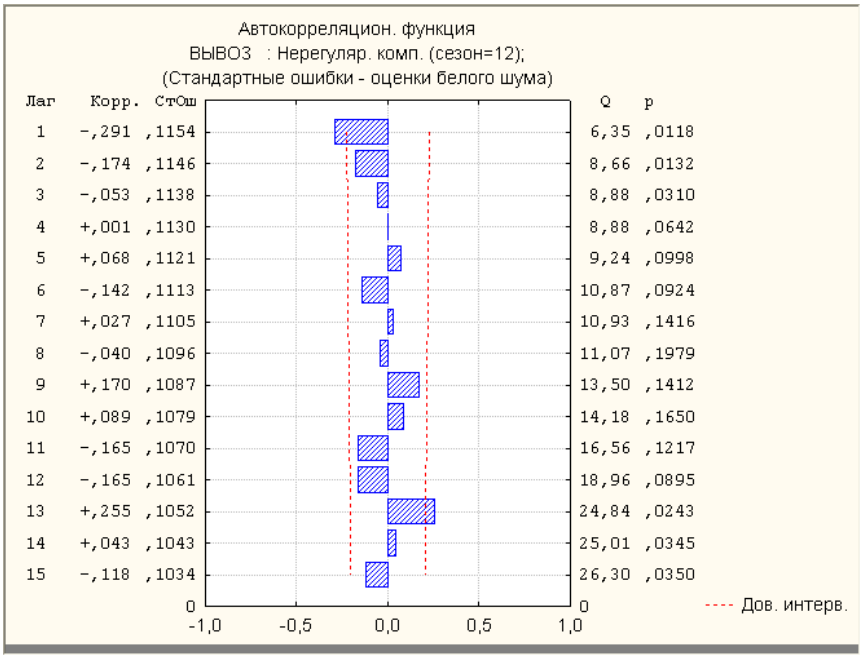


Рис. 8.12. Автокорреляционная функция остатков ряда после исключения тренда, сезонной и циклической составляющих

Проверка гипотезы о нормальности распределения остатков**.** В математической статистике разработаны разные способы проверки нормальности распределения (см. главу 3). В этой главе для проверки соответствия распределения остатков нормальному закону распределения мы будем пользоваться *графиком остатков, построенным на т.н. нормальной вероятностной бумаге*. Это один из наиболее простых способов проверки нормальности распределения. Он сводится к оценке отклонений специальным образом рассчитанной эмпирической функции распределения от прямой линии. Достоинством этого метода является легкость оценки близости распределения к нормальному – чем лучше точки этой эмпирической функции группируются около прямой линии, тем лучше данные согласуются с нормальным законом распределения.

Пример 8.5. Вновь вернемся в файлу Cotton.sta. Для проверки соответствия распределения полученных остатков (напомним, что в результате выполнения двух предыдущих примеров из данных были удалены тренд и сезонная компонента) нормальному распределению построим соответствующий график остатков, перейдя на вкладку **Описательные** и щелкнув по графической кнопке **Нормальный график**. Результат представлен на

рис. 8.13. Видно, что распределение остатков довольно хорошо согласуется с нормальным распределением за исключением его правой части, где имеются заметные отклонения.

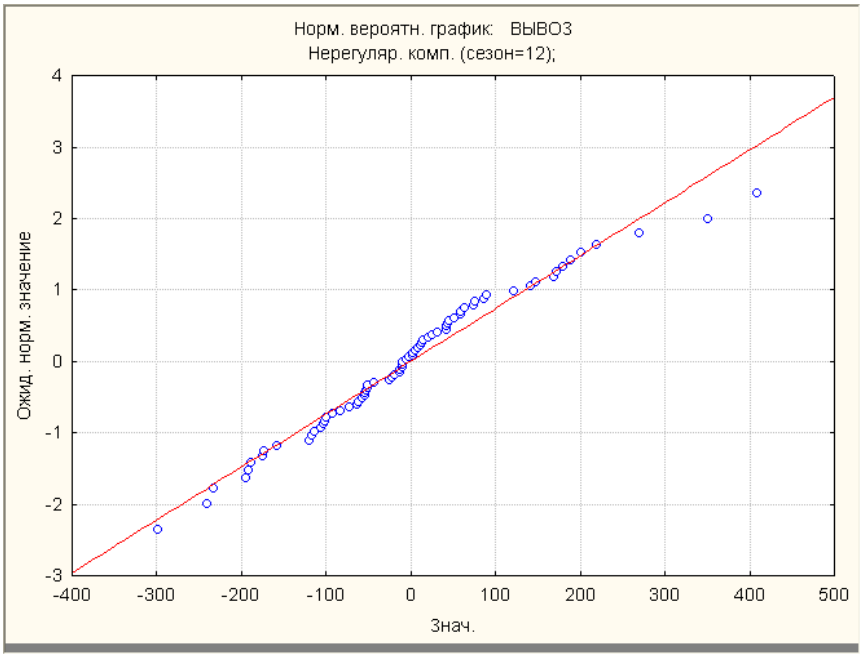


Рис. 8.13. Графический способ проверки гипотезы о нормальности распределения остатков ряда

Подводя итоги анализу, можно сказать, что разбиение временного ряда на детерминированную (тренд, сезонная и циклическая составляющие) и случайную компоненты получилось не идеальным, но вполне удовлетворительным, позволяющим судить о поведении ряда. В принципе, в подобных ситуациях можно иногда добиться лучших результатов, но для этого нужно использовать более тонкие методы исследования, изложение которых выходит за границы введения в анализ временных рядов.

ВОПРОСЫ

1. Назовите основные показатели динамики.
2. Чем базисные показатели отличаются от цепных? Каков информативный смысл базисных индексов?
3. Назовите составляющие временного ряда.

4. Перечислите основные этапы анализа временного ряда.
5. Расскажите о методе скользящих средних. Что такое центрирование?
6. Сколько точек исходного временного ряда вы потеряете при использовании 7-членной скользящей средней?
7. Что понимается под трендом временного ряда?
8. Что понимается под сезонной компонентой ряда?
9. Аддитивная и мультипликативная модели сезонной компоненты.
10. Расскажите о случайной составляющей временного ряда.
11. Зачем нужен анализ остатков?
12. Основные моменты анализа остатков.
13. Что собой представляет коррелограмма? Как ее интерпретировать?
14. Что собой представляет график остатков на нормальной вероятностной бумаге? Его интерпретация.

ЗАДАНИЯ

1. В задании 8.1 были вычислены характеристики динамики добычи каменного угля в Западной Сибири. Рассчитайте эти показатели динамики для а) Донецкого, б) Подмосковного, в) Уральского бассейна (на выбор) и проведите сравнительный анализ развития любых двух бассейнов. Исходные данные – в файле Coal.xls.
2. Вычислите средние абсолютные приросты и средние темпы роста для всех рядов файла Coal.sta. Что вы можете сказать о динамике добычи каменного угля в четырех регионах Российской империи?
3. Подберите тренды для данных о численности населения в США (файл Us_popul.sta) и в России (файл Rus_pop.sta). Какие выводы можно сделать о статистике населения?
4. Подберите модели тренда для данных об урожае а) ржи, б) пшеницы, в) овса, г) ячменя и сделайте анализ остатков. Сравните динамику сбора различных хлебов. Исходные данные – в файле Harvest1.sta.
5. Выявите закономерную составляющую в поведении урожайности зерновых культур в СССР (файл Harvest2.sta). Проведите анализ остатков. Интерпретируйте полученные результаты.

ПРИЛОЖЕНИЕ

ТАБЛИЦЫ, СОДЕРЖАЩИЕ ДАННЫЕ ИСТОРИЧЕСКИХ ИСТОЧНИКОВ, ИСПОЛЬЗУЕМЫХ НА ПРАКТИЧЕСКИХ ЗАНЯТИЯХ

Все представленные таблицы имеются в электронной форме и входят в банк заданий по II практикуму курса "Информатика и математика"

СОЦИАЛЬНО-ЭКОНОМИЧЕСКАЯ ИСТОРИЯ

1. Потребительская кооперация в России и ведущих странах Европы в 1908 году (*файл coopera.sta*)

Государства	Число членов на один кооператив	Проц. отношение кооператоров к населению
Россия	157	1,5
Германия	437	9,4
Англия	1671	21,6
Франция	306	7,2
Венгрия	133	2,1
Дания	110	19,2
Голландия	319	8,7
Финляндия	202	13,4
Норвегия	257	14,9
Швеция	184	3,9
Австрия *	461	3,1
Швейцария	684	22,8
Испания	160	0,6
Италия **	224	4,1
Бельгия ***	804	7,8
Итого	395	6,1

* Данные о потребительских обществах, объединенных в центральные союзы.

** Данные на 1910 г.

*** Данные только о социалистических кооперативах.

Источник: РГИА. Ф. 32. Оп. 2. Д. 53. Л. 2. Записка К. Комаровского "О состоянии потребительской кооперации в России".

2. Основные показатели общего уровня развития стран в 1987 году
(файл *tab_1987.sta*)

	Показатели в 1987 г.							
Страна	1	2	3	4	5	6	7	8
СССР	5975	5911	914	829	12	70	45	311
Австрия	13421	6711	118	702	50	409	97	440
Великобритания	12583	2778	436	412	52	348	75	433
США	15012	11479	968	1305	92	718	112	798
ФРГ	14384	6909	360	427	66	463	97	383
Франция	13327	6823	394	984	61	438	100	389
Япония	12770	5705	102	130	55	374	36	579

Источник: Миронов Б.Н. История в цифрах. Математика в исторических исследованиях. М., 1991. С. 145, 146, 147, 150, 154, 155, 142, 140.

Показатели:

1 - национальный доход на душу
2 - произ-во электроэнергии на душу
3 - военные расходы на душу
4 - производство зерна на душу
5 - кол-во телефонов на душу
6 - кол-во автомобилей на 1000 жителей
7 - потребление мяса на душу
8 - кол-во телевизоров на 1000 жителей.

3. Численность занятых в обрабатывающей промышленности СССР и США в 1987 г. (тыс. чел.) (файл *workers.sta*)

	СССР	США	СССР в % к
Производство напитков	311,8	172,9	180,3
Табачная промышленность	38,7	63,5	60,9
Пищевая	2910,3	1384,2	210,2
Текстильная	1997,2	698,2	286,1
Швейная	2336,2	1113,8	209,8
Кожевенно-обувная	618,9	135,7	456,1
Лесная и деревообрабатывающая	2564,9	1235,1	207,7
Целлюлозно-бумажная	290,2	654,8	44,3
Химическая	1375,2	1028,4	133,7
Резиновая и изделия из пластмас	477,4	863,3	55,3
Нефтеперерабатывающая и продукты из угля	170,1	153,6	110,7
Стр.мат-лы, стекольная и фарфоро-фаянсовая	2741,4	554,2	494,7
Металлургия	2752,3	2228,9	123,5
Промышленное и транспортное оборудование	5364,9	3752,1	143,6
Электротехническая	3763,8	2732,1	137,8
Прочие	5292,8,	2195,5	241,1
Всего	33006,0	18950,6	174,2

Источник: Кудрин В.М. Советская экономика в ретроспективе. Опыт переосмысления . М., 1997. С. 293.

4. Сопоставление производительности труда в обрабатывающей промышленности СССР и США в 1987 г. (по товарной продукции) (файл *product1.sta*)

Отрасль	В руб. (тыс.)			В долл. (тыс.)		
	СССР	США	СССР к США, %	СССР	США	СССР к США, %
Производство напитков	35,5	151,6	23,4	56,0	273,7	20,5
Табачная промышленность	122,1	98,7	123,7	372,2	326,9	113,9
Пищевая	45,7	167,5	27,3	83,6	204,0	41,0
Текстильная	31,4	96,3	32,6	31,3	89,9	34,8
Швейная	14,2	47,9	29,6	20,1	57,7	34,8
Кожевенно-обувная	15,0	20,7	72,2	46,7	66,9	69,7
Лесная и деревообрабатывающая	12,0	46,9	25,6	26,1	86,8	30,1
Целлюлозно-бумажная	26,9	129,2	20,8	36,9	166,4	22,1
Химическая	33,6	193,1	17,4	40,6	223,2	18,2
Резиновая и изделия из пластмасс	25,3	58,7	43,1	47,5	100,3	47,3
Нефтеперерабатывающая и продукты из угля	129,6	252,2	51,4	450,0	849,0	53,0
Стройматериалы, стекольная и фарфоро-фаянсовая	13,6	40,4	33,6	32,4	110,9	29,2
Металлургия	32,0	59,0	54,3	76,0	120,1	63,3
Промышленное и транспортное оборудование	19,5	33,5	58,2	84,9	140,4	60,5
Электротехническая	13,7	104,9	13,0	20,1	102,0	19,7
Прочие	19,3	53,7	35,9	45,0	88,6	50,7
Всего	22,9	79,3	29,0	51,9	130,7	40,0

Источник: Кудрин В.М. Советская экономика в ретроспективе. Опыт переосмысления. М., 1997. С. 294.

5. Годовая квартирная плата в городах России за квартиру в 1-3 комнаты * (файл *apartmen.sta*)

Группа городов и район	Число городов		Население, тыс. чел.		Квартирная плата за год, руб.		Прирост,%
	1904 г.	1910 г.	1904 г.	1910 г.	1904 г.	1910 г.	
Поселки	175	278	2803	4497	71	109	53,5
Города, до 10 тыс. жителей	552	512	2731	2500	79	116	46,8
10-50 тыс. жителей	318	355	6666	7494	118	172	45,9
50-100	45	57	3070	4107	186	277	48,9
100-200	14	17	2563	2667	220	286	30,0
200-1000	4	7	1627	2897	241361	49,6	
свыше 1000	2	2	2532	3038			
Итого по России (89 губ.)	1110	1228	21992	27200	152	224	47,5
В границах СССР (71 губ.)	875	678	17375	21782	150	208	38,7
В том числе по районам							
Европейская Россия	786	861	15696	19253	141	201	42,5
Царство Польское	121	128	2568	3070	173	328	88,9
Кавказ	104	135	1997	2563	181	211	16,6
Средняя Азия	49	52	1090	1355	197	306	54,8
Сибирь	50	52	641	959	174	282	62,6

* Подсчитано по изданию ЦСК "Города России в 1904 и 1910 гг.". Цифра по Петрограду - 215 руб. для 1900 г. - повышена для 1904 и 1910 гг. на средний процент вздорожания квартир по всем другим городам. Средние квартирные платы по всем группам и районам взвешены по их населению.

Источник: Струмилин С.Г. Статистика и экономика. М., 1979. С. 322.

6. Номинальная и реальная заработная плата в различных западных странах, 1953-1979 гг. (Индекс: 1970 г.=100) (файл wages1.sta)

Вид заработной платы	1953 г.	1955 г.	1960 г.	1965 г.	1970 г.	1973 г.	1976 г.	1979 г.
ФРГ								
Номинальная	27,0	30,7	44,0	66,5	100,0	139,0	181,1	215,9
Реальная	41,8	46,4	55,8	76,2	100,0	115,3	126.6	135.8
Италия								
Номинальная	30,0	25,5	35,5	63,5	100,0	150,0	268,4	437,3
Реальная	37,5	41,3	52,5	74,0	100.0	118.6	126,3	134,8
Бельгия								
Номинальная	36,1	36,8	47,4	68,0	100.0	143.9	228,0	286,1
Реальная	53,5	54,4	64,3	78,8	100,0	123.8	144.7	156,5
Нидерланды								
Номинальная	21,9	26,1	36,5	59,5	100,0	147,4	214,4	264,5
Реальная	40,2	44,7	54,9	73,9	100,0	115,4	126,7	134,4
Франция								
Номинальная	26,3	24,9	40,6	65,0	100,0	138,4	221,1	315,6
Реальная	55,5	51,2	61.9	81,7	100,0	116.0	133.8	146,2
Великобритания								
Номинальная	34,5	37,7	50.8	68,0	100,0	142.3	257,5	366,3
Реальная	61,1	63,4	75,4	85,2	100,0	113,3	121,8	123.1
США								
Номинальная	47,7	50,4	62,9	74,7	100,0	121,6	152,6	192,7
Реальная	69,0	72,0	81,0	89,5	100,0	106,7	105,9	105,8
Япония								
Номинальная	19,3	21,6	29,4	52,9	100,0	158,5	261,3	325,1
Реальная	38,2	40,9	50,9	68,9	100,0	129,6	146,3	154,8

Примечание. Номинальная заработная плата равняется всей совокупности заработка, выплаченного в течение года, разделенного на число занятых в этот период. Реальная заработная плата соответствует номинальной с учетом индекса цен на потребительские товары.
Источники: база данных COMET, по данным национальной статистики, использованной ЕЭС и ОЭСР.

7. Данные об объеме внешней торговли и численности населения по 16 странам мира в 1938 г. (файл *trade.sta*)

Страна	Объем внешней торговли (млрд. ф.ст.)	Ранг по объему торговли	Население (млн. чел.)	Ранг по населению
Великобритания	1,330	1	47,6	5
США	1,024	2	130,0	2
Дания	0,142	11	3,8	15
Франция	0,450	4	42,0	8
Германия	0,882	3	79,2	3
Греция	0,045	16	7,1	13
Голландия	0,276	7	8,7	11
Италия	0,232	8	43,4	7
Япония	0,309	5	72,8	4
Норвегия	0,098	13	2,9	16
Испания	0,052	15	25,6	9
Швеция	0,201	9	6,3	14
Аргентина	0,180	10	13,0	10
Бельгия	0,307	6	8,4	12
Бразилия	0,121	12	44,1	6
Китай	0,085	14	410,0	1

Источник: Кендэл М. Ранговые корреляции. М., 1975. С. 25.

8. Динамика ВВП * России и СССР (1913 г. = 100) (файл *gdp.sta*)

Год	Индекс	Год	Индекс	Год	Индекс	Год	Индекс
1870	36,0	1945	143,6	1961	383,8	1977	720,1
1890	43,2	1946	143,2	1962	394,2	1978	738,2
1900	66,3	1947	159,2	1963	385,2	1979	734,7
1928	99,8	1948	181,0	1964	435,0	1980	735,6
1929	102,6	1949	200,4	1965	459,7	1981	742,3
1930	108,6	1950	219,6	1966	482,0	1982	760,6
1931	110,7	1951	220,6	1967	503,3	1983	784,9
1932	109,5	1952	234,9	1968	532,8	1984	795,0
1933	114,0	1953	245,0	1969	540,3	1985	802,1
1934	125,2	1954	256,9	1970	581,8	1986	835,1
1935	144,1	1955	278,9	1971	597,3	1987	845,9
1936	155,5	1956	305,6	1972	600,7	1988	863,9
1937	171,3	1957	311,8	1973	651,2	1989	876,8
1938	174,4	1958	335,2	1974	670,1	1990	855,6
1939	185,2	1959	331,5	1975	672,0	1991	726,0
1940	180,8	1960	363,0	1976	703,5	1992	587,8

* ВВП – внутренний валовый продукт.

Источник: Maddison A. Monitoring the World Economy, 1820-1992. OECD Development Centre, 1995. P. 154,155. (Наиболее достоверные, по мнению автора, из западных оценок.)

9 Динамика ВНП, занятости и производительности труда в народном хозяйстве СССР (файл *econ.sta*)

Год	ВНП	Занятость	Производительность труда
1928	100,0	100,0	100,0
1929	102,7	99,8	102,9
1930	108,7	96,2	113,0
1931	110,9	98,9	122,1
1932	109,7	100,9	110,9
1933	114,2	105,7	108,0
1934	125,5	112,7	111,4
1935	144,4	122,9	117,5
1936	155,9	119,5	130,5
1937	171,6	125,0	137,3
1938	174,7	130,9	133,5
1939	185,5	135,8	136,6
1940	202,5	159,5	127,0
1945	160,9	168,0	95,8
1946	160,4	156,8	102,3
1947	178,3	155,1	115,0
1948	202,7	162,5	124,7
1949	224,5	172,0	130,5
1950	246,0	183,0	134,4

Примечание. Все данные в виде индексов, 1928 г. = 100.

Источник: Moorsteen R., Powell R. The Soviet Capital Stock, 1928-1962. Homewood Illin., 1966. P.361,365. Данные о занятости рассчитаны с учетом поправки на отработанное число человеко-часов; включена численность вооруженных сил.

10. Индексы цен за 1885-1914 гг. (1913 г. = 100) * (файл *index1.sta*)

Год	Розничные цены				Оптовые цены		
	Институт экон. исследований (27 товаров) по СПб	М.П.Кохна			Проф. М.Е.Подтягина (66 товаров)	С.П.Боброва (62 товара)	
		(24 товара) по СПб	(15 товаров) по Москве	ср. по двум городам		в кредитной валюте	в золотой валюте
1885	74,3	75,6	84,9	80,2	75,4	–	–
1886	71,2	71,8	78,3	75,1	73,1	–	–
1887	66,9	72,4	78,0	75,2	74,1	80,9	67,6
1888	68,0	72,2	80,9	76,5	75,4	86,2	77,0
1889	70,2	73,7	82,5	78,1	72,8	83,4	82,5
1890	67,5	71,1	79,8	75,4	70,4	73,7	80,2
1891	71,1	73,5	82,7	83,6	85,6	74,8	74,9
1892	74,4	77,9	84,8	81,4	88,2	73,2	69,2
1893	74,7	79,4	80,9	80,2	78,8	75,9	74,3
1894	72,2	74,2	78,2	76,2	66,4	68,5	68,9
1895	72,5	70,7	74,8	72,7	62,3	69,9	70,8
1896	72,7	70,1	72,8	71,4	62,2	72,0	72,0
1897	72,9	71,1	74,5	72,8	72,2	74,4	74,4
1898	78,0	78,1	76,5	77,3	79,2	78,2	78,2
1899	79,3	80,3	75,6	78,0	78,3	80,7	80,7
1900	77,8	80,0	77,4	78,7	76,0	84,2	84,2
1901	79,1	78,9	79,0	78,9	77,9	79,8	79,8
1902	80,3	79,4	79,9	79,7	79,0	75,8	75,8
1903	80,8	79,2	78,9	79,0	77,8	72,4	77,4
1904	80,0	80,2	81,7	80,9	80,1	80,1	80,1
1905	80,9	80,4	86,3	83,3	84,6	81,9	81,9

10. Продолжение

Год	Розничные цены				Оптовые цены		
	Институт экон. исследований (27 товаров) по СПб	М.П.Кохна			Проф. М.Е.Подтягина (66 товаров)	С.П.Боброва (62 товара)	
		(24 товара) по СПб	(15 товаров) по Москве	ср. по двум городам		в кредитной валюте	в золотой валюте
1906	82,8	84,0	92,4	88,2	91,0	89,3	89,3
1907	86,1	89,9	96,3	93,1	102,7	97,0	97,0
1908	90,0	95,0	100,8	97,7	103,3	93,0	93,0
1909	90,5	94,2	98,5	96,3	98,6	91,0	91,0
1910	89,0	90,9	96,7	94,3	94,4	91,6	91,6
1911	91,1	91,8	96,4	94,1	94,6	94,8	94,8
1912	97,1	98,1	100,7	99,4	101,5	98,0	98,0
1913	100,0	100,0	100,0	100,0	100,0	100,0	100,0
1914	105,4	–	–	–	107,8	109,2	106,9

* Струмилин С.Г. История черной металлургии в СССР. С. 514-515 (субиндексы по 27 товарам за 1853-1913 гг.).

Примечание. Оптовые индексы Подтягина и Боброва построены примерно на одной и той же ценовой базе ряда крупнейших товарных рынков (с 1890 года по своду товарных цен). Первый взвешен по весам рабочих бюджета, второй – по доле участия каждого товара в грузообороте страны (по водным и железнодорожным перевозкам). Последний индекс дается не только в текущей (кредитной) валюте, но и в переводе по курсу на золото за те годы, когда кредитные рубли отличались от золотых. Все остальные индексы рассчитаны в кредитной валюте.

11. Индексы розничных цен в Санкт-Петербурге с 1867 по 1916 гг. (1913 г. = 100) (файл *index2.sta*)

Год	Продукты			Итого по сельскому хозяйству	Промтовары	Всего
	землед.	животн.	лесовод.			
1867	77,6	54,6	45,2	58,3	133,6	65,0
1868	86,6	55,7	49,3	62,3	132,5	68,5
1869	77,2	63,0	53,4	64,3	130,1	70,2
1870	71,5	66,7	51,1	64,1	143,0	71,1
1871	73,7	63,9	65,7	67,0	128,6	72,4
1872	83,2	66,7	59,8	69,4	142,8	75,9
1873	84,0	63,5	54,8	66,7	133,3	72,6
1874	83,6	64,6	65,3	69,8	123,8	74,6
1875	76,3	63,2	75,8	69,8	123,2	74,6
1876	75,9	63,9	81,3	71,4	122,7	76,0
1877	79,3	65,6	71,7	70,7	115,8	74,7
1878	88,8	65,3	63,0	71,0	119,2	75,2
1879	87,1	71,0	63,0	73,2	144,6	79,6
1880	106,5	75,4	71,2	82,6	109,7	85,0
1881	121,1	78,4	73,5	88,5	120,2	91,3
1882	103,4	79,9	72,6	84,3	134,0	88,7
1883	88,4	75,6	68,5	77,2	136,2	82,5
1884	90,1	74,0	68,9	77,0	122,2	81,0
1885	89,2	67,0	63,9	72,1	105,7	75,1
1886	86,6	65,1	59,8	69,5	89,1	71,3
1887	80,6	70,3	59,8	70,4	85,6	71,7
1888	75,0	71,2	60,3	69,5	96,3	71,9
1889	78,4	65,6	67,1	69,4	97,4	71,9
1890	75,9	63,9	63,5	67,0	99,8	69,9
1891	89,2	66,5	55,2	69,7	103,3	72,7
1892	99,1	71,9	57,1	75,4	100,9	77,7
1893	87,8	72,8	68,5	75,5	112,7	78,8
1894	72,8	69,8	63,9	69,1	106,9	72,5
1895	69,1	65,6	59,4	64,9	99,9	68,0
1896	67,7	63,9	63,9	64,9	96,4	67,7
1897	71,8	64,6	63,0	65,9	103,4	69,3
1898	80,2	68,8	79,0	74,4	106,9	77,3
1899	78,4	71,0	89,9	77,7	98,7	79,5
1900	77,6	75,4	81,3	77,4	95,3	79,0
1901	78,0	71,0	76,7	74,3	106,9	77,2
1902	79,3	73,1	75,3	75,3	102,2	77,7
1903	75,9	73,3	80,8	75,8	94,0	77,5
1904	75,9	77,3	74,4	76,2	97,5	78,1
1905	80,2	74,5	69,9	74,8	101,0	77,2
1906	83,2	78,7	74,4	78,8	103,4	81,0
1907	93,5	86,2	82,2	87,1	98,8	88,2
1908	104,3	94,4	84,5	94,5	98,7	94,9
1909	100,4	95,5	79,4	92,8	99,9	93,4
1910	91,4	93,2	77,6	88,8	98,6	89,7
1911	90,9	93,0	86,8	90,9	98,7	91,6
1912	100,4	96,0	103,6	99,1	94,0	98,6
1913 *	100,0	100,0	100,0	100,0	100,0	100,0
1914	100,0	106,6	105,2	106,0	103,5	105,4

10. Продолжение

Год	Продукты			Итого по сельскому хозяйству	Промтовары	Всего
	землед.	животн.	лесовод.			
1915	120,7	119,2	133,8	123,2	125,8	123,5
1916	158,2	205,8	181,3	187,1	324,6	199,3

* Для 1913 г. веса в наборе: а) продуктов земледелия – 24,1%; продуктов животноводства – 44,3%; продуктов лесоводства – 22,7%; промтоваров – 8,9%.

Источник: Струмилин С.Г. Очерки экономической истории России и СССР. М., 1966. С. 90.

12. Динамика поденной платы строительных рабочих в Санкт-Петербурге и индекса цен с 1853 по 1913 гг. (файл *wages.sta*)

Год	Номинальная поденная плата (коп. серебром)	Индекс цен (1913 г.=100)	Год	Номинальная поденная плата (коп. серебром)	Индекс цен (1913 г.=100)
1853	82,5	46,6	1884	102,8	76,6
1854	82,8	47,5	1885	102,0	74,3
1855	70,7	47,8	1886	85,3	71,2
1856	76,0	49,5	1887	92,6	66,9
1857	71,4	51,5	1888	106,3	68,0
1858	63,0	49,8	1889	116,7	70,2
1859	87,0	49,5	1890	117,3	67,5
1860	114,2	50,7	1891	119,3	71,1
1861	114,0	52,9	1892	120,6	74,4
1862	81,1	55,3	1893	120,4	74,7
1863	90,5	54,5	1894	119,1	72,2
1864	87,5	53,1	1895	120,0	72,5
1865	85,1	55,1	1896	123,3	72,7
1866	80,8	56,0	1897	120,5	72,9
1867	88,3	57,8	1898	136,7	78,0
1868	83,8	61,3	1899	143,3	79,3
1869	94,5	64,3	1900	122,0	77,8
1870	97,6	65,0	1901	137,3	79,1
1871	81,6	66,2	1902	140,1	80,3
1872	83,3	68,5	1903	140,5	80,8
1873	87,2	67,5	1904	137,3	80,0
1874	95,3	69,6	1905	143,6	80,9
1875	98,7	69,7	1906	143,6	82,8
1876	102,2	71,7	1907	150,5	86,1
1877	96,0	71,8	1908	142,8	90,0
1878	94,1	73,2	1909	150,0	90,5
1879	97,2	76,0	1910	141,6	89,0
1880	114,0	78,5	1911	169,2	91,1
1881	112,2	82,9	1912	174,9	97,1
1882	130,3	80,6	1913	189,8	100,0
1883	111,0	80,4			

Источник: Струмилин С.Г. Очерки экономической истории России и СССР. М., 1966. С. 82.

13. Валовая добыча угля в некоторых угольных бассейнах Российской империи, 1887-1913 гг. (тыс. тонн) (файлы *coal.sta, coal.xls*)

	Донецкий	Подмосковный	Уральский	Зап. Сибирь		Донецкий	Подмосковный	Уральский	Зап. Сибирь
1887	2055,5	288,1	163,3	14,4	1901	10889,9	255,1	528,1	236,1
1888	2240,2	276,2	208,9	18,0	1902	10727,7	211,3	547,6	208,0
1889	3110,1	306,3	262,7	17,0	1903	11583,3	217,8	491,1	249,3
1890	3001,7	233,7	249,4	19,3	1904	13080,9	215,3	516,5	308,8
1891	3139,5	180,5	245,5	21,1	1905	12863,4	214,1	492,4	439,6
1892	3571,9	179,7	252,8	21,3	1906	14241,7	320,2	697,5	494,4
1893	3928,6	179,3	260,5	18,3	1907	18184,8	348,0	699,1	516,2
1894	4846,2	194,0	278,6	22,1	1908	17907,6	328,4	749,2	596,6
1895	4886,5	166,4	288,8	23,5	1909	17735,8	253,3	812,7	555,8
1896	5106,8	157,8	365,2	24,9	1910	16707,1	227,8	780,5	516,8
1897	6793,5	202,3	356,2	11,6	1911	19800,9	177,1	697,8	534,9
1898	7565,9	161,5	385,8	22,9	1912	21369,4	226,0	941,8	707,8
1899	9218,9	224,2	362,1	72,0	1913	25288,2	300,4	1203,3	878,0
1900	11002,0	288,5	371,7	153,8					

Источник: Кафенгауз Л.Б. Эволюция промышленного производства России (последняя треть XIX в. – 30-е годы XX в.). М., 1994. С. 476.

14. Вывоз хлопка из Средней Азии, со станций ж.д., 1902-1908 гг. (в тыс. пудов) (файл *cotton.sta*)

Месяц	Год					
	1902/3	1903/4	1904/5	1905/6	1906/7	1907/8
Сентябрь	134	100	91	108	153	51
Октябрь	1084	1002	615	392	887	801
Ноябрь	959	1510	1201	542	1173	1145
Декабрь	911	1221	1141	633	1028	1662
Январь	900	763	1023	1036	1248	1131
Февраль	546	1156	1024	1432	1426	890
Март	609	1101	900	1144	1981	853
Апрель	278	747	606	1280	783	512
Май	106	415	282	1081	1053	440
Июнь	328	280	197	504	414	232
Июль	140	166	170	155	143	100
Август	108	94	100	76	60	85

Источник: Малаховский Н. Хлопок в мировой торговле // Вестник финансов, промышленности и торговли. 1909, №34. С. 289.

15. Сводные данные об аграрном развитии 50 губерний Европейской России на рубеже XIX-XX вв. (файл *typol.sta*)

		Показатели							
	Губерния	1	2	3	4	5	6	7	8
1	Архангельская	0,033	1,07	0,000	0,334	0,053	0,19	9,1	55,6
2	Астраханская	0,135	2,70	0,038	0,334	0,295	0,38	8,3	25,3
3	Бессарабская	0,053	1,13	0,224	0,287	0,127	0,99	37,8	44,3
4	Виленская	0,073	0,91	0,391	0,319	0,025	0,67	25,4	35,7
5	Витебская	0,056	1,25	0,319	0,463	0,047	0,60	22,0	36,5
6	Владимирская	0,024	1,63	0,108	0,293	0,110	0,57	22,4	40,2
7	Вологодская	0,017	2,89	0,007	0,424	0,021	0,45	20,4	46,1
8	Волынская	0,030	0,62	0,358	0,256	0,043	0,57	32,5	58,6
9	Воронежская	0,014	1,58	0,177	0,293	0,195	0,89	34,4	38,8
10	Вятская	0,018	2,62	0,024	0,111	0,022	0,99	41,8	42,7
11	Гродненская	0,055	1,14	0,406	0,119	0,031	0,59	24,3	38,0
12	Донская	0,040	4,38	0,071	0,372	0,087	1,55	45,8	33,0
13	Екатеринославская	0,043	1,35	0,213	0,313	0,244	1,26	47,6	40,9
14	Казанская	0,010	1,61	0,083	0,258	0,052	0,86	31,5	39,2
15	Калужская	0,023	1,13	0,159	0,275	0,109	0,50	23,5	37,7
16	Киевская	0,035	0,68	0,330	0,318	0,092	0,52	27,9	57,9
17	Ковенская	0,208	1,12	0,241	0,370	0,065	0,61	26,1	44,1
18	Костромская	0,029	1,65	0,134	0,356	0,092	0,64	27,8	42,7
19	Курляндская	0,298	1,79	0,384	0,334	0,128	0,76	38,7	57,7
20	Курская	0,021	1,14	0,217	0,296	0,209	0,87	37,6	43,5
21	Лифляндская	0,305	1,26	0,407	0,334	0,255	0,66	33,9	62,6
22	Минская	0,050	1,01	0,504	0,467	0,062	0,60	26,1	37,7
23	Могилевская	0,033	1,05	0,342	0,309	0,104	0,56	25,6	39,6
24	Московская	0,023	1,22	0,158	0,310	0,215	0,35	10,0	44,8
25	Нижегородская	0,024	1,36	0,151	0,316	0,102	0,65	26,0	40,9
26	Новгородская	0,032	2,25	0,140	0,394	0,109	0,48	21,0	43,1
27	Олонецкая	0,029	11,39	0,005	0,081	0,002	0,37	19,6	57,6
28	Оренбургская	0,045	7,37	0,037	0,595	0,035	0,94	29,5	34,2
29	Орловская	0,025	1,12	0,210	0,276	0,178	0,81	34,0	41,7
30	Пензенская	0,019	1,36	0,230	0,306	0,169	0,94	38,7	42,7
31	Пермская	0,028	2,96	0,235	0,324	0,047	0,64	31,4	51,1
32	Петербургская	0,068	1,49	0,267	0,446	0,149	0,42	7,7	46,3
33	Подольская	0,033	0,93	0,354	0,469	0,063	0,66	26,9	43,8
34	Полтавская	0,047	0,87	0,257	0,404	0,340	0,73	34,8	48,9
35	Псковская	0,027	1,39	0,168	0,425	0,163	0,54	21,4	41,5
36	Рязанская	0,020	1,15	0,186	0,270	0,196	0,75	34,5	45,1
37	Самарская	0,044	2,58	0,071	0,264	0,413	1,14	34,0	30,6
38	Саратовская	0,031	1,59	0,193	0,421	0,275	1,06	37,6	39,3
39	Симбирская	0,026	1,16	0,158	0,346	0,225	0,91	36,2	41,2
40	Смоленская	0,033	1,16	0,202	0,486	0,324	0,58	28,9	48,1
41	Таврическая	0,097	1,69	0,165	0,339	0,273	1,95	54,4	35,9
42	Тамбовская	0,020	1,15	0,195	0,292	0,196	0,88	42,7	50,9
43	Тверская	0,029	1,64	0,116	0,295	0,224	0,50	23,7	48,0
44	Тульская	0,024	1,05	0,295	0,250	0,196	0,90	43,1	46,0
45	Уфимская	0,027	2,99	0,130	0,401	0,067	0,71	29,5	42,6
46	Харьковская	0,021	1,25	0,180	0,443	0,083	0,90	30,9	40,0
47	Херсонская	0,052	1,17	0,202	0,338	0,576	1,62	42,0	35,7
48	Черниговская	0,041	1,11	0,207	0,337	0,155	0,69	25,6	34,3
49	Эстляндская	0,177	1,17	0,666	0,334	1,162	0,56	40,1	61,6
50	Ярославская	0,040	1,53	0,130	0,281	0,141	0,52	28,7	55,2

15. Продолжение

	Показатели										
	9	10	11	12	13	14	15	16	17	18	19
1	0,09	0,66	2,33	0,130	0,770	0,026	0,40	59	10	1,0	90
2	0,09	0,19	1,85	0,090	0,675	0,079	0,62	47	20	1,0	50
3	0,01	0,18	0,50	0,220	0,505	0,079	0,42	44	137	10,1	72
4	0,05	0,20	0,84	0,160	0,718	0,023	0,50	30	53	4,9	64
5	0,04	0,28	0,93	0,180	0,629	0,031	0,48	38	46	3,6	64
6	0,02	0,22	0,56	0,130	0,825	0,013	0,28	58	56	1,9	64
7	0,02	0,31	1,20	0,170	0,711	0,015	0,52	39	14	1,9	71
8	0,02	0,24	0,43	0,190	0,561	0,046	0,23	36	155	11,9	56
9	0,01	0,17	0,53	0,240	0,582	0,109	0,44	51	105	10,5	47
10	0,01	0,18	0,46	0,250	0,554	0,087	0,44	34	28	6,0	47
11	0,04	0,18	0,75	0,140	0,747	0,016	0,39	32	58	2,5	67
13	0,01	0,18	0,79	0,390	0,463	0,175	1,07	80	87	7,0	60
13	0,01	0,17	0,51	0,280	0,408	0,177	0,57	78	141	7,8	68
14	0,00	0,19	0,33	0,200	0,692	0,033	0,26	34	69	7,9	61
15	0,02	0,34	0,63	0,230	0,581	0,063	0,35	53	74	4,3	64
16	0,03	0,24	0,55	0,150	0,680	0,035	0,25	40	122	13,2	53
17	0,15	0,28	0,96	0,220	0,540	0,143	0,51	37	72	4,9	66
18	0,02	0,24	0,62	0,180	0,794	0,010	0,37	49	19	2,5	62
19	0,24	0,30	1,01	0,220	0,334	0,392	0,59	50	75	6,6	70
20	0,01	0,21	0,36	0,270	0,482	0,149	0,29	52	141	15,0	50
21	0,27	0,28	1,22	0,170	0,411	0,191	0,57	43	65	7,5	75
22	0,03	0,25	1,03	0,180	0,625	0,039	0,56	39	32	2,9	60
23	0,03	0,41	0,84	0,270	0,382	0,148	0,43	40	53	12,7	56
24	0,03	0,54	0,75	0,110	0,791	0,017	0,14	61	163	2,5	64
25	0,02	0,19	0,40	0,170	0,769	0,034	0,24	43	47	6,2	55
26	0,02	0,37	0,98	0,220	0,632	0,041	0,44	42	20	1,5	79
27	0,04	0,42	1,42	0,190	0,707	0,030	0,50	46	13	4,5	85
28	0,02	0,29	0,75	0,380	0,359	0,340	0,64	51	22	2,0	40
29	0,02	0,22	0,37	0,230	0,571	0,095	0,26	40	108	11	53
30	0,01	0,17	0,38	0,240	0,592	0,079	0,32	41	83	9,5	45
31	0,02	0,16	0,86	0,140	0,541	0,114	0,52	43	30	1,0	49
32	0,06	0,59	0,72	0,090	0,757	0,023	0,10	51	64	1,7	84
33	0,02	0,32	0,78	0,250	0,426	0,157	0,42	31	62	5,6	57
34	0,04	0,14	0,55	0,150	0,777	0,028	0,36	47	143	9,5	49
35	0,02	0,35	1,03	0,200	0,611	0,041	0,48	40	41	4,8	75
36	0,01	0,19	0,45	0,200	0,688	0,055	0,31	54	99	9,5	54
37	0,02	0,25	0,45	0,400	0,430	0,251	0,48	47	45	6,0	54
38	0,01	0,18	0,56	0,250	0,545	0,127	0,52	52	78	11,3	58
39	0,01	0,17	0,34	0,210	0,652	0,053	0,29	39	71	8,8	64
40	0,03	0,41	0,93	0,290	0,430	0,119	0,50	44	55	3,6	66
41	0,02	0,14	0,35	0,340	0,311	0,316	0,55	69	110	6,1	72
42	0,01	0,18	0,41	0,240	0,579	0,114	0,33	46	109	13,4	48
43	0,03	0,38	0,92	0,210	0,667	0,026	0,41	48	44	2,9	63
44	0,01	0,20	0,37	0,260	0,552	0,080	0,30	55	115	11,0	52
45	0,01	0,33	0,64	0,340	0,541	0,171	0,44	40	29	3,0	47
46	0,01	0,16	0,53	0,170	0,704	0,037	0,41	52	127	9,3	50
47	0,01	0,16	0,34	0,280	0,414	0,183	0,39	64	142	8,4	65
48	0,03	0,29	0,53	0,250	0,495	0,134	0,33	40	93	7,4	54
49	0,18	0,30	1,14	0,200	0,543	0,117	0,52	47	70	6,0	70
50	0,04	0,27	0,81	0,150	0,834	0,008	0,36	55	39	2,0	58

Источник: Ковальченко И.Д., Бородкин Л.И. Аграрная типология губерний Европейской России на рубеже XIX - XX веков (Опыт многомерного количественного анализа) // История СССР. 1979. N 1. С. 92-93.

Показатели:

1 - Наемные с.-х. рабочие по отношению к местным работникам;

2 - Земельный надел на душу (дес.);
3 - Доля дворянской земли в удобной земле;
4 - Отношение проданных частновладельческих земель к общей их площади;
5 - Отношение арендованной крестьянами земли к надельной земле;
6 - Посев (дес.) на душу населения;
7 - Сбор (пуд.) хлебов и картофеля на душу населения;
8 - Урожайность зерновых (пуд. с дес.);
9 - Наемные рабочие на дес. посева;
10 - Лошадей на дес. посева:
11 - Продуктивный скот на дес. посева;
12 - Лошадей на душу населения;
13 - В общем числе дворов доля безлошадных и однолошадных дворов;
14 - Доля дворов с 4 и более лошадьми;
15 - Продуктивный скот на душу населения;
16 - Поденная плата с.-х. рабочим в уборку урожая (коп.);
17 - Цена десятины земли (руб.);
18 - Арендная плата за дес. пашни (руб.);
19 - Осенние цены ржи (коп./пуд.).

16. Урожайность хлебов в России и других странах в 1913 г. (пудов с десятины) (файлы *harvest.sta, harvest*))

Страна	Пшеница	Рожь	Ячмень	Овес	Картофель
Россия	55	56	62	63	491
Австрия	89	92	107	94	602
Венгрия	88	82	92	91	470
Великобритания	149	0	127	117	1086
Бельгия	168	147	180	171	1338
Болгария	80	75	81	60	–
Германия	157	127	148	146	1057
Голландия	160	122	168	148	1176
Испания	52	61	64	45	–
Румыния	94	70	71	68	641
Сербия	72	58	65	46	–
Франция	89	71	92	86	571
Швейцария	153	123	130	150	1038
Швеция	161	94	139	123	–
Канада	94	61	108	102	750
США	68	68	85	70	408

Источник. Сборник статистико-экономических сведений по сельскому хозяйству России и иностранных государств. Год десятый. Пг., 1917. С.117,118.

17. Урожай хлебов в 64 губерниях Европейской России, 1890-1913 гг. (в тыс.пудов) (файл *harvest1.sta*)

Год	Пшеница	Рожь	Ячмень	Овес	Всего хлебов
1890	352762	1054383	226915	544585	2178645
1891	286967	791444	195516	434006	1707933
1892	538722	974317	279443	473075	2265557
1893	733225	1166718	449255	675884	3025082
1894	694648	1350865	364493	675565	3085571
1895	626017	1215224	327682	648700	2817623
1896	606512	1190025	324955	645949	2767441
1897	475590	969889	306308	527773	2279560
1898	678189	1107306	397806	556347	2739648
1899	654174	1365310	289870	805166	3114520
1900	657712	1401736	309364	720225	3089037
1901	667297	1145848	313366	527821	2654332
1902	931546	1387023	442102	786131	3546799
1903	916784	1364452	465873	645170	3392279
1904	1013991	1516567	451552	943790	3925900
1905	944349	1098993	450398	754714	3248454
1906	749440	990413	404452	561199	2705504
1907	727449	1200556	457375	728473	3113853
1908	812832	1176503	488430	739501	3217266
1909	1182249	1360317	621753	946121	4110440
1910	1162266	1308356	602828	856255	3929705
1911	742907	1151202	537271	702631	3134011
1912	1036532	1567754	606085	862482	4072853
1913	1392158	1507229	741054	979744	4620185

Источники: Брошниовский А.К. «Вероятные условия сбыта русских хлебов…» СПб., 1902; Материалы по статистике хлебной торговли. 1911 г. СПб., 1911; Сборник статистико-экономических сведений по сельскому хозяйству… Год девятый» Пг., 1916.

18. Факторы урожайности (погодный индекс, обрабатываемая площадь, мощность двигателей) в СССР в 1925-1940 гг. (файл *hunter.sta*)

Год	Погодный индекс	Обрабатываемая площадь (млн га)	Мощность двигателей (л.с.)			Оценка сбора зерновых (млн тонн) по данным:			
			лошади	трактора	всего	N.Jasny	D.Johnson& A.Kahan	официальным ЦСУ	J.Karcz
1925	87	87,3	24,6	0,026	24,7				
1926	112	93,7	26,4	0,073	26,5				
1927	89	94,7	28,4	0,203	28,6				
1928	101	92,2	28,9	0,262	29,2	73,3	73,3	73,3	73,1
1929	104	96,0	29,2	0,294	29,5	71,7	71,7	71,7	71,7
1930	115	101,8	25,7	0,428	26,1	83,5	83,5	83,5	77,2
1931	79	104,4	21,5	1,082	22,6	66,0	66,0	69,5	68,0
1932	95	99,7	18,1	2,056	20,2	66,4	63,0	69,9	67,1
1933	108	101,6	15,8	2,636	18,4	70,1	67,1	89,8	68,4
1934	94	104,7	14,9	3,866	18,8	72,2	67,3	89,4	67,6
1935	118	103,4	14,0	5,483	19,5	76,6	69,3	90,1	75,0
1936	86	102,6	13,7	8,092	21,8	63,6	60,0	83,0	56,1
1937	113	104,5	14,0	10,390	24,4	96,0	91,9	120,3	97,4
1938	95	102,4	14,4	11,410	25,8	75,9	70,7	94,9	74,5
1939	91	99,9	15,3	12,800	28,1				
1940	113	100,4	15,8	13,870	29,7				

Примечание: погодный индекс был построен С. Витскрофтом (Stefen G.Wheatscroft) на основе данных местных метеостанций (личная корреспонденция). Остальные данные получены на основе советских статистических данных и западных оценок. Каждый трактор 15 л.с. считался равным 20 лошадям, грузовик или комбайн – 10 лошадей.

Источник: H. Hunter and J. Szyrmer. Faulty Foundations. Soviet Economic Policies, 1928-1940. С. 107, 97.

19. Итоги выборов в Учредительное собрание по избирательным округам (число голосов)

(файл *uchred.sta*)

Округ	эсеры	энесы	меньш.	больш.	социал.	кадеты	нац.сп.	правые	прочие	всего
Алтайский	621377	6066	3585	45268	0	12108	8048	17292	0	713946
Архангельский	106570	0	7335	36522	0	12086	0	0	1160	163673
Астраханский	100482	942	2220	36023	0	13017	25023	0	16400	194107
Бессарабский	112886	1367	4179	28614	10797	19050	49018	6317	116467	348695
Витебский	150279	3599	12471	287101	32108	8134	41227	15117	10504	560538
Владимирский	196886	6917	13139	345306	0	38058	0	9193	1659	611158
Вологодский	348239	8340	3606	0	84358	27357	0	0	0	469900
Волынский	27575	0	16947	35612	570073	22397	113992	1746	15866	804208
Воронежский	875300	6116	8658	151517	11871	36488	0	7281	769	1098000
Вятский	612525	37621	19167	236952	0	48106	55585	9396	55476	1074828
Донской	478901	5049	17504	205497	5718	43345	0	13640	636966	1406620
Екатеринославский	231717	9496	39155	213163	565150	27551	72152	34665	0	1193049
Енисейский	233345	8000	4581	96138	0	12263	0	0	2452	356779
Забайкальский	104220	4260	3245	17260	0	7200	26155	218	15078	177636
Закавказский	117522	514	661934	93581	825672	25673	728206	0	0	2453102
Иркутский	127834	14935	6899	33576	0	9393	39248	3267	0	235152
Казанский	273978	0	4906	51936	382640	31728	99080	12322	2001	858591
Калужский	127313	601	6996	225378	1067	24125	0	4409	0	389889
КВЖД	5079	0	13138	10613	0	6322	0	0	0	35152
Киевский	19220	3072	32685	60693	1179234	28667	133766	48758	3624	1509719
Костромской	249838	0	19488	226905	0	41448	0	17901	0	555580
Курский	869498	8591	6043	120094	0	47221	0	8715	0	1060161
Кубано-Черноморск.	2828	0	794	9167	0	3226	98	0	5733	21846

19. Продолжение

Округ	эсеры	энесы	меньш.	больш.	социал.	кадеты	нац.сп.	правые	прочие	всего
Лифляндский	0	0	7046	97781	0	0	0	0	31253	136080
Минский	181673	0	16277	579087	14054	10724	101928	13505	0	917248
Могилевский	511998	0	21664	93060	6600	19316	62278	10136	0	725052
Москва	62260	2508	21597	366148	37651	263859	0	0	10740	764763
Московский	172229	6978	27928	368264	0	44478	0	8458	31536	659871
Нижегородский	314472	5186	7660	133970	19959	34724	402	67305	2216	585894
Новгородский	220665	10314	10196	203658	0	31484	0	8704	2301	486422
Олонецкий	0	0	0	0	127120	20278	0	0	2813	150211
Орловский	510628	0	16301	241785	3338	18345	0	12911	10096	813404
Оренбургский	112209	6550	9575	166121	0	24847	158663	0	226604	704569
Пензенский	518228	4336	4726	54731	0	25407	29821	0	0	636247
Пермский	665118	0	27502	268292	29012	111241	77861	83734	15129	1277889
Петроград	156936	19109	29820	424024	4377	246506	0	42308	18001	941081
Петроградский	119761	12048	6100	229698	16904	64859	15963	5661	841	471835
Подольский	11052	852	12487	32942	660432	9371	113588	0	7530	848254
Полтавский	198437	4391	5993	64460	761313	18105	33340	0	63217	1149256
Приамурский	61967	0	16772	43534	0	17799	3275	0	97556	240903
Псковский	295012	4059	4870	173631	0	25961	3859	3209	4703	515304
Рязанский	427364	5695	5039	272153	0	30734	0	9368	4216	754569
Самаркандский	4238	0	1586	0	0	0	87059	0	1913	94796
Самарский	690341	4369	6125	195132	51212	44507	192861	20180	3083	1207810
Саратовский	612094	10243	15152	261308	0	27226	103470	51774	6379	1087646
Семиреченский	0	0	0	0	120150	0	200639	0	28272	349061
Симбирский	424185	7953	4785	90388	0	18303	71931	11130	2116	630791
Смоленский	250134	2855	7901	361062	0	29274	1708	5300	0	658234
Ставропольский	291395	670	1836	17430	0	10898	0	3078	2609	327916
Степной	14930	0	2500	18901	7713	7886	135386	1469	5869	194654
Таврический	300100	4643	17449	31612	63286	38791	110248	7715	885	574729
Тамбовский	837497	7412	22424	240652	6222	47548	0	12494	889	1175138

19. Продолжение

Округ	эсеры	энесы	меньш.	больш.	социал.	кадеты	нац.сп.	правые	прочие	всего
Тверской	245997	3739	18752	398479	0	37453	0	7723	2974	715117
Терско-Дагестанский	11542	443	1461	29889	386	11330	1881	0	29924	86856
Тобольский	392061	50780	12061	0	0	13793	25830	0	0	494525
Томский	541153	15802	5769	51456	0	18618	0	0	2686	635484
Тульский	256069	1991	10940	237558	0	22782	0	0	9516	538856
Тургайский	63650	0	6758	0	0	0	211274	0	0	281682
Уральский	5076	0	0	0	0	0	278014	0	87535	370625
Уфимский	322166	11429	2614	48151	304844	15825	224817	11178	15372	956396
Ферганский	0	0	0	0	0	0	770284	0	0	770284
Харьковский	838873	11852	20529	114743	1802	61302	11587	24335	11448	1096471
Херсонский	368078	6177	17371	107975	65210	57699	117805	21609	143647	905571
Черниговский	105565	10089	10813	271174	486964	28864	28308	16715	15154	973646
Эстляндский	3200	64704	0	119863	0	0	68085	0	17022	272874
Якутский	1208	0	247	0	0	586	0	0	1541	3582
Ярославский	197465	5637	16803	176035	0	53730	0	4427	4421	458518
Балтийский флот	45016	0	0	66810	0	0	0	0	3997	115823
Черноморский флот	22251	0	1943	10771	12895	0	0	0	4769	52629
Северный фронт	249832	5868	10420	471828	88956	13687	0	0	0	840591
Западный фронт	180582	0	9700	653430	68455	16750	0	0	47083	976000
Юго-Зап. фронт	402930	1125	79630	300112	177354	13724	9465	0	23083	1007423
Румынский фронт	670047	4004	36485	173728	188760	20956	31624	0	0	1125603
Кавказский фронт	229705	20574	61783	0	10285	8640	0	0	0	330987
Всего	19110024	459873	1564360	10828742	7003942	2181171	4674851	673763	1885110	48381836

Примечание: эсеры – ПСР – партия социалистов-революционеров; меньшевики – РСДРП (м); большевики – РСДРП (б); социалисты – совокупность партий социалистической и социал-демократической ориентации (за исключением приведенных выше); кадеты – ПНС – партия народной свободы; правые – партии либеральной, консервативной и религиозной ориентации.
Источник. Протасов Л.Г. Итоги выборов в Учредительное собрание // Вопросы выборов и избирательного права. 1998. №1. С. 68-70.

20. Социально-экономические показатели и результаты голосования по выборам в Учредительное Собрание в 1917 г. по уездам Тамбовской губернии (файл *tambov.sta*)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Борисоглеб.	267	11,1	0,6	61,5	35,6	24,2	1,23	1,58	19,8	2,00	58	29,2	0,41	81,2	0,9	2,3	0,3	3,7	0,05	11,4	0,19
Елатомский	127	6,5	3,3	89,4	24,3	15,5	1,17	1,90	15,9	1,49	50	14,2	0,39	72,3	0,7	0,6	0,6	3,0	0,05	19,4	3,37
Кирсановск.	152	10,4	3,0	85,1	34,3	27,6	1,23	1,44	18,0	2,31	99	4,7	0,13	64,9	0,6	0,7	0,3	2,3	0,05	31,2	0,03
Козловский	183	11,9	2,0	79,8	40,9	28,9	1,17	1,45	22,0	2,15	65	14,9	0,21	64,5	1,1	2,5	0,5	5,4	0,08	25,9	0,09
Лебедянск.	127	13,1	1,4	77,8	36,9	22,3	1,23	1,63	18,7	2,00	18	34,9	0,34	78,2	2,2	0,4	1,2	3,5	0,05	14,4	0,05
Липецкий	248	13,6	0,8	77,2	38,7	23,8	1,01	1,44	25,4	1,67	46	26,1	0,56	48,8	1,0	3,3	0,8	6,5	0,06	39,4	0,07
Моршанск.	655	26,2	0,8	85,8	40,9	30,4	1,16	1,42	25,1	1,55	82	47,8	1,09	81,2	1,0	3,8	0,6	4,3	0,10	8,8	0,09
Спасский	207	14,0	2,3	9,2	35,8	21,5	0,99	1,69	11,2	1,59	14	45,4	0,38	89,0	0,4	0,8	0,4	2,3	0,41	6,6	0,06
Тамбовский	184	17,4	1,4	75,3	33,5	24,7	1,20	1,40	22,8	1,97	62	75,3	0,89	68,1	2,0	3,4	1,4	5,4	0,05	19,3	0,23
Темниковс..	99	14,8	0,6	92,9	24,9	14,7	1,08	2,01	11,6	1,34	25	61,7	0,91	78,9	0,3	0,6	0,3	4,8	0,04	10,8	4,23
Усманский	169	9,2	1,7	73,8	37,2	22,8	1,25	1,59	17,6	2,04	23	14,3	0,11	61,4	0,4	0,2	0,4	2,3	0,05	35,2	0,09
Шацкий	143	20,4	2,0	82,3	29,9	18,6	1,18	1,82	14,1	1,76	11	90,3	0,46	74,5	0,8	0,5	0,2	3,2	0,08	20,0	0,68

Примечание: Данные собраны проф. Л.Г. Протасовым

Список признаков:

1 - средний размер селения (чел.)
2 - процент селений с промыслами
3 - средний процент дворов без земли
4 - средний процент дворов без инвентаря
5 - средний процент дворов без рабочего скота
6 - средний процент дворов без продуктивного скота
7 - среднее число лошадей на двор
8 - среднее число коров на двор
9 - средний процент дворов без посева
10 - рожь (дес./двор)
11 - число фабрик
12 - среднее число рабочих на фабрике
13 - процент рабочих среди населения
14 - процент голосов, поданных за эсеров
15 - процент голосов, поданных за СЗС
16 - процент голосов, поданных за меньшевиков
17 - процент голосов, поданных за энесов
18 - процент голосов, поданных за кадетов
19 - процент голосов, поданных за крестьянский список
20 - процент голосов, поданных за большевиков
21 - процент голосов, поданных за мусульманскую партию

21. Социальные движения в городах Италии в XIV в.
Матрица экспертных оценок показателей (файл *bragina.sta*)

Восстание		Показатели				
№	Название	I	II	III	IV	V
1	Движение городских низов во Флоренции осенью 1342 г.	2	8	4	2	6
2	Восстание во Флоренции в июле 1343 г.	8	6	8	8	6
3	Выступление наем. рабочих во Флоренции 23 сент. 1343 г.	1	3	5	6	8
4	Волнения городских низов во Флоренции в октябре 1343 г.	2	3	3	4	0
5	Волнения городских низов во Флоренции в 1344 г.	3	2	3	2	1
6	Восстание во Флоренции в мае 1345 г.	4	5	4	4	0
7	Восстание в Сьене в 1346 г.	3	4	4	5	0
8	Выступление городских низов во Флоренции в авг. 1368 г.	1	2	3	3	0
9	Волнения наемных рабочих во Флоренции осенью 1368 г.	2	3	3	4	1
10	Восстание чомпи во Флоренции - июнь 1378 г.	5	5	8	7	6
11	Восстание чомпи во Флоренции - июль 1378 г.	7	8	8	10	10
12	Восстание чомпи во Флоренции - август 1378 г.	10	10	8	10	8
13	Выступление наем. рабочих во Флоренции в марте 1382 г.	4	4	3	6	0
14	Выступление город. низов во Флоренции в июле 1383 г.	4	5	3	1	0
15	Восстание горожан в Сьене в марте 1355 г.	3	4	7	1	0
16	Восстание в Сьене в сентябре 1368 г.	8	7	8	4	8
17	Восстание в Сьене в декабре 1368 г.	9	2	3	9	10
18	Восстание в Сьене в январе 1369 г.	8	2	3	5	8
19	Восстание наемных рабочих в Сьене в июле 1371 г.	7	4	4	9	10
20	Восстание в Перудже в мае 1371 г.	5	4	3	7	6

Источник: Брагина Л.М. О методике количественного анализа социальных движений в средние века // Математические методы и ЭВМ в историко-типологических исследованиях. М., 1989.

Показатели и их экспертные оценки [1]:

Показатель	*Оценка*
I. Цели и требования движения	
1. Экономические требования (цены, зарплата и т.д.)	1-2
2. Социально-политические требования, не включающие изменения политической власти (создание цехов и т.д.)	3-4
3. Политический компромисс (включение представителей восставших в правительство)	5-6
4. Социально-экономические требования и политический компромисс	6-7
5. Смена власти без социально-экономических перемен	7-8
6. Смена власти с социально-экономическими переменами	9-10
II. Степень организованности	
1. Стихийное движение	1-2
2. Элементы руководства, возникшие в ходе стихийного движения	3-4
3. Элементы предварительной подготовки движения (создание братства, заговор и т.д.)	4-5
4. Стихийное выдвижение требований	6-7
5. Выдвижение требований, подготовленных как программа	8-10
III. Социальный состав и массовость движения	
1. Отдельные группы наемных рабочих, плебс	1-3

[1] Необходимо отметить, что имеет значение не абсолютная величина баллов, а их соотношение "больше – меньше". Каждое из 20 изученных восстаний получило оценки по указанным пяти показателям.

2. Городские низы – наемные рабочие, подмастерья, мелкие ремесленники и торговцы	4-5
3. Средние слои горожан	5-6
4. Средние слои и нобили (гранды)	6-7
5. Городские низы и средние слои	8-10
IV. Методы борьбы	
1. Мирные демонстрации и переговоры с властью	1-2
2. Забастовка и другие формы невооруженной борьбы	3-4
3. Отдельные акты насилия	4-5
4. Сочетание мирных путей борьбы с актами насилия	6-7
5. Вооруженная борьба как основная форма движения	8-10
V. Результаты борьбы	
1. Поражение без дальнейших реформ	0
2. Поражение с последующими реформами (экономическими, социально-экономическими, социально-политическими)	1-3
3. Частичное удовлетворение экономических, социально-экономических и социально-политических требований	4-6
4. Смена власти без социально-экономических и социально-политических реформ	6-8
5. Смена власти с социально-экономическими и социально-политическими реформами	8-10

22. Распределение случаев выступлений по формам борьбы в "приговорном" и остальной части крестьянского движения в 1905-1907 гг. в Воронежской и Самарской губ. * (файл *bukhovez.sta*)

Формы движения	Количество случаев выступлений			
	Воронежская губерния		Самарская губерния	
	в "приг." движении	в ост. части	в "приг. движении	в ост. части
разгромы	51	232	22	114
захват земли	3	10	11	11
порубки	13	142	24	125
потравы, покосы	27	299	22	113
выступления на почве аренды	12	36	15	27
отказ от уплаты податей	3	6	10	15
выступления против земских начальников и полиции	9	13	21	25
отказ от выборов в землеустроительные комиссии	0	1	28	108
вооруженные столкновения с войсками и полицией	35	72	6	14
выст.против местных сел. властей и мелких служащих	9	39	1	11
установление самоуправления	3	0	1	0
подготовка воор. восстаний	2	0	4	0
попытки создания организаций	2	0	14	4,
полит. демонстрации и митинги	8	2	14	8
принятие политических приговоров и наказов	56	0	190	0
прочие	2	8	11	35
Итого:	235	860	394	610

*"Приговоры" – петиции, прошения, требования, наказы посылаемые в письменном виде крестьянами в Государственную Думу. Приведены данные об активности участия крестьян двух губерний в различных формах борьбы во время революции 1905-1907 гг. отдельно по селам, охваченным "приговорным" движением, и селам, отн. к остальной части крестьянского движения.

Источник: Буховец О.Г. Социальные конфликты и крестьянская ментальность в Российской Империи начала XX века: новые материалы, методы, результаты. М., 1996.

23. Криминальная статистика США (данные XIX – начала XX вв.)

(файл *criminal.sta*)

Город, штат, период	Число задержанных на 1000 жителей			Полицейских на 1000 жит.	
	Всего	Пьянство	Убийство	Всего	Патрульных
Lowell, Massachusetts (1862-1920)	44,36	33,58	0,01	1,46	1,16
Louisville, Kentucky (1870-1915)	39,31	25,12	0,17	1,61	1,25
Buffalo, New York (1872-1920)	80,44	48,29	0,02	2,22	1,61
New Haven, Connecticut (1863-1920)	57,07	33,86	0,01	1,62	1,07
Cleveland, Ohio (1872-1915)	52,19	27,14	0,06	0,98	0,77
Boston, Massachusetts (1860-1920)	60,00	34,34	0,06	2,16	1,69
Baltimore, Maryland (1864-1920)	60,56	29,04	0,06	1,93	1,38
Washington, D.C. (1862-1920)	93,08	34,42	0,06	2,05	1,88
Detroit, Michigan (1862-1920)	27,28	14,11	0,02	1,95	
Cincinnati, Ohio (1862-1916)	50,11	11,10	0,09	1,62	1,16
Chicago, Illinois (1860-1920)	41,47	20,58	0,02	1,90	1,43
St.Louis, Missouri (1861-1919)	40,82	18,38	0,05	2,18	1,48
San Francisco, California (1862-1917)	77,16	45,10	0,25	1,72	1,43
Richmond, Virginia (1872-1919)	57,14	24,47	0,09	1,18	0,98
Providence, Rhode Island (1863-1915)	53,76	36,17	0,03	1,92	1,34
Philadelphia, Pennsylvania (1887-1915)	50,73	32,02	0,05	2,15	1,57
New York, New York (1860-1920)	38,64	0,00	0,00	2,15	0,00
Newark, New Jersey (1870-1920)	29,39	15,14	0,07	1,55	1,02
New Orleans, Louisiana (1880-1915)	62,13	26,64	0,21	1,28	0,85
Milwaukee, Wisconsin (1868-1920)	16,86	10,17	0,02	1,13	0,96

Источник: Данные Национального архива правосудия США по криминальным делам, представленные в Internet.

24. Распространенность заразных болезней в России в 1912 г. (чел.)

(файл *deseases.sta*)

Болезнь	Число заболевших							
	Европ. Россия	Кавказ	Сибирь	Средняя Азия	Итого	Польша	Всего по империи	На 100 больных
Оспа	63801	3983	4783	1156	73723	7865	81588	0,40
Скарлатина	294288	15070	18758	5898	334014	16242	350256	1,72
Дифтерит	380993	16832	14770	8478	421073	10772	431845	7,02
Корь	348763	22867	25319	6550	404499	16308	419807	2,06
Коклюш	451283	19322	31599	10826	513030	17155	530185	2,60
Грипп	3073041	91252	183307	55612	3403212	37070	3440282	16,87
Тиф разн.	485065	26837	27275	12137	551314	18026	569340	2,79
Дизентерия	345575	35696	34059	9481	424811	11309	436120	2,14
Холера	3131	403	247	20	3792	1317	5109	0,03
Эпид. гастроэнтерит	252232	36030	36597	11748	336571	16837	353444	1,73
Заушница	209812	17217	13189	5818	246036	6100	252136	1,24
Рожа	166850	12145	6771	4782	190548	5512	196060	0,96
Ревматизм остр.	589651	80807	44750	19724	734932	17780	752712	3,69
Цинга	45627	5560	17187	34773	103147	657	103804	0,51

Источник: Россия. 1913 год. СПб., 1995. С. 323.

25. Грамотность населения в России (в тыс.) (файл *edu_1897.sta*)

Сословие	в уездах				в городах			
	неграмотных		грамотных		неграмотных		грамотных	
	об. пола	м. пола	об. пола	м. пола	об. пола	м. пола	об. пола	м. пола
А	358	160	444	228	177	77	872	422
Б	139	50	284	141	24	9	142	77
В	4043	1703	2109	1325	4255	1711	3603	2109
Г	77391	33970	15737	11525	3963	1888	2735	2035
Итого	81931	35883	18574	13218	8419	3665	7552	4642

Источник: Общий свод по империи результатов разработки данных первой всеобщей переписи населения, произведенной 28 января 1897 г. СПб., 1905. Табл.IX-а. С. 190-194.

Обозначения:
А – дворяне потомственные и личные; чиновники и их семьи;
Б – лица духовного звания и их семьи;
В – почетные граждане, купцы, мещане и другие городские сословия;
Г – лица сельского состояния (крестьяне, казаки, иностранные поселенцы)

26. Динамика уровня образования населения республик СССР за 1959-1979 гг. (файл *educat.sta*)

	На 1000 населения в возрасте 10 лет и старше приходится лиц с высшим и средним полным образованием		
	1959	1970	1979
РСФСР	146	233	390
Украинская ССР	140	258	408
Белорусская ССР	117	224	386
Узбекская ССР	129	237	422
Казахская ССР	126	221	386
Грузинская ССР	253	371	514
Азербайджанская ССР	169	262	419
Литовская ССР	96	185	327
Молдавская ССР	88	170	334
Латвийская ССР	171	271	404
Киргизская ССР	130	222	393
Таджикская ССР	106	193	345
Армянская ССР	208	315	486
Туркменская ССР	124	213	375
Эстонская ССР	172	265	392
СССР	143	212	397

Источник: Вестник статистики, 1984, №6. С. 43-46.

27. Распределение учащихся учебных заведений Мѣнистерства народного просвещения по вероисповеданиям и сословиям на 1 января 1914 года (ôайл *religsoc.sta*)

Учебные заведения	Общее чис-ло учеников к 1.01.1914	По вероисповеданиям						
		православ-ного	римско-католическ.	армянского	иных христ.	иудейского	магоме-танского	иных нехрист.
Мужские гимназии	147751	104876	15323	4312	8402	13463	1024	351
Мужские прогимназии	4359	2575	911	185	269	300	107	12
Реальные училища	80800	62044	4590	1377	6381	4960	1294	154
Средние техн. училища	8272	6829	482	77	379	465	32	8
Низшие техн. училища	2671	2585	30	3	23	12	18	-
Женские гимназии	311637	238252	11768	5412	14050	40805	722	628
Женские прогимназии	11940	9564	489	367	419	1030	34	37
Учительские институты	2249	2090	8	6	61	81	3	-
Учительские семинарии	12190	10773	513	61	440	1	340	62
Высшие начальные, гор., уездные училища	189511	149652	10686	3492	9880	12356	3099	346
Начальные училища, муж	3722007	3038288	297220	40479	202861	20821	112141	10197
жен	1837673	1431650	166078	11437	153998	42001	20333	2176
Ремесл. учеб. заведения	19108	16640	1173	184	350	494	227	40

27. Продолжение.

Учебные заведения	По сословиям								
	потомств. дворян	лич.дв.и чиновн.	из духов. звания	поч. граж. и купцов	мещан и цеховых	казаков	крестьян	иностран-цев	прочих
Мужские гимназии	12618	35659	8360	14832	39625	2766	29167	1494	3230
Мужские прогимназии	287	494	159	266	1278	342	1251	36	246
Реальные училища	4776	13465	2296	7715	23953	3835	22094	1188	1478
Средние техн. училища	259	627	153	440	2892	195	3471	60	175
Низшие техн. училища	24	93	37	77	630	161	1575	5	69
Женские гимназии	17005	51250	15114	29889	109787	5750	72220	2393	8229
Женские прогимназии	364	1223	484	632	4218	376	4203	25	415
Учительские институты	23	36	29	41	465	180	1451	1	23
Учительские семинарии	147	201	135	202	1714	928	8651	2	210
Высшие начальные, гор., уездные училища	3962	5477	1858	4473	65031	10322	95149	1046	2193
Ремесл. учеб. заведения	293	288	175	218	5805	388	11634	40	267

Источник. Всеподданнейший отчет Министра народного просвещения за 1913 г. Пг., 1916. Приложение. С. 50-53, 80, 102-105, 126-129, 150-151, 159-161, 174-179, 195-197, 215-221.

28. Распределение книг, вышедших в 1913 г., по видам изданий и содержанию (файл *books1.sta*)

Разделы по видам изданий и содержанию*	Количество названий	Тираж	Стоимость (руб.)
1. Учебные пособия	2761	22556928	9768063
2. Народные издания	2506	21625709	936573
3. Отчеты	2158	1391250	7165
4. Уставы	2120	1362321	4439
5. Беллетристика	1878	7653723	2707662
6. Земское и городское дело	1840	500230	14450
7. Религия	1764	5731975	2144695
8. Детские издания	1396	6549530	2034071
9. Сельское хозяйство	1329	3579146	835194
10. Музыка, пение, сцена	1300	1303419	430785
11. Медицина, ветеринария, гигиена	1007	1942176	1009614
12. Путеводители	979	2802855	1554730
13. Сборники	843	1698781	1060955
14. Библиография	773	3124983	138054
15. Драма	767	1025090	287437
16. Календари	676	13703665	2287886
17. Текущая жизнь	662	1087879	96862
18. Программы, правила	615	641545	92575
19. Педагогика	581	834531	425246
20. Природоведение	577	862161	671037
21. Технология	552	639063	593849
22. Поэзия	540	909876	343840
23, Военное и морское дело	524	1215826	688128
24. История	516	950875	1154284
25. Статистика	412	292810	31435
26. Юбилей Дома Романовых	402	4100670	854202
27. Право	385	997266	1493537
28. Биографии	358	597092	147176
29. География, этнография, путешествия	354	510792	372127
30. Социология	341	844841	247398
31. Собрания сочинений	290	1215887	1942544
32. Железнодорожное дело	269	245497	106655
33. Спорт	247	247725	120402

* Учтены разделы, насчитывающие более 240 названий.

Источник: Статистика произведений печати, вышедших в России в 1913 г. Пг., 1915. С. 5-7.

29. Распределение книг, вышедших в 1913 г., по языкам
(файл *books2.sta*)

Языки*	Количество названий	Тираж	Стоимость (руб.)
1. Польский	1882	4887037	1377730
2. Латышский	1133	37960718	583618
3. Немецкий	717	1541388	364632
4. Еврейский	574	1541015	282069
5. Древнееврейский	518	866520	725000
6. Эстонский	321	1102635	235507
7. Татарский	267	1052100	153635
8. Армянский	263	404407	144004
9. Грузинский	236	478338	129256
10. Малороссийское наречие	228	725585	151660
11. Литовский	194	668990	118346
12. Французский	148	307433	172778
13. Церковно-славянский	101	512620	311368
14. Азербайджанский	95	111540	62108
15. Чувашский	57	106900	9314
16. Киргизский	37	156300	21270

Источник: Статистика произведений печати, вышедших в России в 1913 г. Пг., 1915.

30. Число этнически смешанных семей в республиках СССР (1959-1979 гг.; на 1000 семей) (файл *mixture.sta*)

Республики СССР	Все население		Городское население		Сельское население	
	1959	1979	1959	1979	1959	1979
РСФСР	83	120	108	132	56	93
Украинская ССР	150	219	263	299	58	93
Белорусская ССР	110	201	237	295	56	92
Узбекская ССР	82	105	147	173	47	47
Казахская ССР	144	215	175	239	119	182
Грузинская ССР	90	104	164	155	37	148
Азербайджанская ССР	71	76	118	121	20	17
Литовская ССР	59	113	104	152	30	56
Молдавская ССР	135	210	269	360	94	113
Латвийская ССР	158	242	213	271	92	180
Киргизская ССР	123	155	181	216	92	107
Таджикская ССР	94	130	167	231	55	59
Армянская ССР	32	40	50	49	14	22
Туркменская ССР	85	123	149	199	25	33
Эстонская ССР	100	158	142	186	51	90

Источник: Население СССР. С. 99.

31. Средний размер семьи в республиках СССР (1959-1979 гг., чел. *) (файл *family.sta*)

Республики СССР	1959	1970	1979
РСФСР	3,6	3,5	3,3
Украинская ССР	3,5	3,4	3,3
Белорусская ССР	3,7	3,6	3,3
Узбекская ССР	4,6	5,3	5,5
Казахская ССР	4,1	4,3	4,1
Грузинская ССР	4,0	4,1	4,0
Азербайджанская ССР	4,5	5,1	5,1
Литовская ССР	3,6	3,4	3,3
Молдавская ССР	3,8	3,8	3,4
Латвийская ССР	3,2	3,2	3,1
Киргизская ССР	4,2	4,6	4,6
Таджикская ССР	4,7	5,4	5,7
Армянская ССР	4,8	5,0	4,7
Туркменская ССР	4,5	5,2	5,5
Эстонская ССР	3,1	3,1	3,1

* *Рассчитано по*: Итоги Всесоюзной переписи населения 1959 г. (сводный том). С. 242-243; Итоги Всесоюзной переписи населения 1970 г.. Т. VII. С. 206-207; Численность и состав населения СССР. С. 220-221.

32. Динамика естественного прироста населения республик СССР (на 1000 человек населения) * (файл *populat.sta*)

	1960	1965	1970	1979	1981
СССР	17,8	11,1	9,2	8,1	8,3
РСФСР	15,8	8,1	5,9	5,0	5,1
Украинская ССР	13,6	7,7	6,4	3,6	3,3
Белорусская ССР	17,8	11,1	8,6	6,3	6,7
Узбекская ССР	33,8	28,8	28,1	27,4	27,7
Казахская ССР	30,6	21,0	17,4	16,3	16,3
Грузинская ССР	18,2	14,2	11,9	9,6	9,6
Азербайджанская ССР	35,9	30,2	22,5	18,1	19,4
Литовская ССР	14,7	10,2	8,7	5,0	4,8
Молдавская ССР	22,9	14,2	12,0	9,7	10,2
Латвийская ССР	6,7	3,8	3,3	1,0	1,4
Киргизская ССР	30,8	24,9	23,1	21,8	22,8
Таджикская ССР	28,4	30,2	28,4	30,1	30,5
Армянская ССР	33,3	22,9	17,0	17,3	18,1
Туркменская ССР	35,9	30,2	28,6	27,3	25,8
Эстонская ССР	6,1	4,1	4,7	2,6	3,1

* *Составлено по*: Народное хозяйство СССР в 1979 г. С. 30; Народное хозяйство СССР. 1922-1982. С. 29.

33. Распределение новобранцев русской армии, призванных в 1911 году, по росту (файл *novobr.sta*)

Рост (см) *	Количество новобранцев
153	3994
155,5	22736
160	66040
164,5	111752
169	115530
173,5	73376
178	29116
182,5	7303
191,5	164
196	12
более 196	4
не подвергались измерениям	155
Итого	431436

* В отчете Военного Министерства фиксировалось изменение роста на 1 вершок; в таблице 1 вершок округленно считается равным 4,5 см.

Источник: Военно-статистический ежегодник армии за 1912 г. СПб., 1914. С. 144-145.

34. Численность населения США в 1902-1914 гг. (тыс. человек)

(файл us_popul.sta)

Год	Численность населения	Год	Численность населения
1902	79231	1909	90557
1903	80849	1910	92175
1904	82467	1911	93793
1905	84085	1912	95411
1906	85703	1913	97028
1907	87321	1914	98646
1908	88939		

Источник: Statistical Abstract of the United States. 1914. Washington, 1915. P. 628.

35. Численность населения России (млн. человек)

(файл *rus_pop.sta*)

Год	Численность населения
1893	120,1
1894	121,6
1895	123,2
1896	124,8
1897	126,4
1898	128,1
1899	129,9
1900	131,7

Источник: Сборник сведений по статистике внешней торговли России / Под ред. В.И. Покровского. СПб., 1902. Т. 1. С. XXXIV.

"БОЛЬШИЕ ТАБЛИЦЫ",
представленные в виде файлов электронного архива
Лаборатории исторической информатики кафедры источниковедения
исторического факультета МГУ

I. Промышленные переписи 1900 и 1908 гг. по Закавказью (1060 предприятий, 12 показателей (файл *industry.sta*).
Источник: Список фабрик и заводов Европейской России. СПб., 1903 и 1912; Статистические сведения о фабриках и заводах по производствам, не обложенным акцизом, за 1903 г. СПб., 1903 и 1908.

II. Макроэкономические показатели по 37 странам мира с 1950 по 1992 гг. (37 стран, 43 года, 30 показателей) (файлы *wor_tabl.sta* и *wor_tabl.xls*).
Источник: Данные International Comparison Programme.

III. Биографические сведения о депутатах I Государственной Думы (431 депутат, 9 показателей) (файл *duma.sta*)
Источник: Григорьева Ю.Г. Дисс. на соискание уч. степени канд. ист. наук. М., 1995.

IV. Биографические сведения по высшему командному составу СА периода II Мировой войны (296 чел., 27 параметров) (файл *general.sta*).
Источник: Юмашева Ю.Ю. Дисс. на соискание уч. степени канд. ист. наук. М., 1994.

V. Продвижения по службе рабочих "Т-ва бр.Нобель" в Баку в к. XIX – н. XX вв. (147 рабочих, 3 показателя) (файл *baku2.sta*).
Источник: Аханчи П. Дисс. на соискание уч. степ. канд. ист. наук. М., 1993.

VI. Демографические данные по странам мира за 1994 г. (216 стран, 18 параметров) (файлы *demo_wor.sta* и *mapstats.xls*).
Источник: Данные ООН.

VII. Демографические данные по странам Европы с 1975 по 1993 гг. (27 стран, 67 параметров) (файлы *demo_eur.sta* и *mapstats.xls*).
Источник: Данные Eurostat.

VIII. Описание погребений раннесарматского времени (1116 погребений, 54 параметра) (файлы *sarmat.sta* и *sarmat.xls*).
Источник: Статистическая обработка погребальных памятников Азиатской Сарматии. Вып. II. Раннесарматская эпоха. М., 1997.