Академия управления МВД России

Б. А. Торопов, Э. Ф. Болтачев, В. В. Баранов

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ СОЦИАЛЬНЫХ СИСТЕМ

Учебное пособие

Москва • 2020

УДК 52-17:916.3
ББК 22.1в6
Т61

Одобрено редакционно-издательским советом Академии управления МВД России

Рецензенты: *Н.М. Дубинина*, начальник кафедры информатики и математики Московского университета МВД России им. В.Я. Кикотя, кандидат юридических наук, доцент; *Р.А. Одинцов,* начальник УНК УМВД России по Оренбургской области.

Торопов Б. А., Болтачев Э. Ф., Баранов В. В.

Т61 Математические методы исследования социальных систем: учебное пособие / Торопов Б. А. и др. – М.: Академия управления МВД России, 2020. – 80 с.
ISBN 978-5-907-187-30-6

Учебное пособие «Математические методы исследования социальных систем» предназначено для методического обеспечения одноименной учебной дисциплины, преподаваемой на факультете подготовки научных и научно-педагогических кадров Академии управления МВД России. При этом изложенный материал подходит для широкого круга обучающихся с различным уровнем начальных знаний в области математики и статистики. Типология рассматриваемых исследовательских задач и примеров их решения такова, что они могут быть легко адаптированы к различным направлениям научного поиска в области социальных, экономических, правовых систем, причем как в системе МВД России, так и за ее пределами.

УДК 52-17:916.3
ББК 22.1в6

ISBN 978-5-907-187-30-6

Оглавление

Введение

В любой своей деятельности человек постоянно оперирует моделями, причем чаще всего непроизвольно – не отдавая себе отчета. Действительно, рассматривая какой-либо объект реального мира, мы выделяем его главные свойства и не замечаем или игнорируем многие другие, ведь любой объект обладает практически бесконечным множеством свойств и характеристик. Восприняв эти главные, с нашей точки зрения, свойства, впоследствии мысленно обращаясь к объекту, мы представляем его не во всем многообразии свойств, а именно как совокупность тех, которые отложились в памяти. Такой упрощенный образ объекта и есть модель.

С древних времен, познавая мир, человек фиксировал полученные знания в виде моделей. Наскальный рисунок с изображением охоты – модель реального процесса, необходимого для выживания общины людей. Такая модель содержит общие сведения о внешнем виде животных, охотников и орудиях охоты, а ее назначение, помимо эстетического, – передать знания о важном виде деятельности подрастающему поколению. Любое изображение в виде графики, текста, словесного описания, характеризующее какое-либо явление, процесс, систему, существующие или могущие существовать в действительности, также есть модель, неизбежно упрощающая описываемый объект, выделяя его главные свойства и опуская второстепенные.

В научных исследованиях моделирование применялось еще в глубокой древности и постепенно захватывало все новые области научных знаний: географию, инженерию и архитектуру, астрономию, физику, химию, биологию и медицину. Наконец и общественные науки, сугубо гуманитарные по своей сути и часто оперирующие абсолютно абстрактными конструкциями, взяли моделирование на вооружение.

Большие успехи и признание практически во всех отраслях современной науки принес методу моделирования XX в. Однако методология моделирования долгое время формировалась независимо друг от друга отдельными отраслями научного знания, вкладывающими в ее содержание свой особый смысл. Отсутствовала единая система понятий, единая терминология. Но постепенно стала осознаваться роль моделирования как универсального метода научного познания.

Особое место в современной науке заняло математическое моделирование – способ формального описания зачастую плохо формализуемых явлений, процессов и систем на основе анализа

статистических данных. Математическое моделирование позволяет устанавливать закономерности в развитии систем, выявлять зависимость одних характеристик и показателей от других, прогнозировать будущее состояние исследуемых явлений, предлагать пути достижения желаемых состояний изучаемой системы.

Сегодня в мировой науке ни одно исследование в области социологии, права, экономики не мыслимо без построения моделей исследуемых явлений и процессов, без опоры на твердую почву статистических выводов. Педагогика и психология успешно применяют методы моделирования и анализа статистических данных при выдвижении и подтверждении научных гипотез. Социологи применяют математический аппарат и специализированные программные средства для изучения результатов опросов. Сугубо гумманитарные отрасли науки, такие как, например, филология, взяли на вооружение контент-анализ текстов – не что иное, как статистический анализ количественных данных, характеризующих текстовые массивы.

Настоящее пособие в первую очередь предназначено для нового поколения исследователей, которые работают над диссертациями в образовательных организациях системы МВД России, для тех, кто готов на новом уровне аргументировать свои выводы.

1. Моделирование как метод исследования социально-правовых явлений и процессов

Моделирование – это неотъемлемая часть любой человеческой деятельности. Будучи не в состоянии во всей полноте воспринимать окружающий мир, мы неизбежно и регулярно упрощаем в своем мыслительном процессе интересующие нас явления, события и объекты, а значит моделируем, отбрасывая второстепенные детали и сосредоточиваясь на важных. Научная деятельность здесь не только не является исключением, но, напротив, оперирует моделями в любых исследованиях. Модели применяются учеными либо в процессе экспериментальных исследований и теоретических построений, либо выступают результатом работы, либо верно то и другое.

Термин «моделирование» в науку первоначально был введен для исследования проблем, которые не удавалось сразу решить теоретическим или экспериментальным методом.

Реализации идеи моделирования способствовало развитие теории подобия изучающей условия подобия физических явлений, систем и процессов, и опирающейся на учение о размерности физических величин. Сначала рассматривался ряд видов подобия: геометрическое (подобие геометрических фигур), матричное (подобие матриц при задании их матрицами линейного преобразования), механическое (характеризующее однотипные механические системы). В последующем были введены термины физического (обобщающего механическое, тепловое и т. п. виды подобия) и его разновидностей – математического, кинематического, динамического, физического и химического подобий.

Основой теории подобия является установление критериев подобия различных явлений и изучение с помощью этих критериев свойств самих явлений. Сходства в сходственные моменты времени в сходственных точках пространства значений переменных величин, параметров, характеризующих состояние одной системы, пропорциональны соответствующим величинам другой системы. Коэффициент пропорциональности для каждой из величин называется коэффициентом подобия.

Основная идея теории подобия состоит в том, что из обширного класса однородных с физической точки зрения процессов, описываемых одной и той же системой дифференциальных уравнений, выбирают более узкую группу таких процессов, в пределах которой возможно распространение результатов единичных экспериментов. Процессы этой группы называются подобными между собой. Сле-

довательно, экспериментально исследуя один из таких процессов, можно экстраполировать полученные результаты и на другие процессы, входящие в группу.

Принципы теории подобия полезно использовать в качестве основы теории моделирования и в настоящее время. При этом способы установления сходства параметров изучаемых явлений и процессов при использовании различных методов моделирования различны.

Например, при применении теоретико-множественных представлений вводятся понятия изоморфизма и гомоморфизма, при использовании логико-лингвистических представлений – сходство предикатов и т. п.

Выходя на более высокий уровень обобщения, моделирование можно определить как замещение одного объекта (оригинала) другим (моделью), фиксацию и изучение свойств модели. Замещение производится с целью упрощения, удешевления, ускорения изучения свойств оригинала. При этом замещение правомерно, если интересующие исследователя характеристики оригинала и модели определяются однотипными подмножествами параметров, имеющих определенные свойства, количественной мерой которых служит множество характеристик, и связаны определенными зависимостями с этими параметрами.

Термин «модель» используется в разных смыслах: экземпляр, вариант какого-либо изделия; макет, повторяющий какие-то особенности определенного объекта; наглядные (уменьшенные, увеличенные или в натуральную величину) копии разных объектов – конструкций машин, зданий, сооружений, кристаллов, атомов и молекул и т. п.; модели одежды, фотомодели (девушки и юноши), т. е. то, что служит образцом для художественного воспроизведения, примером для подражания или сравнения и т.д.

Модель также может быть представлена как «один из важнейших инструментов научного познания, условный образ объекта исследования (или управления)», и поясняется, что «модель конструируется субъектом исследования («наблюдателем», по Эшби) так, чтобы отобразить характеристики объекта (свойства, взаимосвязи, структурные и функциональные параметры и т. п.), существенные для цели исследования. Поэтому вопрос о качестве такого отображения – адекватность модели объекту – правомерно решать лишь относительно определенной цели». При этом подчеркивается, что наиболее строгое и общее определение модели должно опираться на понятия гомоморфизма и изоморфизма.

Познание любого явления и процесса сводится, по существу, к созданию ее модели. Для выбора моделей разрабатывают их классификации.

Первоначально все модели делили на две группы – физические (вещественные, реальные) и математические (абстрактные, мыслимые). Затем – в соответствии с видами подобия были введены термины физических моделей кинематического и динамического; математического, химического и физико-химического подобия.

В последующем физические модели стали иногда делить на натуральные (макеты, опытные образцы); квазинатуральные (совокупность натуральных и математических моделей); масштабные (модели той же физической природы, что и оригинал, но отличающиеся от него масштабами; методологической основой таких моделей является теория подобия); аналоговые (модели, имеющие физическую природу, отличающуюся от оригинала, но сходные с оригиналом процессы функционирования). Математические модели классифицировали различными способами, но при этом интерпретируют неодинаково.

Выбор типа модели зависит от целей моделирования, а также от объема и характера исходной информации о рассматриваемом объекте и возможностей исследователя.

Моделирование всегда предполагает принятие допущений той или иной степени важности. При этом должен удовлетворяться ряд требований к моделям: адекватности, достаточной точности, целесообразности, экономичности и т. п., определение состава и уточнение формы реализации которых зависят от характера задачи, вида моделей и условий моделирования.

Необходимость в методах моделирования возникает в различных ситуациях для исследования, анализа, прогнозирования. Особое значение моделирование имеет для исследования социально-правовых явлений и процессов, где проведение натуральных экспериментов нереализуемо в принципе.

Традиционный подход, применяющийся в математических исследованиях: определить элементы (переменные, константы) и связать их соответствующим соотношением (формулой, уравнением, системой уравнений), отображающим принцип взаимодействия элементов.

Когда задачи усложнились и такое соотношение не удавалось сразу получить, то предлагалось формировать «пространство состояний» элементов и вводить «меры близости» между элементами этого пространства. Однако первые же попытки применить такой

подход показали, что «перечислить» элементы явлений и процессов практически невозможно.

Учитывая трудности «перечисления» систем, предлагались различные подходы к их исследованию и проектированию.

Применение философских категорий – индуктивный и дедуктивный подходы, анализ и синтез – позволяет определить основные принципы исследования. Однако эти категории могут трактоваться и реализовываться по-разному.

Рассматривая социально-правовые явления и процессы как системную целостность, можно рассмотреть подходы, применимые и к исследованию систем.

С самого начала возникновения системных теорий предлагались подходы, в большей мере ориентированные на прикладные задачи. Приведем основные из них:

– в начальный период становления теории систем развивался бихевиористский подход (behavior – поведение), основанный на исследовании поведения (т. е. функционирования) систем; однако этот подход весьма трудоемок и не всегда реализуем;

– американский ученый М. Месарович предложил подходы, которые, соответственно, назвал целенаправленным и терминальным (от терм – элементарная частица, интересующая исследователя);

– польский ученый Р. Куликовски предложил называть аналогичные подходы декомпозицией и композицией системы;

– швейцарский астроном Ф. Цвикки предложил и развил морфологический подход, который помогает искать полезные объединения элементов путем их комбинаций;

– американская корпорация RAND предложила подход к созданию сложных программ и проектов, названный «дерево целей»;

– в практике проектирования сложных технических комплексов возникли термины «язык моделирования», «язык автоматизации проектирования», применяющиеся для отображения взаимосвязей между компонентами проекта; при разработке языков моделирования применяют математическую логику и математическую лингвистику, в которой есть удобный термин для описания структуры языка – «тезаурус», и этот подход называют иногда лингвистическим или тезаурусным;

– при исследовании и формировании структур были предложены следующие подходы: путем поиска связей между элементами или, напротив, путем устранения лишних связей.

С учетом рассмотренных подходов на основе обобщения предшествующего опыта сформировалось два основных подхода к ото-

бражению систем, первоначально предложенных для формирования структур целей:

а) «сверху» – методы структуризации или декомпозиции, целевой или целенаправленный подход;

б) «снизу» – подход, который называют морфологическим (в широком смысле), лингвистическим, тезаурусным, терминальным, методом «языка» системы. С помощью этого подхода определяется «пространство состояний» системы и реализуется поиск взаимосвязей (мер близости) между элементами.

Подход «снизу» можно реализовать, применяя не только комбинаторные приемы (морфологический и т. п.), но и бихевиористский подход, вариант которого при автоматизации моделирования поведения объектов в настоящее время иногда называют процессным, статистические методы, лежащие в основе бизнес-аналитики, методы представления и извлечения знаний, основанные на применении математической логики и математической лингвистики.

Подходы «сверху» и «снизу» называют также аксиологическим и каузальным соответственно.

Аксиологическое представление системы – отображение системы в терминах целей и целевых функционалов. Этот термин используют в тех случаях, когда необходимо выбрать подход к отображению системы на начальном этапе моделирования и противопоставить это отображение описанию системы в терминах «перечисления» элементов системы и их непосредственного влияния друг на друга, т. е. каузального представления.

Каузальное представление системы – описание системы в терминах влияния одних переменных на другие, без употребления понятий цели и средств достижения целей. Этот термин происходит от понятия «cause» – причина, т. е. подразумевает причинно-следственные отношения. Применяют каузальное представление в случае предварительного описания системы, когда цель сразу не может быть сформулирована и для отображения системы или проблемной ситуации не может быть применено аксиологическое представление.

В 1970–1980-е гг. при проектировании организационных структур были предложены три подхода к решению этой проблемы.

1. Нормативно-функциональный подход направлен на унификацию организационных форм управления в рамках отрасли. Разработка типовых организационных структур явилась первым шагом на пути внедрения принципов их научно обоснованного построения. Однако ориентация на типовую номенклатуру функций управления

и структурных управленческих подразделений не позволяет учесть особенностей конкретных предприятий и условий их деятельности.

2. Функционально-технологический подход основан на рационализации потоков информации и технологии ее обработки, на формировании и анализе организационно-технологических процедур подготовки и реализации управленческих решений. Этот подход обеспечивает возможность достаточно полно учесть особенности конкретного предприятия (организации), отличается гибкостью и универсальностью. Вместе с тем он характеризуется высокой трудоемкостью, использованием стабильной номенклатуры сложившихся функций управления, подчинением оргструктуры схеме документооборота.

3. Системно-целевой подход заключается в построении структуры целей, определении на ее основе функций управления и их организационном оформлении. Преимущества этого подхода заключаются в возможности учитывать особенности объекта управления и условия его деятельности, изменять и расширять состав функций, проектировать разнообразные организационно-правовые формы предприятий. Трудности в использовании подхода связаны с проблемой перехода от совокупности целей и функций к составу и подчиненности структурных звеньев, обеспечивающих их реализацию.

Обобщающий подход «сверху», называемый целевым, целенаправленным, системно-целевым, основан на структуризации или декомпозиции системы в пространстве. Этот подход позволяет расчленить исходную большую неопределенность на более обозримые и выбрать методы их анализа и проектирования, сохраняя целостность представления об исследуемой системе или решаемой проблеме на основе иерархической структуры (древовидной, стратифицированной).

Подход «снизу», основанный на анализе пространства состояний, поиске «мер близости» между компонентами с помощью различных, в том числе статистических, методов, морфологического моделирования, отличается большой трудоемкостью. В настоящее время для анализа пространства состояний разработаны методы представления и извлечения знаний, основанные на применении статистических методов, математической логики и математической лингвистики.

В настоящее время для моделирования систем широкое применение нашел подход, кратко называемый процессным. Этот подход, который можно считать развитием функционально-технологического подхода, основан на структуризации во времени, на представлении процессов в форме графов.

Применение функционально-технологического подхода долгое время было практически нереализуемым из-за большой трудоемкости, отсутствия правил и средств автоматизации формирования графов, отображающих процессы в системах. В 1990-е гг. была разработана методология SADT (Structured Analysis and Design – структурный анализ и проектирование; предложена Дугласом Россом), представляющая собой совокупность методов, правил и процедур, предназначенных для построения функциональной модели объекта какой-либо предметной области. На ее основе разработаны и стали широко применяться функционально-ориентированные и объектно-ориентированные CASE – и RAD-технологии. Компьютерная реализация методологии SADT получила название IDEF (Icam Definition). Основными структурными моделями являются модели процессов IDEF0 и IDEF3, модель данных IDEF1X. Созданы стандарты IDEF и DFD, ориентированные на анализ процессов (в том числе бизнес-процессов). Для реализации моделей применяются автоматизированные средства – BPWin, ARIS, язык UML (Unified Modeling Language – унифицированный язык моделирования). Популярность CASE-методологии и RAD-технологий базируется на разработке принципов и автоматизации формирования процессов, на развитии методов их формирования (на основе анализа «жизненного цикла» производства, обслуживания или других процессов, причинно-следственных связей и т. п.), что и обеспечило развитие процессного подхода, преимущества которого заключаются в возможности учитывать особенности конкретного объекта и условий его деятельности.

Постановка любой задачи заключается в том, чтобы перевести ее словесное, вербальное описание в формальное.

Если полученная формальная модель (математическая зависимость между величинами в виде формулы, уравнения, системы уравнений) опирается на фундаментальный закон или подтверждается экспериментом, то этим доказывается ее адекватность отображаемой ситуации, и модель рекомендуется для решения задач соответствующего класса.

По мере усложнения задач получение модели и доказательство ее адекватности усложняется. Вначале эксперимент становится дорогим и опасным (например, при создании сложных технических комплексов, при реализации космических программ и т. д.), а применительно к социальным и экономическим объектам эксперимент становится практически нереализуемым, задача переходит в класс проблем принятия решений, и постановка задачи, формирование модели, т. е. перевод вербального описания в формальное, становит-

ся важной составной частью процесса принятия решения. Причем эту составную часть не всегда можно выделить как отдельный этап, завершив который, можно обращаться с полученной формальной моделью так же, как с обычным математическим описанием, строгим и абсолютно справедливым. Большинство реальных ситуаций управления социально-экономическими системами необходимо отображать классом самоорганизующихся систем, модели которых должны постоянно корректироваться и развиваться.

При этом возможно изменение не только подхода, но и метода моделирования, что часто является средством развития представления ЛПР о моделируемой ситуации.

Иными словами, перевод вербального описания в формальное, осмысление, интерпретация модели и получаемых результатов становятся неотъемлемой частью практически каждого этапа моделирования сложной развивающейся системы.

Для решения проблемы перевода вербального описания в формальное в различных областях деятельности стали развиваться специальные приемы и методы. Так, возникли методы типа «мозговой атаки», «сценариев», «дерева целей» и т. п.

В свою очередь развитие математики шло по пути расширения средств постановки и решения трудноформализуемых задач. Наряду с детерминированными, аналитическими методами классической математики возникла теория вероятностей и математическая статистика (как средство доказательства адекватности модели на основе представительной выборки и понятия вероятности использования результатов моделирования). Для задач с большей степенью неопределенности инженеры стали привлекать теорию множеств, математическую логику, математическую лингвистику, теорию графов, что во многом стимулировало развитие этих направлений. Иными словами, математика стала постепенно накапливать средства работы с неопределенностью, со смыслом, который классическая математика исключала из объектов своего рассмотрения.

Таким образом, к XX в. сложилось две формы культуры – естественно-научная и гуманитарная, различающиеся методами познания. Гуманитарное познание формирует образ, целостность, а формальное мышление обеспечивает отображение элементов и законов их взаимодействия. Гуманитарное знание связано с определением смысла, назначения, целесообразности (телеология), цели. Вершиной гуманитарного знания традиционно считается философия. Формальное – традиционно базируется на математике.

Формальные методы не позволяют выявить содержание исследуемых процессов, понять их целостность, хотя и могут помочь

ускорить обработку имеющейся информации, активизировать интуицию и опыт специалистов, в том числе с гуманитарным мышлением, для выявления новой информации, отобразить законы взаимодействия компонентов, полученные эмпирически.

Постепенно между неформальным, образным мышлением человека и формальными моделями классической математики сложился как бы «спектр» методов, которые помогают получать и уточнять (формализовать) вербальное описание проблемной ситуации, с одной стороны, и интерпретировать формальные модели, связывать их с реальной действительностью – с другой.

Развитие методов моделирования, разумеется, шло не так последовательно. Методы возникали и развивались параллельно.

Существует еще одна особенность, связанная с развитием математики, где отмечается возникновение новых областей, математические теории отмирают или вливаются в другие устаревающие разделы. Исследованием структуры математики занимаются многие ученые.

Несмотря на то, что в практике моделирования широко используются теория множеств, математическая логика, математическая лингвистика и другие направления современной математики, долгое время многие ученые-математики были не склонны включать в число математических некоторые из этих направлений. Благодаря работам французских ученых теорию множеств и математическую логику стали признавать разделами математики, а математическую лингвистику и семиотику есть основания не относить к математике. Поэтому, чтобы не обсуждать различные точки зрения (которые постепенно изменяются, развиваются), вместо термина «математические методы» удобнее применять термин «методы формализованного представления систем».

В большинстве первоначально применявшихся при исследовании систем классификаций выделяли детерминированные и вероятностные (статистические) методы или классы моделей, которые сформировались в конце прошлого столетия. Затем появились классификации, в которых в самостоятельные классы выделились теоретико-множественные представления, графы, математическая логика и некоторые новые разделы математики. Так, в классификации современного математического аппарата инженера выделяются: множества, матрицы, графы, логика, вероятности.

Поэтому, методы моделирования систем можно разделить на два больших класса: 1) методы формализованного представления систем (МФПС) и 2) методы, направленные на активизацию использования интуиции и опыта специалистов (МАИС).

В литературе можно встретить различные подходы к классификации методов моделирования систем. Так, наряду с выделением таких уровней математического абстрагирования, как общеалгебраический, теоретико-множественный, логико-лингвистический, рассматриваются информационный и эвристический уровни изучения сложных систем.

Выделяются также и другие обобщенные группы (классы) методов:

– аналитические (методы классической математики, включая интегро-дифференциальное исчисление, методы поиска экстремумов функций, вариационное исчисление, методы математического программирования; первые работы по теории игр и т. п.);

– статистические (включающие и теоретические разделы математики – теорию вероятностей, математическую статистику, и направления прикладной математики, использующие стохастические представления – теорию массового обслуживания, методы статистических испытаний (основанные на методе Монте-Карло), методы выдвижения и проверки статистических гипотез А. Вальда и другие методы статистического моделирования);

– теоретико-множественные;

– логические, лингвистические, семиотические представления (методы дискретной математики), составляющие теоретическую основу разработки языков моделирования, автоматизации проектирования, информационно-поисковых языков;

– графические (включающие теорию графов и разного рода графические представления информации типа диаграмм, гистограмм и других графиков).

Разделение методов на МАИС и МФПС находится в соответствии с основной идеей системного анализа, которая состоит в сочетании в моделях и методиках формальных и неформальных представлений, что помогает в разработке методик, выборе методов постепенной формализации отображения и анализа проблемной ситуации.

Наибольшее распространение получили следующие специальные методы моделирования систем.

1. Имитационное динамическое моделирование (System Dynamics Symulation Modeling). Использует удобный для человека структурный язык, помогающий выражать реальные взаимосвязи, отображающие в системе замкнутые контуры управления, и аналитические представления (линейные конечно-разностные уравнения), позволяющие реализовать формальное исследование полученных моделей на ЭВМ.

2. Ситуационное моделирование. Это направление базируется на отображении в памяти ЭВМ и анализе проблемных ситуаций с применением специализированного языка, разрабатываемого с помощью выразительных средств теории множеств, математической логики и теории языков.

3. Лингво-комбинаторное моделирование. Предложено для моделирования плохо формализованных систем. Заключается в том, что формальная модель строится на основе ключевых слов, характеризующих ту или иную систему. На основе ключевых слов строятся лингвистические уравнения, составленные из суммы произведений ключевых слов на смыслы. Эти лингвистические уравнения разрешаются путем введения произвольных коэффициентов, число которых определяется как число сочетаний из n по m, где n – число переменных, число разных слов, m – число ограничений, число лингвистических уравнений. Произвольные коэффициенты и их распределение по матрице эквивалентных уравнений определяет структурированную неопределенность, эти произвольные коэффициенты могут быть использованы для адаптации системы к окружающей среде. Лингво-комбинаторное моделирование – универсальный метод моделирования, с его помощью получены новые модели атомно-молекулярных структур, социально-экономических систем и процессов, биологических систем и геологических структур и т. д.

4. Логико-лингвистическое моделирование. Является развитием структурно-лингвистического моделирования, широко распространенного в 1970-е гг. в инженерной практике и основанного на использовании для реализации идей комбинаторики структурных представлений разного рода, с одной стороны, и средств математической лингвистики – с другой.

5. Системно-структурный синтез. Системно-структурные методы моделирования разрабатывались на основе иерархических и сетевых структур как средства исследования объектов и процессов с неопределенностью, когда не могут быть сразу получены математические модели.

6. Когнитивный подход (от лат. cognitio – знание, познание). Базируется на идеях когнитивной психологии. Истоки когнитивного подхода прослеживаются начиная с работ древнегреческих мыслителей (учение об универсалиях Платона). Оформление когнитивного подхода как особой дисциплины связывают с именем У. Найссера, опубликовавшего в 1967 г. книгу с изложением этого подхода, которая стала в определенном смысле программной. В настоящее время наблюдается обилие моделей, предлагаемых для интерпретации различных аспектов мыслительного процесса.

Отображение социально-правовых явлений и процессов в виде статистических закономерностей находит широкое применение при моделировании. Однако при определении и использовании закономерностей необходимо также определять правомерность их применения.

Проблемным ситуациям с большой начальной неопределенностью в большей мере соответствует представление объекта классом самоорганизующихся или развивающихся систем, который характеризуется рядом признаков, особенностей, приближающих их к реальным развивающимся объектам. Эта особенность приводит к необходимости сочетания формальных методов и методов качественного анализа. Поэтому при моделировании социально-правовых явлений и процессов необходимо использовать не только формальное, но и гуманитарное знание, не только формальные методы, но и методы, обеспечивающие активизацию интуиции и опыта субъекта, экспертов, лиц, формирующих модель и принимающих на ее основе решение.

Наиболее широко употребимый тип математических моделей, который применяется сегодня в различных исследованиях социальных, правовых, экономических явлений и процессов, – это регрессионные модели.

Регрессионным моделям будет посвящена отдельная тема дисциплины. Здесь лишь отметим, что такие модели позволяют выявлять взаимозависимости между уровнями различных явлений.

Рассмотрим некоторые результаты, полученные учеными в области криминологии на основе регрессионного моделирования.

Авторы (Год)	**Зависимые переменные**	**Влияющие переменные**	**Основные результаты**
Fleisher. The Effect of Income on Delinquency (1966)	Количество осужденных и арестов	Доходы населения, уровень безработицы, доля разведенных женщин, расовое соотношение, а также фиктивная переменная для выделения типа городов (южный и северный)	Наибольшее влияние фактора доходов населения

Авторы (Год)	Зависимые переменные	Влияющие переменные	Основные результаты
Ehrlich. Participation in Illegitimate Activities: A Theoretical and Empirical Investigation (1973)	Преступления: насильственные и против собственности	Количество арестов, осуждений, доходы населения, плотность населения, безработица, а также расовое соотношение	Подтверждена теория Г.Беккера
Sjoquist. Property Crime and Economic Behavior (1973)	Преступления против собственности: грабежи, кражи, мошенничества	Количество арестов, осуждений, доходы населения, средний срок наказания, плотность населения, образование, безработица, доля белого населения	Доказана теория Беккера: высокая вероятность ареста и степени наказания приводит к снижению преступности
Morgan Kelly. Inequality and Crime (2000)	Общий уровень преступности, а также имущественные и насильственные преступления	Население, плотность населения, индекс Джини, небелое население, уровень безработицы, уровень бедности, процент молодежи, образование, а также расходы полиции	На насильственные преступления мало влияет деятельность полиции или бедность, но они сильно зависят от неравенства (дохода или образования). Имущественная преступность не зависит от неравенства, но она стимулируется бедностью и несколько сдерживается расходами полиции
Daly, Wilson, Vasdev. Income Inequality and homicide rates in Canada and the United States (2001)	Убийства	Индекс Джини, средние доходы домохозяйств	Высокая корреляционная связь между убийствами и неравенством доходов

Авторы (Год)	Зависимые переменные	Влияющие переменные	Основные результаты
Fajnzylber, Lederman, Loayza. Inequality and Violent Crime (2002)	Убийства и грабежи	Джини, ВНП, уровень образования, уровень урбанизации	ВНП, урбанизация и уровень образования не оказывают существенного влияния на преступность
Eric Neumayer. Inequality and Violent Crime: Evidence from Data on Robbery and Violent Theft (2005)	Число грабежей на 1 млн населения	Индекс Джини, ВВП на душу населения, темпы роста ВВП, безработица, урбанизация, мужчины в возрасте 15–64	Наибольшее влияние – неравенство в доходах
Lena Edlund, Hongbin Li, Junjian Yi, Junsen Zhang. Sex Ratios and Crime: Evidence from China's One-Child Policy (2007)	Общий уровень преступности	Соотношение мужчин и женщин, доход на душу населения, безработица, неравенство в доходах, урбанизация, возрастная структура, расходы на социальное обеспечение, расходы полиции	Наибольшее значение оказывает урбанизация и численность мужского населения
A. H. Baharom, Muzafar Shah Habibulla. Crime and Income Inequality: The Case of Malaysia (2009)	Общий уровень преступности, кражи, насильственная преступность и преступления против собственности	Неравенство в доходах	Ни одна из переменных не оказывает статистически заметного влияния на уровень преступности

Авторы (Год)	Зависимые переменные	Влияющие переменные	Основные результаты
Luis Guilherme Scorzafave and Milena Karla Soares. Income Inequality and Pecuniary Crimes (2009)	Общий уровень преступности, а также преступления, связанные с незаконным оборотом наркотиков	Индекс Джини, доходы, безработица, население в возрасте 15–17 лет, миграция, урбанизация.	Неравенство в доходах оказывает наибольшее влияние, остальные факторы тоже являются статистически значимыми
Kristin Ross Balthazar. The Socioeconomic Determinants of Crime: the Case of Texas (2012)	Преступления против личности (убийства, изнасилования, хулиганство) и собственности (кражи, грабежи, угоны)	Индекс Джини, население, плотность населения, безработица, раса, бедность, расходы полиции, население в возрасте 15–24, владельцы оружия	Три фактора, оказались значимыми во всех регрессиях – плотность населения, неустойчивость семьи и соотношение между женщинами и мужчинами
Гаврилов О.А., Колемаев В. А. Математические модели в криминологии (1970)	Количество осужденных за кражи	Количество городского и сельского населения	Наибольшее влияние численности городского населения
Состояние преступности в РФ и ее долгосрочный прогноз. ВНИИ МВД России (1998)	Убийства		Регионы разбиты на 5 групп

Авторы (Год)	Зависимые переменные	Влияющие переменные	Основные результаты
Востроκнутов А.В., Коимшиди Г.Ф., Яковлев О.В. Моделирование воздействия социально-экономических факторов на криминогенную обстановку (1998)	Общий уровень преступности, отдельные виды преступлений (убийства, кражи, грабежи, разбои и др.)	Демографические показатели, уровень жизни населения, экономические показатели, состояние миграции и занятости	По силе воздействия: 1) доля несовершеннолетних 2) безработица и доля городского населения
Андриенко Ю.В. В поисках объяснения роста преступности в России (2001)	Преступления против личности	Доходы, индекс Джини	Убийства сокращаются с ростом доходов на душу населения, но растет с неравенством в распределении доходов

При проектировании моделей и их применении в гуманитарных научных разработках необходимо придерживаться некоторых принципов, соблюдение которых позволит получить адекватное и точное отображение исследуемого события или процесса. К числу этих принципов следует отнести следующие:

– принцип компромисса между ожидаемой точностью результатов моделирования и сложностью модели;

– принцип точности, выражающийся в соразмерности исходных данных и точностью в отображении объекта моделирования;

– принцип разнообразия элементов модели, позволяющий отразить многофункциональный характер исследовательских задач;

– принцип наглядности, то есть способности отобразить объект моделирования не только точно, но и максимально просто для наблюдателя;

– принцип непрерывности, охватывающий переход от максимально полного описания объекта моделирования к более простым

формам. Методологическим выражением действия этого принципа является метод *декомпозиции;*

– принцип верификации, предусматривающий возможность соответствия образа объекта его содержанию и возможности проверки этого соответствия на адекватность.

Соблюдение принципов моделирования является важнейшим условием построения модели, проектирования ее свойств, что позволит не только адекватно отобразить исследуемый объект, но и сформировать при помощи модели условия его существования и развития, направляя динамику этого объекта.

Контрольные вопросы:

1. Роль и место моделей в научных исследованиях.
2. Основания классификации моделей.
3. Примеры абстрактных моделей.
4. Примеры выводов, полученных на основе моделирования при проведении исследований в области криминологии.

2. Основы математической статистики

Последующие разделы учебного пособия посвящены построению математических моделей и их применению в исследовании различных социальных явлений, процессов и систем. Однако прежде чем приступить к построению моделей, необходимо обзавестись способом описания этих социальных объектов, пригодным для дальнейшего математического моделирования. Единственным таким способом выступает применение инструментария математической статистики. Не претендуя на глубину постижения этого инструментария, в настоящем разделе рассмотрим основные его элементы.

2.1. Задачи математической статистики

Наиболее общим образом термин «статистика» определяется как отрасль знаний, фокусирующаяся на вопросах сбора, систематизации, измерения и анализа массовых статистических данных.

Прикладная статистика — это наука о том, как обрабатывать данные произвольной природы. Действительно, в зависимости от сферы человеческой деятельности и данные, которые ее характеризуют, будут иметь различную природу, размерность, представление и источники. Поэтому существуют различные отраслевые прикладные статистики: медицинская, судебная, правоохранительная, финансовая и т. п.

Отдельно следует рассматривать статистику математическую. Математическая статистика – это отрасль научного знания, занимающаяся разработкой методов систематизации и обработки статистических данных для формирования обоснованных научных и практических выводов.

Опираясь на методы и понятия теории вероятностей, математическая статистика при этом решает в каком-то смысле обратные теории вероятностей задачи.

В теории вероятностей рассматриваются случайные величины с известным, заранее заданным распределением или случайные эксперименты, свойства которых также известны. Предмет теории вероятностей – свойства и взаимосвязи случайных величин (или их распределений).

В исследованиях социальной направленности истинные распределения значений случайных величин часто неизвестны, а эксперимент либо невозможен, либо условия его проведения неизвест-

ны или плохо формализуемы. В распоряжении исследователя, как правило, имеются лишь некие результаты, по которым требуется сделать вывод о свойствах истинного распределения случайной величины. Этим занимается математическая статистика.

В социальных и гуманитарных исследованиях статистические методы позволяют давать ответы на ряд ключевых вопросов для подтверждения или опровержения гипотез исследования.

Например, если наблюдается проявление одновременно двух (или более) характеристик у исследуемых объектов — что можно сказать о взаимозависимости этих характеристик? Имеется ли нет? И если да, то как ее можно охарактеризовать?

Либо если имеется набор числовых значений, характеризующих объекты, принадлежащие к одному классу, то можно ли выявить среди них группы объектов, похожих по совокупности показателей друг на друга?

Если имеется временной ряд, отображающий изменение во времени одного показателя, то каково наиболее вероятное значение этого показателя в следующий период времени?

Таким образом, основные задачи математической статистики следующие:

1. Создание методов сбора и обработки статистических данных, полученных в результате наблюдений за случайными процессами.

2. Разработка методов анализа статистических данных.

3. Формирование выводов по результатам анализа данных.

Анализ статистических данных включает оценку вероятностей события, функции распределения вероятностей или плотности вероятностей, оценку параметров известного распределения, оценку связей между случайными величинами.

2.2. Генеральные совокупности и выборки

Основными понятиями математической статистики, касающимися самих статистических данных, являются генеральная совокупность и выборка.

Прежде чем приступать к рассмотрению методов принятия решений на основе статистических процедур, следует рассмотреть некоторые ключевые понятия в области математической статистики и теории вероятностей. Некоторые расчетные примеры, приведенные ниже, проиллюстрируют важные концепции, которые будут полезны как сами по себе, так и в свете решения задач из последующих разделов.

В первую очередь рассмотрим понятия генеральной совокупности и выборки, а также некоторых связанных с ними определений.

Статистическая совокупность – набор некоторых объектов, которому присуща массовость и качественная однородность, но вместе с тем – наличие вариации (отличий между объектами по одному или нескольким признакам).

Объекты, образующие статистическую совокупность, – это реально существующие единицы (сотрудники, организации, подразделения, регионы, районы и т. п.), в отношении которых проводится статистическое исследование.

Единица совокупности – каждый отдельный объект статистической совокупности.

Признак – это свойство, характерная черта единиц статистической совокупности, которая может быть зафиксирована. Признаки делятся на количественные и качественные. Многообразие и изменчивость величины признака у отдельных единиц совокупности называется вариацией.

Качественная однородность – сходство всех единиц совокупности по некоторому признаку и различие по всем остальным.

В статистической совокупности отличия одной единицы совокупности от другой чаще имеют количественную природу. Количественные изменения значений признака разных единиц совокупности называются вариацией.

Атрибутивные (качественные) признаки не поддаются числовому выражению (состав населения по полу). Количественные признаки имеют числовое выражение (состав населения по возрасту).

Показатель – это обобщающая количественно-качественная характеристика какого-либо свойства единиц совокупности в целом в конкретных условиях времени и места.

Система показателей – это совокупность показателей, всесторонне отражающих изучаемое явление.

Любое статистическое исследование основывается на данных, полученных в результате измерения одного или нескольких признаков для каждого из исследуемых объектов. При этом наблюдаемая совокупность объектов, статистически представленная рядом наблюдений случайной величины или нескольких величин, является выборкой, а полная совокупность этих объектов (домысливаемая, гипотетически существующая) – генеральной совокупностью. Генеральная совокупность может быть как конечной ($N = const$), так и бесконечной ($N = \infty$), выборка же – это всегда результат ограниченного ряда наблюдений.

Выборкой (выборочной совокупностью) называется совокупность случайно отобранных объектов из генеральной совокупности.

Выборка должна быть репрезентативной (представительной), то есть ее объекты должны достаточно хорошо отражать свойства генеральной совокупности.

Выборка может быть повторной, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность, и бесповторной, при которой отобранный объект не возвращается в генеральную совокупность.

Количество объектов, вошедших в выборку, называется ее объемом. В случае когда объем выборки достаточно велик ($n \to \infty$), она считается большой, в противном случае она называется выборкой ограниченного объема.

2.3. Проблема репрезентативности выборок

В свете сказанного выше неизбежно возникает вопрос: можно ли на основе изучения ограниченной выборки экстраполировать полученные выводы на всю генеральную совокупность? Чтобы это было возможно, выборка должна обладать свойством репрезентативности.

Репрезентативность – это степень соответствия характеристик выборки характеристикам генеральной совокупности. Только данные по репрезентативным выборкам можно экстраполировать на генеральную совокупность.

Существуют различные способы достижения репрезентативности выборки. Основной – это случайная выборка, когда все объекты генеральной совокупности нумеруются и при помощи генератора случайных чисел из нее извлекаются n случайных элементов. При достаточном n выборка окажется репрезентативной, поскольку у каждого элемента генеральной совокупности равные шансы быть отобранным, а значит, распределение объектов внутри выборки будет стремиться к пропорциональному от генеральной совокупности. Как правило, случайный отбор на практике не осуществим, но есть методы, позволяющие приблизить выборку к истинно случайной или другими способами обеспечить ее близость к репрезентативной. Например, при проведении социологических опросов известны следующие способы формирования выборки: стратифицированная выборка, механический отбор, серийная выборка, метод снежного кома, стихийная выборка и др.

Правильно сформированная выборка позволяет прийти к достоверным статистическим выводам о генеральной совокупности, а следовательно и к более выверенным и адекватным управленческим решениям.

Применяют различные способы получения выборки.

1) Простой отбор – случайное извлечение объектов из генеральной совокупности с возвратом или без возврата.

2) Типический отбор, когда объекты отбираются не из всей генеральной совокупности, а из ее «типической» части.

3) Серийный отбор – объекты отбираются из генеральной совокупности не по одному, а сериями.

4) Механический отбор – генеральная совокупность «механически» делится на столько частей, сколько объектов должно войти в выборку, и из каждой части выбирается один объект.

Рассмотрим несколько задач, связанных с типическим отбором. Для формирования выборки в данном случае генеральная совокупность разбивается на группы по отдельной характеристике или по совокупности характеристик. Впоследствии из каждой группы случайно отбираются объекты в таком количестве, которое пропорционально численности данной группы.

Пример.

Дано: генеральная совокупность объектов исследования – это весь личный состав сотрудников и работников правоохранительных органов страны, в их числе из 900 000 человек 700 000 – это сотрудники (имеют специальные звания полиции или внутренней службы), 20 000 – федеральные служащие, 180 000 – работники по контракту.

Требуется: выяснить, какое количество сотрудников каждой категории войдут в выборку из 100 человек.

Решение подобной задачи состоит в следующем расчете.

Для того, чтобы выборка оказалась репрезентативной, по виду службы в нее должны войти различные категории в долях, пропорциональных генеральной совокупности. Так, если в выборку войдут 100 человек, то среди них должно быть 700 000*100/900 000 ≈ 78 сотрудников, 20 000*100/900 000 ≈ 2 федеральных служащих и 180 000*100/900 000 ≈ 20 работников.

Далее рассмотрим вопрос определения необходимого объема выборки – n, при котором ее можно считать репрезентативной. В общем случае необходимый объем зависит от двух основных показателей: вариативность исследуемого признака и желаемая достоверность результатов.

Формула для расчета объема выборки следующая:

$$n = \frac{t^2 p(100-p)}{\Delta^2}, \qquad (2.1)$$

где: t – доверительный уровень, статистическая величина, значение которой для исследований в социально-правовой сфере принято 1,96 (при 95 % точности статистического вывода);

p – % объектов, у которых предположительно проявляется признак, важный для проводимого исследования;

Δ – допустимая ошибка в %, задается произвольно при планировании исследования.

Пример.

Дано: проводится исследование устойчивости к психологическим перегрузкам сотрудников правоохранительных органов. Важным фактором является наличие или отсутствие у сотрудников опыта службы в условиях военного конфликта, чрезвычайной ситуации, контртеррористической операции. Предполагаемая доля сотрудников, несших службу в указанных условиях, – 10 %. Допустимая ошибка принята за 3 %.

Требуется: определить достаточный объем выборки.

Решение.

Пользуясь формулой (2.1) получаем необходимый объем выборки:

$$n = \frac{1{,}96^2 * 10 * (100-10)}{3^2} = 384{,}16$$

Таким образом, для проведения исследования с заданной допустимой ошибкой необходимо опросить не менее 385 человек.

Из расчетного примера видно, что необходимый объем выборки растет в первую очередь с увеличением изменчивости изучаемого признака. Так, самая объемная выборка потребуется при изменчивости в 50 %. Во-вторых, с увеличением точности желаемого результата (уменьшением допустимой ошибки) объем выборки также начинает расти.

Иногда, если объем генеральной совокупности точно известен, например – это все работники определенной организации или все автомобили определенной марки и года выпуска, то возможно еще сильнее снизить необходимый объем выборки, пользуясь для его расчета следующей формулой:

$$n = \frac{t^2 pN(100-p)}{\Delta^2 N + t^2 p(100-p)}, \qquad (2.2)$$

где: N – объем генеральной совокупности.

Обычно, если доступная для исследования выборка составляет менее 5 % от генеральной совокупности, то эта совокупность считается большой и расчеты проводятся по вышеприведенным правилам. Но в задаче 1.8. расчетный объем выборки составляет достаточно большую долю от генеральной совокупности. Если объем доступной выборки превышает 5 % от генеральной совокупности, то в объем выборки, рассчитанный по формулам (2.1) или (2.2), вводится понижающий коэффициент:

$$n_0 = n * \sqrt{\frac{N-n}{N-1}} \qquad (2.3)$$

2.4. Группировка данных

Часто для принятия управленческих решений необходимо сгруппировать объекты по некоторому признаку. Например, разделить районы города на группы с высоким и низким уровнями преступности, или разделить сотрудников на группы по возрасту и т. п. Эти задачи решаются в том числе в целях формирования оценочных показателей, в целях организации контрольно-инспекционных мероприятий, оказания методической помощи отстающим и поощрения преуспевающих. Зачастую управленческое решение содержательно состоит в том, чтобы выделить группу тех управляемых объектов, которые по определенному признаку являются отстающими или наоборот опережающими.

Разбиение на группы также необходимо для визуального отображения статистических данных, которое позволяет наглядно продемонстрировать объекты, с каким уровнем определенного показателя они встречаются и с какой частотой.

Рассмотрим примеры формирования групп объектов, опирающиеся исключительно на количество этих объектов и не учитывающие вариативность наблюдаемых параметров.

Пример.

Дано: имеется 60 обучающихся, каждый из которых набрал некоторое количество баллов по результатам прохождения теста.

Требуется: определить, на какое количество групп следует разбить обучающихся.

Решение.

В нашем распоряжении нет никаких данных о том, какие результаты показали обучающиеся. В этом случае можно воспользоваться правилом Стерджеса, предполагающим, что количество групп будет равно единице плюс логарифм от количества объектов по основанию 2:

$$k = 1 + log_2 n, \tag{2.4}$$

где n – количество объектов в выборке;

k – количество формируемых групп.

Логарифм показывает, в какую степень нужно возвести основание, чтобы получить n. В какую степень необходимо возвести двойку, чтобы получить число не менее 60? 2 в степени 6 дает 64, таким образом, по правилу Стерджеса количество групп обучающихся будет равно 7.

И если в нашем распоряжении все же есть результаты теста, то эти данные можно визуализировать в виде диаграммы плотности распределения, например так:

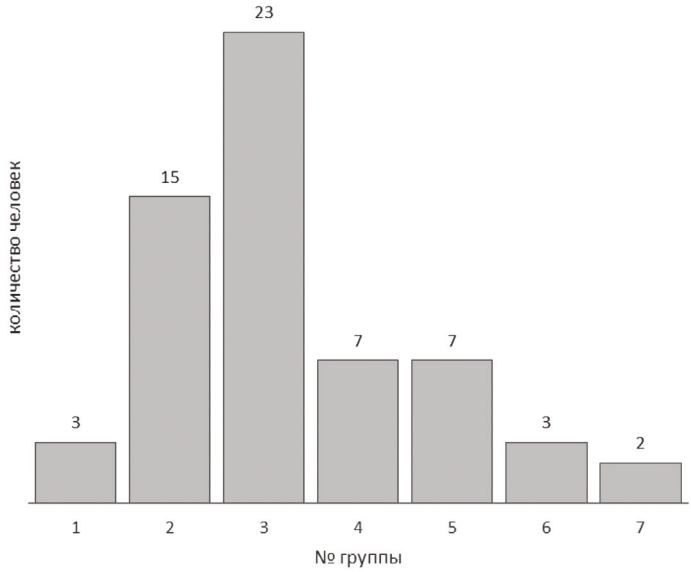

Рис. 2.1. Диаграмма плотности распределения обучающихся по группам

Границы между группами устанавливаются таким образом: вычисляется размах выборки L по исследуемому показателю, он равен разности между максимальным и минимальным значениями.

$$L = x_{max} - x_{min}, \tag{2.5}$$

где x_{max} и x_{min} – это максимальное и минимальное значения показателя.

Например, если наименьший из имеющихся результатов теста равен 15, а наибольший 100, то размах равняется $L = 85$. Далее рассчитаем длину интервала для группы:

$$l = L/k \tag{2.6}$$

Разделим 85 на количество групп: $85/7 \approx 12{,}14$.

Интервалы групп будут такими:

1: от x_{min} до $x_{min} + l$

2: от $x_{min} + l$ до $x_{min} + 2*l$

3: от $x_{min} + 2*l$ до $x_{min} + 3*l$

4: от $x_{min} + 3*l$ до $x_{min} + 4*l$

....

k: от $x_{min} + (k - 1)*$l до $x_{min} + k*l$.

В первую группу попадут обучающиеся, набравшие балл от 15 до 27,14;

во вторую – от 27,14 до 39,28;

...

в седьмую – от 87,86 до 100.

Другой способ расчета количества интервалов группировки данных, также основанный на количестве наблюдений, заключается в извлечении квадратного корня из n:

$$k = \sqrt{n} \tag{2.7}$$

Группировка данных на основе их распределения внутри выборки.

Другие способы расчета количества интервалов группировки данных связаны не с объемом выборки, а с распределением значений наблюдаемого параметра.

Например, можно рассчитать стандартное отклонение (среднеквадратичное отклонение) для выборки:

$$\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})}, \tag{2.8}$$

где: x_i – отдельное наблюдение;

$\bar{x}$ – среднеарифметическое по выборке.

Теперь можно для каждого наблюдения рассчитать нормированное значение:

$$x_i^* = \frac{(x_i - \bar{x})}{\sigma} \qquad (2.9)$$

Пример.

Сгруппируем данные из задачи 2.1. на основе стандартного отклонения. Сначала рассчитаем, какую долю от стандартного отклонения по выборке составляет расстояние от среднего в каждом наблюдении. Формула в Excel примет вид:

	A	B	C	D	E	F	G
1	Ф.И.О.	Результат тестирования	в стандартных отклонениях				
2	Бабкин А.В.	87	=(B2-СРЗНАЧ($B$2:$B$15))/СТАНДОТКЛОН.В($B$2:$B$15)				
3	Галимов А.О.	70	-0,17				
4	Гномов Н.Л.	45	-1,56				
5	Граблина А.Ф.	39	-1,89				
6	Гришин Н.Н.	92	1,04				
7	Елкина О.Т.	82	0,49				
8	Иванова И.И.	59	-0,78				
9	Лютов В.П.	64	-0,51				
10	Матюшенко Л.Р.	73	-0,01				
11	Наумова А.Ю.	75	0,10				
12	Ратова П.П.	60	-0,73				
13	Саблин Е.С.	89	0,88				
14	Семушкин Г.И.	90	0,93				
15	Филипов В.Д.	99	1,43				

Рис. 2.2. Расчет нормированных значений по выборке на основе стандартного отклонения

Теперь на листе Excel ниже полученной таблицы введем диапазоны, в которых данные будут сгруппированы, и воспользуемся функцией «ЧАСТОТА» для определения того, какое количество наблюдений оказалось в каждом интервале. Обратите внимание, что данная функция Excel предназначена для расчета массива значений. Перед вводом функции задайте диапазон ячеек, а функцию вводите при помощи сочетания клавиш Ctrl+Shift+Enter.

-3	=ЧАСТОТА($C$2:$C$15;$B$18:$B$24)
-2	0,00
-1	2,00
0	5,00
1	5,00
2	2,00
3	0,00

Рис. 2.3. Расчет количества наблюдений в каждом интервале

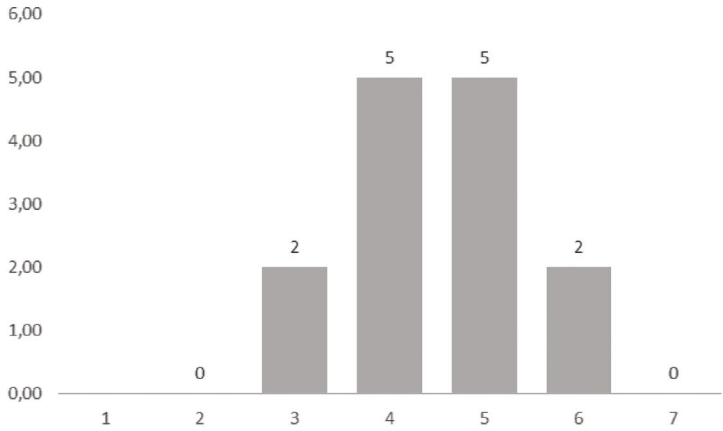

Рис. 2.4. Диаграмма

2.4. Ранжирование выборки

Для обработки данных используют операцию ранжирования, которая заключается в том, что результаты наблюдений над случайной величиной, то есть наблюдаемые значения случайной величины, располагают в порядке возрастания.

Пример.

Дана выборка: 2, 4, 7, 3, 1, 1, 3, 2, 7, 3

Проведем ранжирование выборки: 1, 1, 2, 2, 3, 3, 3, 4, 7, 7

После проведения операции ранжирования значения случайной величины объединяют в группы, то есть группируют так, что в каждой отдельной группе значения случайной величины одинаковы. Каждое такое значение называется *вариантом*. Варианты обозначаются строчными буквами латинского алфавита с индексами, соответствующими порядковому номеру группы $x_i, y_j, \ldots$.

Изменение значения варианта называется *варьированием*.

Последовательность вариантов, записанных в возрастающем порядке, называется вариационным рядом.

Число, которое показывает, сколько раз встречаются соответствующие значения вариантов в ряде наблюдений, называется частотой или весом варианта и обозначается n_i, где i – номер варианта.

Отношение частоты данного варианта к общей сумме частот называется относительной частотой или *частостью* (*долей*) соответствующего варианта и обозначается $p_i^* = (\frac{n_i}{n})$ или $p_i^* = \frac{n_i}{\sum_{i=1}^{m} n_i}$, где

m – число вариантов. Частость является статистической вероятностью появления варианта x_i. Естественно считать частость p_i^* аналогом вероятности p_i появления значения x_i случайной величины X.

Дискретным статистическим рядом называется ранжированная совокупность вариантов (x_i) с соответствующими им частотами (n_i) или частостями (p_i^*).

Дискретный статистический ряд удобно записывать в виде табл. 2.1.

Таблица 2.1

x_i	1	2	3	4	7
n_i	2	2	3	1	2
$\frac{n_i}{n}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

$$\sum_{i=1}^{5} n_i = 10;$$

$$\sum_{i=1}^{5} p_i^* = 1$$

Характеристики дискретного статистического ряда:

1. *Размах варьирования* R = $x_{max} - x_{min}$.

2. *Мода* (M_0^*) – вариант, имеющий наибольшую частоту (в примере 1. $M_0^* = 3$).

3. *Медиана* (M_e^*) – значение случайной величины, приходящееся на середину ряда.

Пусть n – объем выборки.

Если $n = 2k$, то есть ряд имеет четное число членов, то

$$M_e^* = \frac{x_k + x_{k+1}}{2} \qquad (2.10)$$

Если , то есть ряд имеет нечетное число членов, то

$$M_e^* = x_{k+1} \qquad (2.11)$$

Если изучаемая случайная величина X является непрерывной или число значений ее велико, то составляют *интервальный статистический ряд*.

Сначала определяют число интервалов m, в зависимости от объема выборки, с помощью табл. 2.2.

Таблица 2.2

Объем выборки	25–40	40–60	60–100	100–200	более 200
Число интервалов	5–6	6–8	7–10	8–12	10–15

Затем определяют длину частичного интервала h:

$$h = \frac{x_{max} - x_{min}}{m}$$ где h – шаг ; m – число интервалов.

Более точно шаг можно рассчитать с помощью формулы Стерджеса:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 lg\ n} \qquad (2.12)$$

Число интервалов $m \approx (1 + 3{,}322\ lg\ n)$.

Если шаг окажется дробным, то за длину интервала берут ближайшее целое число или ближайшую простую дробь (обычно берут интервалы одинаковые по длине, но могут быть интервалы и разной длины).

За начало первого интервала рекомендуется брать величину $x_{нач} = x_{min} - \frac{h}{2}$, а конец последнего должен удовлетворять условию $x_{кон} - h \le x_{max} < x_{кон}$. Промежуточные интервалы получают, прибавляя к концу предыдущего интервала шаг.

Просматривая результаты наблюдений, определяют, сколько значений случайной величины попало в каждый конкретный интервал. При этом в интервал включают значения, большие или равные нижней границе интервала, и меньшие – верхней границы.

В первую строку таблицы статистического распределения вписывают частичные промежутки $[x_0, x_1)$, $[x_1, x_2)$, ... , $[x_{m-1}, x_m)$.

Во вторую строку статистического ряда вписывают количество наблюдений n_i (где $i = \overline{1,m}$) попавших в каждый интервал; то есть частоты соответствующих интервалов.

При вычислении интервальных частостей округление результатов следует производить таким образом, чтобы сумма частостей была равна 1.

Иногда интервальный статистический ряд, для простоты исследований, условно заменяют дискретным. В этом случае серединное значение i-го интервала принимают за вариант x_i, а соответствующую интервальную частоту n_i – за частоту этого варианта.

2.5. Меры центральной тенденции

Среднее арифметическое – это условная величина. Реально она не существует. Реально существует общая сумма. Поэтому среднее арифметическое не есть характеристика одного наблюдения; она характеризует ряд в целом.

Среднее значение можно трактовать как центр рассеивания значений наблюдаемого признака, т. е. значения, около которого колеблются все наблюдаемые значения, причем алгебраическая сумма отклонений от среднего всегда равна нулю, т. е. суммы отклонений от среднего в большую или меньшую сторону равны между собой.

Среднее арифметическое является абстрактной (обобщающей) величиной. Даже при задании ряда только из натуральных чисел среднее значение может выражаться дробным числом. Пример: средний балл контрольной работы 3,81.

Среднее значение находится не только для однородных величин. Средняя урожайность зерновых по всей стране (кукуруза – 50–60 ц. с га. и гречиха – по 5–6 ц. с га, рожь, пшеница и т. д.), среднее потребление продуктов питания, средняя величина национального дохода на душу населения, средний показатель обеспеченности жильем, средний взвешенный показатель стоимости жилья, средняя трудоёмкость возведения здания и т. д. – это характеристики государства как единой народнохозяйственной системы, это так называемые системные средние.

В статистике широкое применение находят такие характеристики, как мода и медиана. Их называют структурными средними, т. к. значения этих характеристик определяются общей структурой ряда данных.

Иногда ряд может иметь две моды, иногда ряд может не иметь моды.

Мода является наиболее приемлемым показателем при выявлении расфасовки некоторого товара, которой отдают предпочтение покупатели; цены на товар данного вида, распространенный на рынке; как размер обуви, одежды, пользующийся наибольшим спросом; вид спорта, которым предпочитают заниматься большинство населения страны, города, поселка школы и т. д.

В строительстве существует 8 вариантов плит по ширине, и более часто применяются 3 вида: 1 м, 1,2 м и 1,5 м. По длине – 33 варианта плит, но чаще других применяются плиты длиной 4,8 м, 5,7 м и 6,0 м, мода на плиты чаще всего встречается среди этих 3-х размеров. Аналогично можно рассуждать и с марками окон.

Моду ряда данных находят тогда, когда хотят выявить некоторый типичный показатель.

Мода может быть выражена числом и словами, с точки зрения статистики мода – это экстремум частоты.

Медиана в математической статистике – число, характеризующее выборку (например набор чисел). Если все элементы выборки различны, то медиана – это такое число выборки, что ровно половина из элементов выборки больше него, а другая половина меньше него. В более общем случае медиану можно найти, упорядочив элементы выборки по возрастанию или убыванию и взяв средний элемент. Например, выборка {11, 9, 3, 5, 5} после упорядочивания превращается в {3, 5, 5, 9, 11}, и ее медианой является число 5. Если в выборке чётное число элементов, медиана может быть не определена однозначно: для числовых данных чаще всего используют полусумму двух соседних значений (то есть медиану набора {1, 3, 5, 7} принимают равной 4), подробнее см. ниже.

Также медиану можно определить для случайных величин: в этом случае она делит пополам распределение. Грубо говоря, медианой случайной величины является такое число, что вероятность получить значение случайной величины справа от него равна вероятности получить значение слева от него (и они обе равны 1/2); более точное определение см. ниже.

Можно также сказать, что медиана является 50-м персентилем, 0,5-квантилем или вторым квартилем выборки или распределения.

В теории вероятностей определили числовые характеристики для случайных величин, с помощью которых можно сравнивать однотипные случайные величины. Аналогично можно определить ряд числовых характеристик и для выборки. Поскольку эти характеристики вычисляются по статистическим данным (по данным, полученным в результате наблюдений), их называют *статистическими характеристиками*.

Пусть дано статистическое распределение выборки объема :

x_i	x_1	x_2	x_3	x_4	...	x_m
n_i	n_1	n_2	n_3	n_4	...	n_m

где m – число вариантов.

Определение. *Выборочным средним* называется среднее арифметическое всех значений выборки:

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{m} x_i n_i \qquad (2.13)$$

Выборочное среднее можно записать и так:

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{m} x_i p_i^*, \qquad \text{где } p_i^* \text{ – частость.}$$

В случае интервального статистического ряда в качестве берут середины интервалов, а n_i – соответствующие им частоты.

2.6. Меры вариативности выборки

Определение. Выборочной дисперсией $D_в$ называется среднее арифметическое квадратов отклонений значений выборки от выборочного среднего $\overline{x_в}$:

$$D_в = \frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x_в})^2 \cdot n_i \qquad (2.14.а)$$

или

$$D_в = \frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x_в})^2 \cdot p_i^* \qquad (2.14.б)$$

Выборочное среднее квадратическое выборки определяется формулой:

$$\sigma_в = \sqrt{D_в} \qquad (2.15)$$

Особенность $\sigma_в$ состоит в том, что оно измеряется в тех же единицах, что и данные выборки.

Если объем выборки мал ($n \le 30$), то пользуются *исправленной выборочной дисперсией*:

$$S^2 = \frac{n}{n-1}D_в \qquad (2.16)$$

Величина $S = \sqrt{S^2}$ называется *исправленным средним квадратическим отклонением*.

Контрольные вопросы

1. Что такое выборка и генеральная совокупность?
2. Назовите основные описательные статистики.
3. Что такое медиана? Дайте определение.
4. Назовите характеристики вариативности случайной величины.
5. Что такое частота (частость) появления случайной величины.

3. Пространственные модели анализа данных

3.1. Линейные регрессионные модели

Линейная регрессия (англ. Linear regression) – используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

Модель линейной регрессии является часто используемой и наиболее изученной в эконометрике. А именно, изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями. Необходимо отметить, что с эконометрической точки зрения более важное значение имеет линейность по параметрам, чем линейность по факторам модели.

Спецификация линейной модели парной регрессии. Основная цель регрессионного анализа – оценка функциональной зависимости между независимыми переменными *X* и условным математическим ожиданием зависимой переменной *Y*. Простая (парная) регрессия представляет собой модель, где теоретическое (среднее) значение зависимой переменной *Y* рассматривается как функция одной независимой переменной *X*: $Y_x = f(x)$. Множественная регрессия представляет собой модель, где теоретическое (среднее) значение зависимой переменной *Y* рассматривается как функция нескольких независимых переменных $X_1, X_2, \ldots X$m: $Y_x = f(x_1, x_2, \ldots, x_m)$.

Спецификация модели – формулирование вида модели, исходя из соответствующей теории связи между переменными. Определяется состав переменных и математическая функция для отражения связи между ними.

Спецификация линейной модели (уравнения) парной регрессии: $Y_i = Y_{xi} + \varepsilon_i$, где Y_i – фактическое значение зависимой переменной *Y*; Y_{xi} – теоретическое (среднее) значение зависимой переменной Y; ε_i – случайная величина (остаток регрессии).

Теоретическое уравнение регрессии (гипотетически для генеральной совокупности): $Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$,

где α – свободный коэффициент; β – коэффициент регрессии;

ε_i – случайное отклонение (возмущение).

Случайное отклонение включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Источники его присутствия в модели: спецификация модели, выборочный характер исходных данных, особенности измерения переменных.

Эмпирическое уравнение регрессии (для выборки наблюдений):

$$Y_i = a + b \cdot x_i + e_i \quad (3.1)$$

где a – эмпирическая (выборочная) оценка свободного коэффициента;

b – эмпирическая (выборочная) оценка коэффициента регрессии;

e_i – эмпирическая (выборочная) оценка теоретического случайного отклонения ε (остаток регрессии).

Метод наименьших квадратов (МНК). Суть метода наименьших квадратов (МНК) – оценки параметров таковы, что сумма квадратов отклонений фактических значений зависимой переменной Y_i от расчетных (теоретических) Y_x минимальна:

$$\sum_{i=1}^{n} (y_i - y_{xi})^2 \to \min \quad (3.2)$$

Интерпретация параметров модели

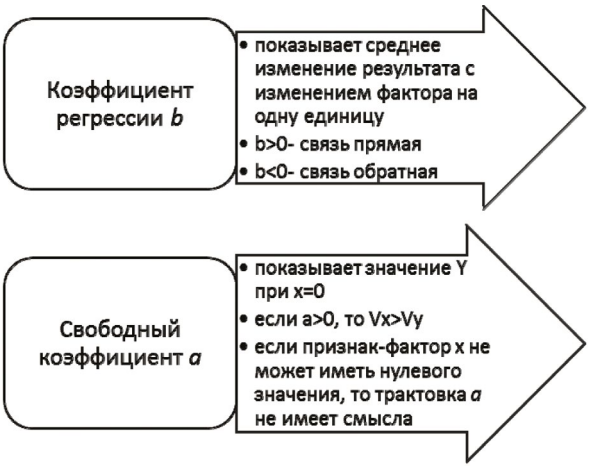

Рис. 1. Интерпретация параметров модели

Коэффициенты корреляции и детерминации в линейной модели парной регрессии. Если все точки лежат на построенной прямой, то регрессия Y на X «идеально» объясняет поведение зависимой

переменной. Обычно поведение *Y* лишь частично объясняется влиянием переменной *X*.

По абсолютной величине, чем ближе значение коэффициента корреляции r_{xy} к единице, тем теснее связь, чем ближе значение r_{xy} к нулю, тем слабее связь между *y* и *x*.

$|r_{yx}| < 0,3$ – слабая

$0,3 \le |r_{yx}| \le 0,7$ – средняя

$|r_{yx}| > 0,7$ – сильная, тесная

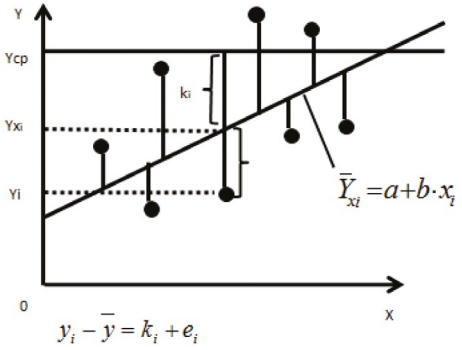

Рис. 3. Геометрическая интерпретация

Коэффициент детерминации определяет долю разброса зависимой переменной *Y*, объясняемую регрессией *Y* на *X*.

Например, возьмем два показателя, характеризующие деятельность территориальных органов МВД России на региональном уровне: 1) количество тяжких и особо тяжких преступлений, совершенных на бытовой почве (в расчете на 100 тыс. населения) и 2) число несовершеннолетних, совершивших преступления (на 1 тыс. несовершеннолетних в возрасте 14–17 лет). Заметим, что эти показатели нормативно закреплены и используются для формирования итоговой оценки результатов деятельности органов внутренних дел.

Полученная модель линейной регрессии в данном случае имеет следующую формулу: $y = 3,0042 + 0,4551*x$. Интерпретировать коэффициенты уравнения можно следующим образом:

Коэффициент a = 3,0042 – в гипотетическом случае, когда *x* равен 0, то есть не регистрируется ни одного несовершеннолетнего, совершившего преступление (на 1 тыс. несовершеннолетних в воз-

расте 14–17 лет), все равно будет зарегистрировано примерно 3 тяжких и особо тяжких преступления на бытовой почве (в расчете на 100 тыс. населения).

Коэффициент b = 0,4551 – характеризует прирост величины y при единичном приращении величины x, то есть при появлении каждых двух зарегистрированных тяжких и особо тяжких преступлений на бытовой почве (в расчете на 100 тыс. населения) модельное значение числа несовершеннолетних, совершивших преступления (на 1 тыс. несовершеннолетних в возрасте 14 – 17 лет) увеличится на 0,4551*2, примерно на единицу.

Коэффициент детерминации R^2 для этой модели равен 0,4896, то есть полученное уравнение $y = 3{,}0042 + 0{,}4551*x$ примерно на 50 % позволяет объяснить дисперсию величины y – числа несовершеннолетних, совершивших преступления (на 1 тыс. несовершеннолетних в возрасте 14–17 лет).

Заметим, что наглядная графическая интерпретации процесса моделирования и оценка построенного уравнения возможна только для парной регрессии.

3.2. Оценка модели и ее параметров

В общем случае, для линейной множественной регрессии оценка точности уравнения регрессии (да и сама возможность его построения адекватным регрессионной модели) производится на основе анализа свойств случайной компоненты ε.

Доказано, для того, чтобы метод наименьших квадратов давал лучшие результаты в поиске значений параметров модели, необходимо, чтобы случайная составляющая удовлетворяла следующим условиям (в литературе их еще называют *условиями Гаусса-Маркова*)[1]:

1. Математическое ожидание значений остатков модели равно нулю для всех n наблюдений: $M(\varepsilon) = 0$. Выполнение этого условия свидетельствует об отсутствии смещения случайной компоненты в ту или иную сторону (в «+» или «–»).

2. Для любых двух наблюдений i и j, дисперсия их остатков постоянна: $D(\varepsilon_i) = D(\varepsilon_j) \ \forall\, i \neq j$.

[1] *Бородич С.А.* Эконометрика: учебное пособие. М.: Новое знание, 2004. 416 с. (с. 122–125).

3. Значения остатков ε_i и ε_j в любых двух наблюдениях i и j независимы друг от друга $\forall\, i \neq j$.

Случайная компонента ε является нормально распределенной случайной величиной.

Значения остатков и объясняющих переменных для одного и того же наблюдения независимы между собой.

Если вышеперечисленные условия Гаусса-Маркова выполняются, то можно проводить идентификацию модели, т. е. оценку точности найденных параметров и качества построенного уравнения регрессии (его близости в регрессионной модели).

Подчеркнем, что эти оценки проводятся статистическими методами с использованием разных статистических критериев, поэтому корректнее говорить не об оценке точности и качества, а о проверках *значимости* отдельных параметров и уравнения регрессии в целом.

Оценка значимости параметров уравнения регрессии заключается в проверке справедливости гипотезы H_0 о равенстве нулю параметра при соответствующей независимой переменной X_i:

$$H_0 : b_i = 0.$$

Если гипотеза H_0 верна, то делается вывод, что полученное в результате расчетов методом наименьших квадратов значение b_i неточно. В действительности $b_i = 0$ и переменная X_i не влияет на Y (т. к. $b_i \cdot X_i = 0 \cdot X_i = 0$).). Другими словами, параметр b_i – незначим.

Для проверки гипотезы H_0 используется t-критерий Стьюдента. При этом вначале находится расчетное значение критерия Стьюдента t_{pi}:

$$t_{pi} = \frac{b_i}{S_{bi}}, \qquad (3.3)$$

где S_{bi} – стандартная ошибка параметра регрессии при переменной X_i;

Далее полученная величина t_{pi} сравнивается со критическим значением $t_{кр}$, взятым из статистической таблицы одностороннего t-распределения, для заданных: уровня значимости α (как правило, α берется равным *0,05*) и числа степеней свободы *ч.с.с.* (*ч.с.с.* = $n - m$).

Здесь следует отметить, что во многих компьютерных программах (в частности, программе Microsoft Excel) оценку значимости параметров уравнения регрессии удобнее проводить не по *t-критерию*, а по так называемому *p-значению*, рассчитываемому для каждого из параметров.

При его использовании b_i считается значимым, если соответствующее ему p–значение меньше *0,05* ($p_i < 0{,}05$). Данная пороговая величина говорит о том, что мы определили b_i с точностью *95 %*.

В таблице 3.1 представлен фрагмент распечатки результатов расчетов параметров уравнения регрессии. В последнем столбце этой таблицы и приведены *p*-значения.

Оценка значимости уравнения регрессии в целом осуществляется на основе проверки гипотезы H_0 об одновременном равенстве нулю всех параметров b_{ii} уравнения регрессии:

$$H_0: b_1 = b_2 = ... = b_n = 0$$

Таблица 3.1

Фрагмент результатов расчетов параметров уравнения регрессии

	Коэффициенты	**Стандартная ошибка**	**t-статистика**	**P-значение**
Y-пересечение	3,004153771	0,422732937	7,106505093	$4{,}43*10^{-10}$
Переменная X_1	0,455055278	0,051943588	8,760566919	$2{,}60*10^{-13}$

Если гипотеза H_0 принимается, то можно считать, что совокупное влияние всех независимых переменных X_i *(i = 1,2, ... , n)* на зависимую переменную *Y* является несущественным и найденное уравнение регрессии в целом незначимо.

Проверка данной гипотезы осуществляется на основе анализа выборочной дисперсии зависимой переменной *Y*, которая характеризует разброс ее значений в имеющемся наборе данных. Эта дисперсия *D(Y)* разбивается на две составляющие: дисперсию, объясняемую уравнением регрессии $D(Y_X)$, и дисперсию («разброс») остатков $D(\varepsilon)$:

$$D(Y) = D(Y_x) + D(\varepsilon)$$

Разделим правую и левую часть выражения (2.9) на D(Y):

$1 = \frac{D(Y_x)}{D(Y)} + \frac{D(\varepsilon)}{D(Y)}$, а затем обозначим $\frac{D(Y_x)}{D(Y)}$ через R^2 и после преобразования получаем:

$$R^2 = 1 - \frac{D(\varepsilon)}{D(Y)} \qquad (3.4)$$

Коэффициент R^2 получил название коэффициента детерминации. Из формулы видно, что если разброс остатков D(ε) равен нулю, то $R^2 = 1$.

Это означает, что построенная по исходным данным линия регрессии идеально соответствует регрессионной модели (полностью ее объясняет), параметры уравнения регрессии определены точно, и гипотеза H_0 *(*$b_1 = b_2 = ... = b_n = 0$*) м*ожет быть отклонена.

Вместе с тем, как мы уже отмечали выше, добиться положения, чтобы ε = 0 на практике невозможно. В этом случае, для проверки значимости уравнения регрессии в целом, на основе коэффициента детерминации R^2 строится F-статистика вида:

$$\frac{R \quad n}{(1-R\ )/(m-n-1)}$$

где n – число независимых переменных; m – число наблюдений.

При выполнении случайной компонентой ε условий Гаусса-Маркова данная статистика имеет распределение Фишера (*F*-распределение) со степенями свободы ч.с.с.1 = n и ч.с.с.2 = m – n – 1. Поэтому, как и в случае оценки значимости отдельных параметров уравнения регрессии, найденную величину $F_р$ сравнивают с табличным значением $F_{кр}$ при заданном уровне значимости α (α = 0,05).

Если $F_р > F_{кр}$., то гипотеза H_0 неверна и построенное уравнение регрессии значимо в целом.

Например, предположим, что на основе анализа статистических данных, взятых за 15 лет, найдено уравнение регрессии, связывающее расходы на обеспечение общественной безопасности и охрану правопорядка (объясняемая переменная *Y*) и величину валового внутреннего продукта (независимая переменная *X*):

$Y = 103,2 + 0,087*X$.

В ходе расчетов было получено и значение R^2 ($R^2 = 0,835$).

По формуле (2.12) построим *F*-статистику ($m = 1$, $n = 15$):

$$F_p = \frac{0,835/1}{0,165/13} = 65,788$$

Из таблицы распределения Фишера найдем критическое значение, соответствующее выбранному уровню значимости α = 0,05 : $F_{кр} = 4,67$.

Из сравнения $F_р$ и $F_{кр}$ можно сделать вывод значимости найденного уравнения регрессии в целом.

Использование компьютерных программ существенно облегчает поверку значимости уравнения регрессии, делая эту процедуру практически автоматической.

Таблица 3.2

Фрагмент таблицы одностороннего *F*-распределения Фишера

Ч.с.с.2	α	**Ч.с.с.1**		
		1	2	3
13	0,1	3,14	2,76	2,56
	0,05	4,67	3,81	3,41
	0,001	9,07	6,70	5,74
14	0,1	3,10	2,73	2,52
	0,05	4,60	3,74	3,34
	0,001	8,86	6,51	5,56

В таблице 3.3 приведены результаты дисперсионного анализа, полученного с помощью Microsoft Excel.

Таблица 3.3

Результаты дисперсионного анализа

	df	**SS**	**MS**	**F**	**Значимость F**
Регрессия	1	42432,1	42432,1	65,82388	0,01231
Остаток	13	8380,2	644,6308		
Итого	14	50812,3			

В строках этой таблицы (согласно уравнению *2.5*) отражены основные характеристики двух компонент регрессионной модели: уравнения регрессии (строка – «регрессия») и случайной составляющей (строка – «остаток»).

В столбце **df** приведены значения числа степеней свободы, связанных с каждой из компонент; в столбце **SS** – величины дисперсии, объясняемой уравнением регрессии *($D(Y_x)$)* и дисперсии остатков *($D(\varepsilon)$);* в столбце **MS** – те же значения, но пересчитанные на одну степень свободы.

Кроме того, в таблице находятся *F*-расчетное значение критерия Фишера (*F = 65,82388)* и результаты оценки его значимости (в столбце **Значимость F**). Если число, характеризующее значимость, меньше *0,05*, то уравнение регрессии признается значимым в целом.

3.3. Проблема мультиколлинеарности и неоднородности данных. Использование фиктивных переменных

Мультиколлинеарность – это линейная взаимосвязь двух или нескольких объясняющих переменных (x_1, x_2, ... x_m). Если объясняющие переменные связаны строгой функциональной зависимостью, то говорят о совершенной мультиколлинеарности. Последствия мультиколлинеарности: увеличиваются стандартные ошибки оценок; уменьшаются *t*-статистики МНК-оценок регрессии; МНК-оценки чувствительны к изменениям данных; возможность неверного знака МНК-оценок; трудность в определении вклада независимых переменных в дисперсию зависимой переменной.

Обнаружение мультиколлинеарности и способы ее устранения или снижения. Признаки мультиколлинеарности: высокий R^2; близкая к 1 парная корреляция между малозначимыми независимыми переменными; высокие частные коэффициенты корреляции; сильная дополнительная регрессия между независимыми переменными. Методы устранения мультиколлинеарности: исключение из модели коррелированных переменных (при отборе факторов); сбор дополнительных данных или новая выборка; изменение спецификации модели; использование предварительной информации о параметрах; преобразование переменных.

Исходные статистические данные называют однородными, если все они зарегистрированы при одних и тех же условиях (время года, регион, образование, пол человека). Если же данные объединяют в себе наблюдения, зарегистрированные при различных условиях, то они могут быть неоднородными. В этом случае в модель включается фактор, имеющий два или более качественных уровней. Фиктивные (dummyvariables, искусственные, двоичные, структурные) переменные отражают в модели влияние качественного фактора, содержащего атрибутивные признаки двух и более уровней. Для того, чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки, то есть качественные переменные необходимо преобразовать в количественные.

Правило использования фиктивных переменных. В случае, когда качественная переменная принимает не два, а большее число значений, может возникнуть ситуация, которая называется ловушкой фиктивной переменной. Она возникает, когда для моделирования k значений качественного признака используется ровно k бинарных (фиктивных) переменных. В этом случае одна из таких переменных линейно выражается через все остальные, и матрица $(X'X)$ становится вырожденной. Тогда исследователь попадает в ситуацию совершенной мультиколлинеарности. Избежать подобной ловушки позволяет правило: если качественная переменная имеет k альтернативных значений, то при моделировании используется только $(k-1)$ фиктивных переменных.

Характерным представителем факторных социальных моделей являются криминологические модели, с помощью которых пытаются формализовать сложные социальные процессы, связанные с противоправным поведением, развитием преступности, формированием механизмов правового регулирования.

Создание пространственной модели представляет собой итерационный процесс, направленный на поиск эффективных независимых переменных. Основная цель заключается в попытке объяснения зависимых переменных, которые подлежат моделированию и осмыслению, запуская инструмент регрессии, и определения того, какие величины являются эффективными предикторами. Затем следует методично удалять и/или добавлять переменные до тех пор, пока вы не найдете наилучшим образом подходящую регрессионную модель. Так как процесс ее создания – занятие творческое, он никогда не должен превращаться в простую «подгонку» данных. Следует учитывать теоретические аспекты, мнение экспертов в этой области и здравый смысл. Современный исследователь социальных и правовых процессов и явлений должен быть способен определить ожидаемую взаимосвязь между каждой потенциальной независимой переменной и зависимой величиной, желательно еще до проведения самого анализа, а если эти связи не совпадают – задавать дополнительные вопросы и находить адекватные решения.

Контрольные вопросы:

1. Предпосылки построения регрессионных моделей.
2. Интерпретация коэффициентов уравнения регрессионной модели.
3. Интерпретация коэффициентов корреляции и детерминации.
4. Проверка значимости уравнения регрессии и его коэффициентов.
5. Визуализация статистической взаимозависимости при помощи диаграммы разброса.

4. Временные модели анализа данных

В отличие от анализа случайных выборок анализ временных рядов основывается на предположении, что последовательные значения в данных наблюдаются через равные промежутки времени (тогда как в других методах нам не важна и часто не интересна привязка наблюдений ко времени).

Изучение взаимосвязей между социально-правовыми явлениями развивается по двум направлениям: в первом из них явления рассматриваются во взаимосвязи относительно одного и того же момента времени или временного промежутка. При этом для имеющихся в распоряжении исследователя данных не важен порядок или последовательность их расположения.

Второе направление связано с рассматриванием последовательности значений данных, характеризующих изучаемое явление, причем эта последовательность имеет принципиальное значение. Значение (или уровень ряда) соответствует определенному моменту t (времени), и эти моменты располагаются в хронологическом порядке.

Существуют две основные цели анализа временных рядов: (1) определение природы ряда и (2) прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям). Обе эти цели требуют, чтобы модель ряда была идентифицирована и более или менее формально описана. Как только модель определена, с ее помощью интерпретируются рассматриваемые данные (например, для понимания сезонного изменения преступности).

В распоряжении исследователя, сфера научных интересов которого связана так или иначе с деятельностью органов внутренних дел в Российской Федерации, сегодня имеется обширный спектр временных рядов, характеризующих развитие криминальной ситуации в СССР, Российской Федерации на достаточно длинном промежутке времени.

Современная система учета различных показателей предполагает формирование рядов различных интервалов: от ежесуточной криминальной статистики до квартальных и годовых сводок как по России в целом, так и по ее регионам.

4.1. Понятие временного ряда, его основные характеристики и компоненты

Под *временным рядом* или *динамическим рядом (рядом динамики)* будем понимать упорядоченную последовательность значений показателей, характеризующих изменение изучаемых явлений во времени. Например, динамика преступности может быть описана таким показателем, как количество зарегистрированных тяжких преступлений и т. п.

Отдельные числовые значения выбранного показателя, связанные с определенными моментами времени, называются *уровнями ряда*.

В таблице 4.1 приведен пример интервального временного ряда, уровни которого фиксируют количество зарегистрированных грабежей за определенный год в одном из субъектов Российской Федерации. На рисунке 4.1 представлен тот же ряд в графической форме (для наглядности все точки соединены отрезками ломанной).

Таблица 4.1

Табличная форма представления временного ряда

Показатель	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Число зарегистрированных грабежей	1631	1219	1403	1244	1464	1319	1701	1848	2050	1783

Числовые значения показателя временного ряда служат исходными данными для построения временной модели. По аналогии с пространственной моделью, она позволяет исследовать влияние независимых переменных X_i на поведение зависимой переменной Y, с тем отличием, что в качестве независимых во временной модели рассматривается только одна переменная – t (время). (Чтобы подчеркнуть эту разницу, в обозначении зависимой переменной будем применять нижний индекс t : Y_t).

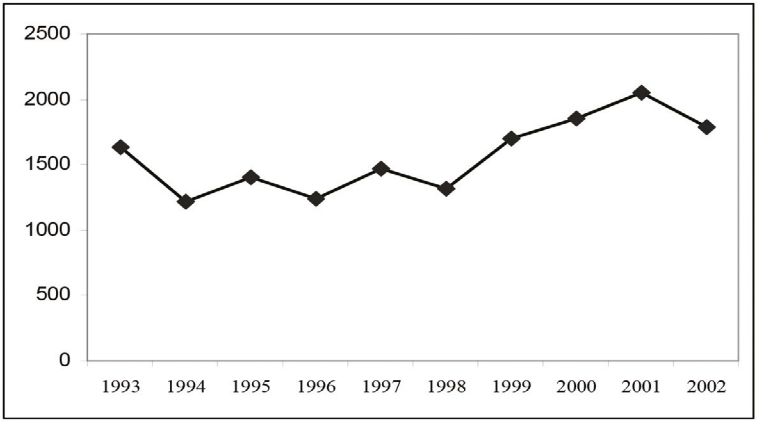

Рис 4.1. Графическая форма представления временного ряда

Использование временных моделей анализа данных, по сравнению с пространственными моделями, имеет определенные преимущества. Как показывает практика, в распоряжении аналитика органов внутренних дел не всегда имеются достаточные информационные массивы о социально-экономических факторах, знание которых необходимо для оценки их влияния на преступность. Эти массивы формируются Департаментами и отделами экономического развития субъектов Российской Федерации на региональном и районном уровнях, зачастую в агрегированном виде.

В случае временных рядов мы, при проведении анализа, можем опираться только на данные ведомственной статистики, что, несомненно, более удобно по причине доступности и оперативности.

В процессе построения временных моделей можно выделить два основных этапа.

На первом этапе необходимо оценить обобщенные характеристики изменений уровня ряда и постараться выявить те или иные его качественные свойства. Как правило, останавливаются на таких характеристиках как:

– *абсолютный прирост (цепной)* – рассчитывается как разность двух соседних уровней ряда: $\Delta Y = Y_t - Y_{t-1}$;

– *темп роста (цепной)* – отношение двух соседних уровней ряда, выраженное в процентах: $T_{pi} = \frac{Y_t}{Y_{t-1}} \cdot 100$; (4.1)

– *темп прироста* – отношение абсолютного прироста к базе сравнения: $T_{np} = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \cdot 100$; (4.2)

– *средний темп роста* – характеризует интенсивность изменения уровней ряда $\overline{T_p} = \sqrt[t]{\prod_{i=1}^{t} T_{pi}}$ (4.3)

Второй этап связан с поиском существующих механизмов изменения уровней ряда.

При этом отметим, что каждый уровень временного ряда включает в себя две основные компоненты: *регулярную R(t)* и *случайную* ε_t:

$$Y_t = R(t) + \varepsilon_t \quad (4.4)$$

Регулярная компонента, в свою очередь, делится на *тренд f(t)* и *циклическую составляющую* c_t.

Тренд *f(t)* формируется под влиянием совокупного воздействия социально-экономических факторов, влияние которых

на динамику исследуемого явления (например преступность) носит долговременный характер (размер валового внутреннего продукта, состояние промышленного и сельскохозяйственного производства, *уровень* безработицы и т. п.). Эти факторы могут оказывать разнонаправленное воздействие, но все вместе, они генерируют возрастающую или убывающую *тенденцию*.

Наличие циклической составляющей c_t является следствием сезонности, которая, как фактор, отражается и в специфике деятельности органов внутренних дел (например, понимание того, что динамика квартирных краж носит сезонный характер, может помочь повысить эффективность их предотвращения).

На формирование случайной компоненты оказывают влияние факторы, которые или непредсказуемы, или очень незначительны. Их воздействие на случайную компоненту проявляется в виде спадов и подъемов последней, в которых традиционными способами трудно установить какие-то закономерности. (По этой причине, а также потому, что в органах внутренних дел подобные факторы крайне редки, вопросы анализа случайной компоненты выходят за рамки настоящего учебника).

Таким образом, задача второго этапа состоит в том, чтобы на основании данных временного ряда обосновать ту или иную аналитическую форму для регулярной компоненты.

Для решения этой задачи можно предложить использовать ряд подходов.

4.2. Методы исследования компонент временного ряда

1. Метод последовательных разностей.

Данный метод основывается на представлении временного ряда Y_t в виде двух компонент: $R(t)$ и ε_t (см. формулу *(2.13)*).

Предположим, что для описания регулярной компоненты предлагается использовать полином первой степени: $R(t) = a + b \cdot t$.

Для каждой из точек $t = 1,2, ..., n$ рассчитаем:

$$Y_1 = a + b + \varepsilon_1$$

$$Y_2 = a + 2 \cdot b + \varepsilon_2$$

....

$$Y_n = a + n \cdot b + \varepsilon_n$$

Далее, найдем первые разности. Ими называют разности между последовательными уровнями ряда:

$$\Delta'1 = Y_2 - Y_1 = b + (\varepsilon_2 - \varepsilon_1)$$
$$\Delta'2 = Y_3 - Y_2 = b + (\varepsilon_3 - \varepsilon_2)$$
$$\ldots$$
$$\Delta'_{n-1} = Y_n - Y_{n-1} = b + (\varepsilon_n - \varepsilon_{n-1})$$

Из выражений для первых разностей видно, что они зависят только от разности ε_t, т. е. от величин случайной компоненты в соседних наблюдениях.

Если они незначительно отличаются друг от друга (т. е. $\forall_t : t = 1,2, \ldots, n : \varepsilon_t - \varepsilon_{t-1} \approx 0$ и $\Delta'_1 \approx \Delta'_2 \approx \ldots \approx \Delta'_{n-1}$), выбор полинома первой степени является правильным.

Таким образом, получаем : $Y_t = a + b \cdot t$ (4.5)

В случае больших расхождений первых разностей, вычисляется вторые разности и т. д.:

$$\Delta^2_1 = \Delta'_2 - \Delta'_1$$
$$\Delta^2_2 = \Delta'_3 - \Delta'_2$$
$$\ldots$$
$$\Delta^2_{n-1} = \Delta'_n - \Delta'_{n-1}$$

Порядок разностей, при котором достигается их равенство, соответствует порядку искомого полинома. Например, если это условие выполняется для разностей второго порядка, то $Y_t = a + b \cdot t + c \cdot t^2$

Дальнейшее вычисление параметров временной модели *a, b, c* осуществляется методом наименьших квадратов.

2. *Метод подбора аналитической функции.*

Данный метод получил большое распространение, поскольку он используется во многих компьютерных программах, в т. ч. и в программе Microsoft Excel.

Его суть состоит в выборе из стандартного набора аналитических функций той, которая лучше соответствует имеющемуся набору данных.

Процедура выбора осуществляется среди следующих аналитических функций:

– линейной: $R(t) = a + bt;$ (4.6)

– параболической: $R(t) = a + bt + ct^2$ (в Microsoft Excel показатель степени может также принимать значения в диапазоне от 3 до 6. В этом случае функция будет представлять собой полином соответствующей степени); (4.7)

– степенной: $R(t) = a \cdot t^b$ (4.8)

– экспоненциальной: $R(t) = a \cdot \exp^{b \cdot t}$ (4.9)

– гиперболической: $R(t) = a + b \cdot \frac{1}{t}$; (4.10)

– логарифмической: $R(t) = a + b \cdot 1nt$. (4.11)

Степень соответствия измеряется с помощью рассчитываемого коэффициента R^2. Функция, имеющая бóльшее значение R^2, и является искомой.

Как и в случае пространственной модели, графическая интерпретация данного метода заключается в поиске линии, которая наиболее близка ко всем точкам, лежащим на плоскости и характеризующим уровни ряда. Показателем этого служит также величина коэффициента R^2. Он должен быть максимально близок к 1.

Технологию выбора аналитической функции, описывающей регулярную компоненту, рассмотрим на примере временного ряда, данные которого изображены в виде точек на рис. 4.2 (весь набор данных включает 23 точки). Для упрощения процедуры воспользуемся программой Microsoft Excel.

После ввода данных в электронную таблицу и построения по этим данным точечной диаграммы в контекстном меню вызовем команду «Добавить линию тренда». В закладке «Тип» представлены шесть возможных типов функций, опции в закладке «Параметры» позволяют вывести на экран аналитическое выражение выбранной функции и соответствующую ей величину коэффициента R^2. На рис. 4.2 представлены также график (сплошная линия) и параметры оцениваемой нами линейной функции ($R(t) = 300401 + 45190 \cdot t$; $R^2 = 0{,}965$).

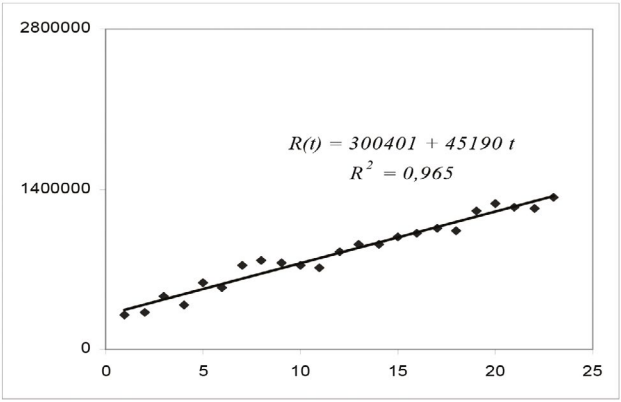

Рис. 4.2. Исходный ряд и параметры выбранной линейной функции

Полученные результаты свидетельствуют о ее хорошем приближении к имеющимся данным и адекватности описания с ее помощью динамики регулярной компоненты.

Вместе с тем, и в случае выбора параболической функции, мы также можем получить похожие результаты (рис. 4.3.).

Для того, чтобы сделать вывод о возможности использования той или иной аналитической функции, имеющей хорошие оценки, в качестве временной модели, необходимо для каждой из них провести дополнительное исследование соответствующей случайной компоненты. Как и в случае пространственных моделей, она должна являться нормально распределенной случайной величиной (см. четвертое условие Гаусса-Маркова).

По формулам, приведенным ниже, для каждого $t = 1, 2, ..., 23$ вычислим остатки для линейной и параболической функции соответственно, и итоги занесем в столбцы электронной таблицы:

$\varepsilon_t = Y_t - R(t) = Y_t - 300401 + 45190 \cdot t$

$\varepsilon_t = Y_t - R(t) = Y_t - R(t) = Y_t - 260399 + 54790\, t - 400{,}02\, t^2$

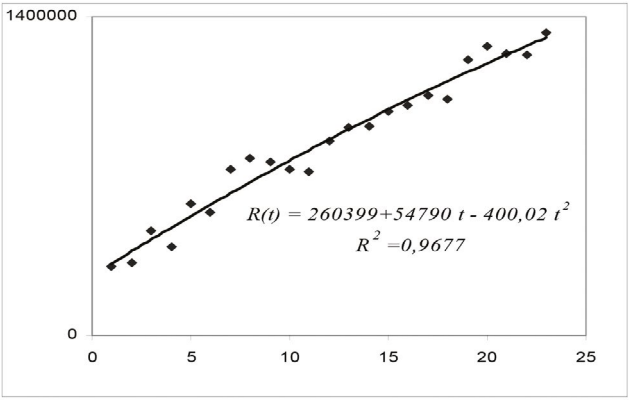

Рис. 4.3. Исходный ряд и параметры выбранной параболической функции

На рисунках 4.4 и 4.5 приведены гистограммы плотности распределения значений остатков для линейной модели и модели, представленной полиномом второй степени соответственно.

Из рисунков видно, что в качестве временной модели следует использовать линейную функцию. График плотности распределения ее остатков больше напоминает вид нормального распределения (рис. 4.5), по сравнению с графиком гистограммы остатков параболической функции. Это означает, что в линейной модели значения случайной компоненты носят действительно более случайный характер, в то

время как в полиномиальной модели наблюдается аномально большое число значений, сильно отличающихся от среднего в большую сторону.

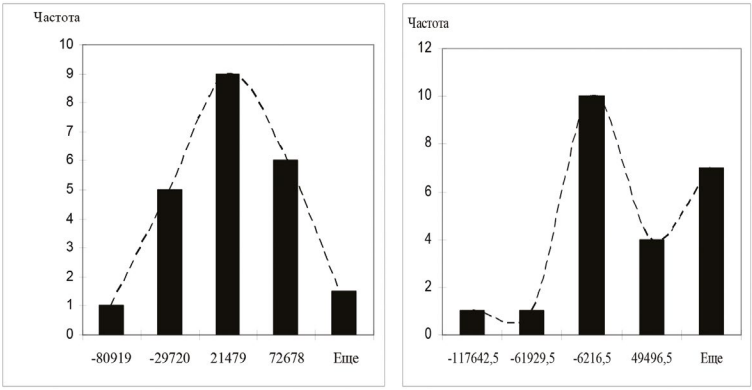

Рис. 4.4. Гистограмма остатков линейной (слева) и параболической (справа) функции

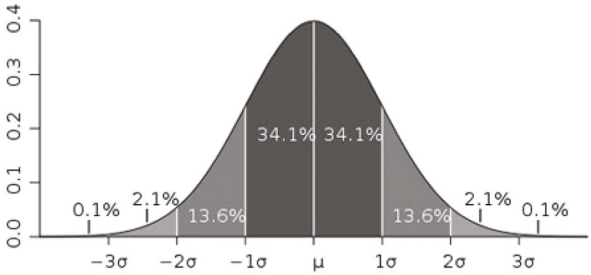

Рис. 4.5. График плотности вероятности нормального распределения и процент попадания случайной величины на отрезки, равные стандартному отклонению

4.3. Адаптивные методы исследования временных рядов

Основное отличие адаптивных методов от методов, описанных выше, состоит в том, что в получаемом с их помощью формальном выражении для регулярной компоненты заложена возможность его

самонастройки в зависимости от степени значимости тех или иных уровней временного ряда.

Достигается эта возможность за счет установления различных весов разным наблюдениям, например в результате проведенного анализа исходных данных поздним уровням ряда (близким к текущему моменту времени) могут присваиваться бóльшие веса (как более значимым), чем ранним. Тем самым подчеркивается, какая часть данных временного ряда устарела, а какая продолжает оставаться актуальной (особенно это качество ценно для длинных временных рядов).

Одним из самых распространенных адаптивных методов, реализованных с помощью компьютерных технологий, является *метод экспоненциального сглаживания*. Его формула имеет вид:

$$S_t = \alpha \cdot Y_t + (1 - \alpha) \cdot S_{t-1}, \qquad (4.12)$$

где St – расчетное сглаженное значение в момент времени t; S_{t-1} – расчетное сглаженное значение в момент времени $t - 1$; α – параметр сглаживания: $0 < \alpha < 1$; $S_0 = Y_1$

Заметим, что если $\alpha = 0$, то $S_t = S_{t-1}$, что означает – значимыми являются более ранние наблюдения. Если $\alpha = 1$, то $S_t = Y_{t-1}$ – игнорируются первые наблюдения. Однако, на практике α не принимает эти крайние значения, а находится в диапазоне от 0 до 1.

После некоторого преобразования соотношение приведенное выше можно представить следующим образом:

$$St = S_{t-1} + \cdot\alpha\,(Y_t - S_{t-1}) \qquad (4.13)$$

Из формулы видно, что каждое новое расчетное сглаженное значение ряда вычисляется как сумма предыдущего сглаженного значения S_{t-1} и взвешенной погрешности сглаживания: $Y_t - S_{t-1}$

Для применения метода экспоненциального сглаживания необходимо научиться правильно определять параметр сглаживания α. Во многих компьютерных программах поиск α осуществляется в автоматическом режиме.

Для оценки параметра α в Microsoft Excel можно использовать подход, основанный на анализе нормальности остатков, который был нами апробирован в методе подбора аналитической функции.

В качестве примера рассмотрим временной ряд, имеющий высокую волатильность и демонстрирующий сезонной характер регистрируемой преступности (рис. 4.6 – сплошная ломаная линия).

Очевидно, что использовать для такого ряда методы определения регулярной компоненты, связанные с подбором гладких аналитических функций, не представляется возможным. Воспользуемся методом экспоненциального сглаживания.

С этой целью в модуле «Анализ данных» выберем «Экспоненциальное сглаживания», и после заполнения входного интервала адресами ячеек, содержащих исходные данные, зададим фактор затухания (так в Microsoft Excel называется параметр сглаживания). На первом этапе установим $\alpha = 0{,}1$. Проведем вычисления и $\forall\, t = 1, 2, ..., 23$, подсчитаем остатки по формуле:

$$\varepsilon_t = Y_t - S_t \qquad (4.14)$$

Повторим эту процедуру, меняя α последовательно от 0,2 до 0,9 с шагом 0,1: $\alpha = 0{,}3;\ 0{,}4;\ 0{,}5;\ 0{,}6;\ 0{,}7;\ 0{,}8;\ 0{,}9$

Далее для каждого ряда остатков, с помощью уже знакомого нам инструмента «Гистограмма», построим свои гистограммы. Значение параметра сглаживания, гистограмма остатков для которого наиболее близка к виду нормального распределения, и является искомым.

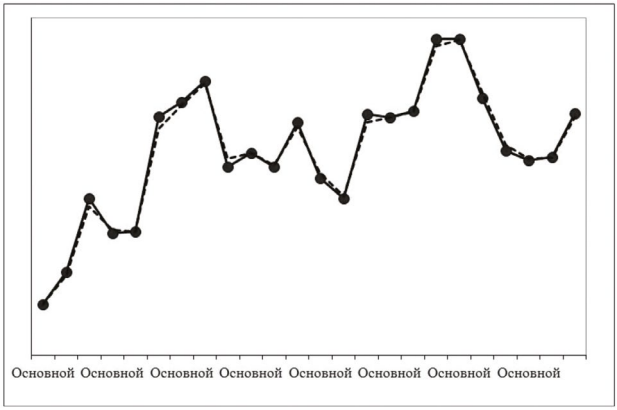

Рис. 4.6. Исходный временной ряд – сплошная линия;
сглаженный временной ряд ($\alpha = 0{,}2$) – пунктирная линия

В качестве примера на рисунке 4.7 приведены гистограммы остатков для $\alpha = 0{,}2$ и $\alpha = 0{,}5$ соответственно.

Поскольку наиболее близка к виду нормального распределения гистограмма остатков для $\alpha = 0{,}2$, то в качестве временной модели анализа данных можно использовать уравнение вида:

$$Y_t = 5 \cdot S_t + 4 \cdot S_{t-1}$$

График сглаженного временного ряда с использованием параметра сглаживания $\alpha = 0{,}2$ представлен на рис. 4.6 пунктирной ломаной линией. Как видно из рисунка, он практически совпадает с первоначальным графиком.

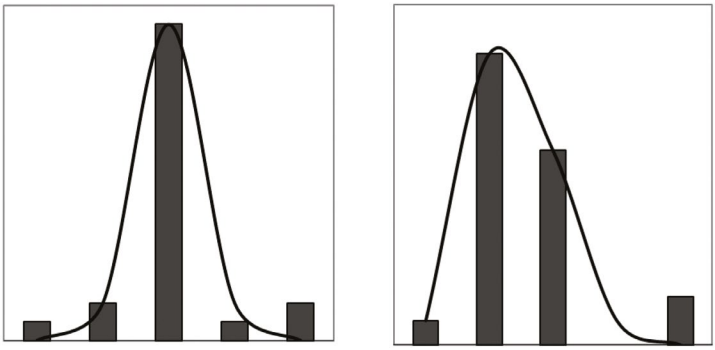

Рис. 4.7. Гистограмма остатков $\alpha = 0{,}2$ (слева) и $\alpha = 0{,}5$ (справа)

Таким образом, с помощью правильного подбора параметра затухания мы получаем возможность построить временную модель, имеющую минимальные отклонения от исходных статистических данных.

Методы анализа временных рядов существенно отличаются от методов анализа данных пространственной модели. При анализе временного ряда исследователя интересуют не только статистические характеристики временного ряда, но и учитывается взаимосвязь измерений со временем. Временные ряды, как правило, возникают в результате измерения некоторого показателя. Это могут быть как характеристики технических систем, так и показатели природных, социально-экономических явлений и процессов. Например, динамика курса валюты или курса акции, при анализе которых пытаются определить основное направление развития, т. е. тренд. Или, например, анализ динамики преступлений с целью планирования мероприятий по противодействию преступности.

Владение исследователей инструментарием анализа закономерностей развития этих показателей во времени – залог качества современных научных разработок в социальной и правовой сфере.

Контрольные вопросы

1. В чем состоит отличие пространственной и временной формы представления данных?

2. Назовите исходные предпосылки построения пространственных моделей анализа данных.

3. Какие существуют проблемы в практике анализа данных?

4. В чем состоит суть метода наименьших квадратов?

5. Как проверить значимость уравнения регрессии и его коэффициентов?

6. Как проверить значимость уравнения регрессии в целом?

7. Назовите основные особенности построения временных моделей анализа данных.

8. Перечислите основные компоненты временного ряда.

9. Что такое тренд?

10. С какими целями проводятся выявление регулярных компонент временного ряда?

11. Какие методы используются для нахождения регулярной компоненты временного ряда?

12. В чем состоит суть метода экспоненциального сглаживания?

5. Компьютерные технологии обработки результатов анкетных опросов

Обработка и анализ информации, по мнению В. А. Ядова (выдающегося отечественного социолога, автора многих работ по методике и методологии исследований), – самый увлекательный этап исследования. Обрабатывая, преобразуя полученные данные, мы делаем из разрозненного, сырого материала некоторую схему, картину, выявляем связи и закономерности. Мы проверяем те предположения, которые были у нас при составлении программы, выдвижении гипотез, которые возникали при сборе данных.

На сегодняшний день в органах внутренних дел вопросы обработки результатов социологических опросов носят самый актуальный характер. По приказу МВД России № 1246 от 30.12.2007 ежегодно во всех субъектах РФ опросы общественного мнения по проблемам безопасности граждан и деятельности органов внутренних дел проводят независимые социологические центры. Ежегодно более тысячи интервьюеров опрашивали более 300 тыс. респондентов в более чем 2 260 населенных пунктов Российской Федерации. После обработки данных в сентябре – декабре представляются сводные массивы по субъектам РФ и по социально-демографическим группам, аналитические обзоры результатов и методические материалы.

Результаты этих опросов не только используются при формировании оценочных показателей результатов деятельности территориальных органов МВД России на региональном уровне, но и представляют существенный научный интерес.

Кроме того, результаты социологических опросов среди действующих и бывших сотрудников органов внутренних дел способны внести существенный вклад в исследования, посвященные совершенствованию различных направлений оперативно-служебной деятельности.

5.1. Технологии первичной обработки данных анкетных опросов

Первичная компьютерная обработка результатов опроса есть последовательная «набивка» каждого заполненного документа (анкеты, бланка и т. д.) – набивка всех вариантов ответов на все

вопросы в единую матрицу. Результатом является возможность получить данные (как в абсолютных числах, так и в процентах) по каждой позиции в каждой группе и по всему массиву (они обычно называются линейными распределениями); данные о связях тех или иных ответов на ту или иную (любую) пару или даже несколько вопросов (эти данные принято называть корреляциями); разного рода коэффициенты и т. п.

До появления компьютерных программ ручная обработка нередко производилась подобным же образом – на больших листах миллиметровой бумаги чертилась так называемая матричная таблица. В каждой строке таблицы собирались все ответы на все вопросы по одному документу, а каждый столбец соответствовал одному варианту ответа на один вопрос. На самом деле эта работа была не только крайне трудоемкой, но и не очень эффективной.

Если обрабатываются исследовательские документы вручную, особенно важно знать (и лучше подумать об этом еще на этапе составления программы исследования), хотите ли вы просто получить информацию о том, как ответят на тот или иной вопрос все ваши респонденты, – или вас интересует специфика ответов представителей той или иной группы.

Но логика исследования нередко требует более тщательной проработки, установления связей между двумя или более характеристиками респондента. Речь в общем случае идет о том, от чего именно зависит появление или превалирование того или иного ответа на тот или иной вопрос. Практика показывает, что чаще всего определяющей, объясняющей характеристикой оказывается социально-демографическая – возраст, пол, образование, место работы или учебы, стаж службы и т. п.

В любом случае, первый шаг в обработке результатов анкетирования – это заполнение матрицы результатов опроса. Существует два основных способа заполнения такой матрицы.

1. Один столбец – один вариант ответа.

В этом случае каждая ячейка матрицы будет содержать бинарный признак, принимающий одно значение (например 1), – в случае если респондент выбрал данный вариант ответа, и другое значение (например, 0), если респондент этот вариант ответа не выбрал. Такое заполнение матрицы подходит как для вопросов с одинарным выбором, так и для вопросов с множественным выбором.

Пример

	Вопрос № 1. Укажите Вашу возрастную группу						**Вопрос № 2. Укажите Ваш пол**		**Вопрос № 3. Как Вы относитесь к увеличению пенсионного возраста?**	
№	1. до 25 лет	2. от 26 до 35 лет	3. от 36 до 45 лет	4. от 46 до 55 лет	5. от 56 до 65 лет	6. от 66 лет	м	ж	положительно	отрицательно
1	0	1	0	0	0	0	1	0	1	0
2	0	0	1	0	0	0	0	1	0	1
3	0	0	0	0	1	0	0	1	0	1
4	0	0	1	0	0	0	1	0	0	1
5	0	0	0	1	0	0	0	1	1	0
6	0	1	0	0	0	0	1	0	0	1

Уже на данном этапе возможен и более глубокий анализ анкетных данных нежели подсчет долей респондентов, выбравших тот или иной вариант ответа.

2. Второй вариант заполнения матрицы результатов анкетирования. Один столбец – один вариант вопроса.

В этом случае каждому варианту ответа ставится в соответствие некоторый код, например:

Вопрос № 1. Каков Ваш возраст (полных лет)?

Вариант ответа:	код
1) от 18 до 25 лет	1
2) от 26 до 35 лет	2
3) от 36 до 45 лет	3
4) от 46 до 55 лет	4
5) от 56 до 65 лет	5
6) свыше 65 лет	6

Вопрос № 10. Каков Ваш стаж службы в органах внутренних дел (полных лет)?

Вариант ответа:	код
1) менее 1 года	1
2) от 1 года до 5 лет	2
3) от 6 до 10 лет	3
4) от 11 до 15 лет	4
5) от 16 до 20 лет	5
6) свыше 20 лет	6

Вопрос № 11. К какой службе Вы принадлежите?

Вариант ответа:	код
1) следствие	1
2) дознание	2
3) оперативное подразделение	3
4) тыловая служба	4
5) кадровая служба	5
6) иное	99

Заполненная матрица в этом случае будет выглядеть следующим образом.

№	Вопрос № 1. Каков Ваш возраст (полных лет)?	…	Вопрос № 10. Каков Ваш стаж службы в органах внутренних дел (полных лет)?	Вопрос № 11. К какой службе Вы принадлежите?
1	3	…	1	3
2	2	…	3	2
3	4	…	2	5
4	5	…	6	99
5	1	…	1	1
6	2		2	2

5.2. Исследование коррелированности вариантов ответов, выбираемых респондентами

Корреляция (от лат. *correlatio* «соотношение, взаимосвязь») или корреляционная зависимость — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих вели-

чин сопутствуют систематическому изменению значений другой или других величин.

Выбор респондентами тех или иных вариантов ответов на различные вопросы также может быть взаимообусловлен, то есть коррелирован. Численная характеристика такой коррелированности выражается коэффициентом корреляции. Ниже рассмотрим некоторые примеры таких коэффициентов, которые могут применяться при обработке результатов опросов.

При первом из приведенных выше вариантов заполнения матрицы результатов анкетирования (один вариант ответа – один столбец матрицы) легко решается задача вычисления коэффициента корреляции между двумя суждениями респондентов, отраженными в вопросах с бинарным выбором (да – нет, положительно – отрицательно, мужчина – женщина).

В этом случае можно применить непараметрический коэффициент корреляции Фехнера.

Подсчитывается количество совпадений и несовпадений знаков отклонений значений показателей от их среднего значения.

$$i = \frac{C - H}{C + H} \qquad (5.1)$$

С — число пар, у которых знаки отклонений значений от их средних совпадают.

Н — число пар, у которых знаки отклонений значений от их средних не совпадают.

В случае же если варианты ответа на вопрос носят ранговый характер — как, например, для вопросов № 1 и № 10, то для выявления взаимозависимости двух свойств респондента или его суждений также могут быть применены различные непараметрические коэффициенты корреляции.

Коэффициент ранговой корреляции Кендалла.

Применяется для выявления взаимосвязи между количественными или качественными показателями, если их можно ранжировать. Значения показателя X выставляют в порядке возрастания и присваивают им ранги. Ранжируют значения показателя *Y* и рассчитывают коэффициент корреляции Кендалла:

$$\tau = \frac{2S}{n(n-1)} \qquad (5.2)$$

где S = P - Q

$$\tau = \in [-1;1]$$

P — суммарное число наблюдений, следующих за текущими наблюдениями с большим значением рангов *Y*.

Q — суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов *Y*.

При этом равные ранги не учитываются, то есть данный метод не подошел бы к данным приведенной таблицы. В этом случае (если имеются объекты с одинаковыми рангами) в расчетах используется скорректированный коэффициент корреляции Кендалла:

$$\tau = \frac{S}{\sqrt{\left[\frac{n(n-1)}{2} - \mathrm{U}_x\right]\left[\frac{n(n-1)}{2} - \mathrm{U}_y\right]}}, \qquad (5.3)$$

$$\mathrm{U}_x = \frac{\Sigma t(t-1)}{2}$$

$$\mathrm{U}_y = \frac{\Sigma t(t-1)}{2}$$

t — число связанных рангов в ряду *X* и *Y* соответственно.

Коэффициент ранговой корреляции Спирмена.

Статистическая взаимозависимость двух показателей (свойств оцениваемых объектов) *X* и *Y* может характеризоваться на основе анализа получаемых пар . Каждому показателю *X* и *Y* присваивается ранг. Ранги значений *X* расположены в естественном порядке *i=1, 2, . . ., n*. Ранг *Y* записывается как *Ri* и соответствует рангу той пары *(X, Y)*, для которой ранг *X* равен *i*. На основе полученных рангов *Xi* и *Yi* рассчитываются их разности и вычисляется коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} \qquad (5.4)$$

Значение коэффициента меняется от −1 (последовательности рангов полностью противоположны) до +1 (последовательности

рангов полностью совпадают). Нулевое значение показывает, что признаки независимы.

Коэффициент множественной ранговой корреляции (конкордации).

$$W = \frac{12S}{m^2(n^3 - n)} \qquad (5.5)$$

$$S = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij}\right)^2 - \frac{\left(\Sigma_{i=1}^{n}\Sigma_{j=1}^{m} R_{ij}\right)^2}{n}$$

m — число групп, которые ранжируются.
n — число переменных.
Rij — ранг i-фактора у j-единицы.
Значимость:

$$\chi^2 = m(n-1) * W$$

$$\chi^2_{kp} = (\alpha; (n-1)(m-1))$$

$\chi^2 > \chi^2_{kp}$, то гипотеза об отсутствии связи отвергается.

В случае наличия связанных рангов:

$$W = \frac{12S}{m^2(n^3 - n) - m\Sigma_{j=1}^{m}(t_j^3 - t_j)} \qquad (5.6)$$

$$\chi^2 = \frac{12S}{mn(n+1) - \dfrac{m\Sigma_{j=1}^{m}(t_j^3 - t_j)}{n-1}} \qquad (5.7)$$

Однако если ответ на вопрос определяет некоторую категорию — как в случае с вопросом № 11 в примере, приведенном выше, — то такой подход не приемлем и необходимы другие способы выявления взаимозависимости двух показателей, оцениваемых на основе анкетирования.

5.3. Распределение «хи-квадрат» в задачах статистического анализа результатов анкетирования

Распределение «хи-квадрат» является одним из наиболее широко используемых в статистике для проверки статистических гипотез.

В 1900 г. Карл Пирсон предложил простой, универсальный и эффективный способ проверки согласия между предсказаниями модели и опытными данными. Предложенный им «хи-квадрат критерий» – это самый важный и наиболее часто используемый статистический критерий. Большинство задач, связанных с оценкой неизвестных параметров модели и проверки согласия модели и опытных данных, можно решить с его помощью.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Критерий χ^2(«хи-квадрат») используется для проверки гипотезы различных распределений. Его расчетная формула такова:

$$\chi^2 = \Sigma_{i=1}^{n} \frac{(m_i - m_i')^2}{m_i'}, \qquad (5.8)$$

где m и m’ — соответственно эмпирические и теоретические частоты

рассматриваемого распределения;

n — число степеней свободы.

Для проверки нам необходимо сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты.

При полном совпадении эмпирических частот с частотами, вычисленными или ожидаемыми, критерий χ^2 будет равен нулю. В противном случае значение критерия укажет на несоответствие вычисленных частот эмпирическим частотам ряда. Тогда необходимо оценить значимость критерия χ^2, который теоретически может изменяться от нуля до бесконечности. Это производится путем сравнения фактически полученной величины $\chi^2_{\text{ф}}$ с его критическим значением $\chi^2_{\text{кр}}$. Нулевая гипотеза, т. е. предположение, что расхождение между эмпирическими и теоретическими или ожидаемыми частотами носит случайный характер, опровергается, если $\chi^2_{\text{ф}}$ больше или равно $\chi^2_{\text{кр}}$ для принятого уровня значимости (a) и числа степеней свободы (n).

Распределение вероятных значений случайной величины χ^2 непрерывно и ассиметрично. Оно зависит от числа степеней свободы (n) и приближается к нормальному распределению по мере увеличения числа наблюдений. Поэтому применение критерия χ^2 к оценке дискретных распределений сопряжено с некоторыми погрешностями, которые сказываются на его величине, особенно на малочисленных выборках.

Точность определения критерия χ^2 в значительной степени зависит от точности расчета теоретических частот (m'_i), для получения разности между эмпирическими и вычисленными частотами следует использовать неокругленные теоретические частоты.

Критерий «хи-квадрат» позволяет сравнивать распределения частот вне зависимости от того, распределены они нормально или нет.

Под частотой понимается количество появлений какого-либо события. Обычно с частотой появления события имеют дело, когда переменные измерены в шкале наименований и другой их характеристики, кроме частоты появления, подобрать невозможно или затруднительно. Другими словами, когда переменная имеет качественные характеристики. Также многие исследователи, например, склонны переводить баллы теста в уровни (высокий, средний, низкий) и строить таблицы распределений баллов, чтобы узнать количество человек по этим уровням. Чтобы доказать, что в одном из уровней (в одной из категорий) количество человек действительно больше (меньше), также используется коэффициент χ^2.

Разберем самый простой пример.

Среди обучающихся было проведено анкетирование для выявления самооценки. Результаты были переведены в три уровня: высокий, средний, низкий. Частоты распределились следующим образом:

Высокий (В) — 27 чел.

Средний (С) — 12 чел.

Низкий (Н) — 11 чел.

Очевидно, что людей с высокой самооценкой большинство, однако это нужно доказать статистически. Для этого используем критерий «хи-квадрат».

Наша задача проверить, отличаются ли полученные эмпирические данные от теоретически равновероятных. Для этого необходимо найти теоретические частоты. В нашем случае теоретические частоты – это равновероятные частоты, которые находятся путем сложения всех частот и деления на количество категорий.

В нашем случае:

(В + С + Н)/3 = (27+12+11)/3 = 16,(6)

Построим таблицу:

	Эмпирические частоты (m_i)	Теоретические частоты (m'_i)	$\chi^2 = \sum_{i=1}^{n} \frac{(m_i - m'_i)^2}{m'_i}$
Высокий	27	16,67	6,41
Средний	12	16,67	1,31
Низкий	11	16,67	1,93

Далее найдем сумму последнего столбца:

$\chi^2 = 9,64$

Теперь нужно найти критическое значение критерия. Для этого нам понадобится число степеней свободы (n) и таблица распределения χ^2. В табличном процессоре Microsoft Excel для вычисления критического значения χ^2 по заданному числу степеней свободы и для заданной вероятности ошибки служит функция ХИ2ОБР.

Число степеней свободы рассчитывается по следующей формуле:

n = (R – 1) * (C – 1),

где R – количество строк в таблице, С – количество столбцов.

В нашем случае — только один столбец (имеются в виду исходные эмпирические частоты) и три строки (категории), поэтому формула изменяется, – исключаем столбцы.

n = (R – 1) = 3-1 = 2.

Для вероятности ошибки p≤0,05 (принятой в статистических исследованиях) и n = 2 критическое значение $\chi^2 = 5,99$.

Полученное эмпирическое значение больше критического – различия частот достоверны ($\chi^2 = 9,64$; p≤0,05), гипотеза о том, что расхождение между эмпирическими и теоретическими (ожидаемыми) частотами носит случайный характер, опровергается.

Как видим, расчет критерия очень прост и не занимает много времени. Практическая ценность критерия «хи-квадрат» огромна. Этот метод оказывается наиболее ценным при анализе ответов на вопросы анкет.

Разберем более сложный пример. Пусть целью анкетирования сотрудников ОВД являлось выявление зависимости между удовлетворенностью условиями прохождения службы и планируемым сроком нахождения в занимаемой должности. Результаты анкетирования занесены в нижеследующую матрицу.

	Возраст	Стаж	Подразделение	Образование	Удовлетворенность условиями службы	Планируемый период нахождения на должности
1	36–40	11 до 15	ГИБДД	высшее	частично	от 1 до 3 лет
2	26–30	до 5	ОООП	высшее	не устраивает	до 1 года
3	20–25	от 5–10	СЛ	высшее	частично	от 3 до 5 лет
4	25–30	от 11–15	БЭП	среднее	не устраивает	до 1 года
5	20–25	до 5	УР	высшее	не устраивает	до 1 года
6	41–45	свыше 20	ОООП	высшее	полностью	от 3 до 5 лет
7	36–40	16–20	Кадры	высшее	частично	до 1 года
8	41–45	16–20	УР	высшее	не устраивает	до 1 года

Для достижения поставленной цели построим сводную таблицу распределения частот ответов на оба изучаемых вопроса. Данную таблицу можно интерпретировать следующим образом: среди тех, кто планирует находиться на должности от 3 до 5 лет, нет ни одного человека, кого условия службы не устраивают; трое тех, кого устраивают полностью; 5 тех, кого – частично, и т. д.

		Планируемый период нахождения на должности			
		от 3 до 5 лет	от 1 до 3 лет	до 1 года	**Общий итог**
Удовл. условиями	не устраивает	0	0	6	**6**
	полностью	3	0	5	**8**
	частично	5	2	11	**18**
	Общий итог	**8**	**2**	**22**	**32**

Это матрица эмпирических частот .

Построим матрицу теоретических (ожидаемых) частот , тех частот, которые наблюдались бы при полной независимости ответов респондентов на два предложенных вопроса.

		Планируемый период нахождения на должности			
		от 3 до 5 лет	от 1 до 3 лет	до 1 года	**Общий итог**
Удовл. условиями	не устраивает	2,81	0,38	2,81	**6,00**
	полностью	3,75	0,50	3,75	**8,00**
	частично	8,44	1,13	8,44	**18,00**
	Общий итог	**15,00**	**2,00**	**15,00**	**32,00**

Рассчитаем фактическое значение критерия «хи-квадрат»:

$$\chi^2_{\Phi} = \sum_{i=1}^{n} \frac{(m_i - m_i')^2}{m_i'}$$

Для этого заполним матрицу $\frac{(m_i - m_i^{'})^2}{m_i^{'}}$

		Планируемый период нахождения на должности		
		от 3 до 5 лет	от 1 до 3 лет	до 1 года
Удовл. условиями	не устраивает	2,81	0,38	3,62
	полностью	2,82	0,50	2,02
	частично	0,02	0,68	0,02

Рассчитаем: $\chi^2_{\Phi} = \sum_{i=1}^{n} \frac{(m_i - m_i^{'})^2}{m_i^{'}} \approx 12,87$

Найдем критическое значение . Для этого зададим вероятность ошибки равную 0,05 и найдем число степеней свободы n = (R – 1) * (C – 1) = (3-1)*(3-1) = 4.

$$\chi^2_{\text{КР}} \approx 9,49$$

Полученное фактическое значение больше критического, а значит нулевая гипотеза о независимости двух показателей отвергается.

Вывод: планируемый период нахождения сотрудника на должности связан с удовлетворенностью условиями службы.

Прикладные социологические исследования в сфере правоохранительной деятельности нацелены на решение задач совершенствования системы управления органами внутренних дел, создания стабильного правоприменительного поля, оценки и повышения качества результатов деятельности ОВД.

Для проведения прикладного социологического исследования необходимо разработать его программу, в которой выделить имею-

щуюся проблемную ситуацию, сформулировать проблему, установить цель исследования как модель ожидаемого решения проблемы, сформулировать задачи, которые необходимо решить для достижения цели. Нужно четко обозначить объект исследования, ограничить его временными рамками, определить количественно, провести его системный факторный анализ. Уточнить предмет исследования, выдвинуть гипотезы – научные предположения, задающие направление всему исследованию.

По результатам проведенного анализа результатов опроса составляется отчет, в котором, в случае прикладного социологического исследования, важнейшим разделом являются выводы и практические рекомендации с обоснованием социальных и экономических последствий внедрения последних.

Владение исследователя методами обработки анкетных данных и математическим аппаратом установления взаимосвязей между различными характеристиками респондентов и их суждениями по той или иной теме, оценкой различных сфер правоохранительной деятельности и криминальной ситуации является залогом качества выводов по проведенному анкетному опросу.

Контрольные вопросы:

1. Способы формирования матрицы результатов анкетного опроса.
2. Установление взаимосвязи между ответами респондентов на два вопроса с бинарным выбором.
3. Непараметрические коэффициенты корреляции при обработке анкетных опросов, возможности применения и ограничения.
4. Способы определения взаимозависимости двух вопросов, предполагающих выставление респондентами ранговых оценок.
5. Применение критерия «хи-квадрат» в обработке результатов анкетных опросов.

Заключение

Следует отметить тот факт, что применение математических методов в научных исследованиях представляет собой комплексную многоаспектную задачу. На различных этапах ее решения могут применяться различные теоретические подходы и различные компьютерные технологии их практической реализации. Так, для отбора независимых показателей, характеризующих объект исследования, для изучения их взаимообусловленности и взаимовлияния имеет смысл применять аппарат корреляционно-регрессионного анализа данных.

Для описания объекта исследования, аргументации актуальности проводимого научного изыскания служат методы расчета описательных статистик – мер центральной тенденции и вариативности выборок и генеральных совокупностей.

Для прогнозирования и изучения логики изменения во времени социальных и правовых явлений существуют модели динамических рядов.

Важной составляющей эмпирической базы гуманитарных исследований выступают результаты анкетных опросов. Для анализа данных, полученных в ходе анкетирований, существуют методы ранговой корреляции и конкордации, а также критерии сопряженности ответов на вопросы, например – критерий «хи-квадрат» Пирсона.

Литература

Айвазян С. А., Бежаев З. И., Староверов О.В. Классификация многомерных наблюдений. М.: Статистика, 1974.

Айзерман М. А., Браверман Э. М., Розеноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

Андерсон Т. Статистический анализ временных рядов. М.: Мир, 1976.

Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. М.: Мир, 1982.

Ашимов А. А., Бурков В.Н. Согласованное управление активными производственными системами. М.: Наука, 1986.

Беляева Л. И., Мартыненко Н. Э., Цепелев В. Ф., Смольянинов Е. С., Трунцевский Ю. В., Черняков С. А., Милехин В. А., Мелехин О.Ю. Уголовная политика и ее реализация органами внутренних дел: учебник / М.: Академия управления МВД России, 2014. Т. 1. 168 с.

Бендат Дж., Пирсол А. Прикладной анализ случайных данных. М.: Мир, 1989.

Васильчикова Н. В., Кухарук В.В. Криминология: конспект лекций, М.: Юрайт, 2010.

Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. М.: Финансы и статистика, 1987.

Гаврилов О.А. Математические методы и модели в социально-правовом исследовании. М.: Наука,1980.

Горошко И.В. Математическое моделирование в управлении ОВД. М.: Академия управления МВД России, 2000.

Горошко И. В., Сичкарук А. В., Флока А.Б. Методы и модели анализа данных в правоохранительной деятельности. М.: Ас-Траст, 2007.

Горошко И. В., Бондаренко Ю.В. Механизмы согласования показателей социально-экономического развития региона и роль органов внутренних дел в их реализации. М.: Академия управления МВД России, 2015. 128 с.

Горошко И. В., Горошко Э. Г., Бондаренко Ю.В. Проблемы управления социально-экономической системой регионального уровня // В сборнике: Стратегическое планирование и развитие предприятий. Материалы симпозиума. Российская академия наук, Федеральное государственное бюджетное учреждение науки «Центральный экономико-математический институт». 2016. С. 34–37.

Горшков М. К., Андреев А. Л., Бараш Р. Э., Бызов Л. Г., Дробижева Л. М., Каравай А. В., Латова Н. В., Латов Ю. В., Лежнина Ю. П., Мареева С. В., Мчедлова М. М., Петухов В. В., Петухов Р. В., Седова Н. Н., Трофимова И. Н., Тихонова Н.Е. Российское общество и вызовы времени: монография. М.: Весь Мир, 2017. Книга 5. 427 с.

Демиденко Е.З. Линейная и нелинейная регрессия. М.: Финансы и статистика, 1981.

Джонстон Дж. Эконометрические методы. М.: Статистика, 1980.

Елисеева И. И., Рукавишников В.О. Группировка, корреляция, распознавание образов. М.: Статистика, 1977.

Иванилов Ю. П., Лотов А.В. Математические модели в экономике. М.: Наука, 1979.

Иванов Н.Н. Оценка эффективности управленческой деятельности ОВД. М.: Академия МВД России, 1993.

Кендалл М. Многомерный статистический анализ и временные ряды. М.: Наука, 1976.

Кокс Д.Р., Оукс Д. Анализ данных типа времени жизни. М.: Финансы и статистика, 1988.

Крамер Г. Математические методы статистики. М.: Мир, 1975.

Левин М. И., Макаров В. Л., Рубинов А.М. Математические модели экономического взаимодействия. М.: Наука, 1993.

Леман Э. Проверка статистических гипотез. М.: Наука, 1964.

Мартыненко Н.Э. Уголовно-правовое обеспечение безопасности личности в Российской Федерации // Публичное и частное право. 2010. № 2. С. 75 — 79.

Рао С.Р. Линейные статистические методы и их применение. М.: Наука, 1968.

Ситковский А. Л., Латов Ю.В. Новая криминальная реальность как результат новых социально-экономических условий развития России // Российский журнал правовых исследований. 2016. № 3 (8). С. 168–174.

Смоляк С. А., Титаренко Б.П. Устойчивые методы оценивания. М.: Статистика, 1980.

Торопов Б. А. О некоторых подходах к многокритериальному оцениванию деятельности полиции // Труды Академии управления МВД России. 2015. № 2 (34). С. 42–45.

Торопов Б. А., Апульцин В.А. Технологии многокритериального оценивания результатов деятельности территориальных органов МВД России на региональном уровне. М.: Академия управления МВД России, 2016. 112 с.

Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981.

Хан Г., Шапиро С. Статистические модели в инженерных задачах. М.: Статистика, 1980.

Харман Г. Современный факторный анализ. М.: Наука, 1972.

Хьюбер П. Робастность в статистике. М.: Мир, 1984.

ДЛЯ ЗАМЕТОК

Учебное издание

Торопов Борис Андреевич
Болтачев Эльдар Филаридович
Баранов Владимир Владимирович

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ СОЦИАЛЬНЫХ СИСТЕМ

Учебное пособие

Редактор *Д. В. Алентьев*
Верстка: *А. А. Мельникова*

Подписано в печать 10.09.2020. Формат 60 × 84 $^{1}/_{16}$.
Усл. печ. л. 4,65. Уч.-изд. л. 3,12. Тираж 61 экз. Заказ № __

Отделение полиграфической и оперативной печати РИО
Академии управления МВД России
125993, Москва, ул. Зои и Александра Космодемьянских, д. 8

ISBN 978-5-907187-30-6