АКАДЕМИЯ УПРАВЛЕНИЯ МВД РОССИИ

Горошко И. В., Торопов Б. А., Гонов Ш. Х.

Математические методы исследования социальных систем

Курс лекций

Москва • 2019

УДК 52-17:316.3
ББК 22.1в6
Г70

Одобрено редакционно-издательским советом Академии управления МВД России

Рецензенты: *И.А. Кубасов,* начальник Вычислительного центра ФКУ «ГИАЦ МВД России», доктор технических наук, доцент; *П.Г. Василенко*, начальник Отдела МВД России по району Якиманка г. Москвы.

Горошко И. В.

Г70 **Математические** методы исследования социальных систем : курс лекций / И. В. Горошко, Б. А. Торопов, Ш. Х. Гонов. – М. : Академия управления МВД России, 2019. – 80 с.

ISBN 978-5-906942-95-1

В настоящее время математические методы исследования социальных систем становятся основным инструментарием поддержки принятия управленческих решений любого уровня. Лицо, принимающее такие решения, обязано уметь анализировать и оценивать преступность как сложное социально-правовое явление, знать теорию управления, методы моделирования и прогнозирования. В научных исследованиях, особенно в гуманитарных науках, применение математического моделирования выступает как вспомогательный инструмент. Отсюда вытекает необходимость формирования знаний, умений, навыков при проверке гипотез и посылок по широкому кругу вопросов деятельности органов внутренних дел. Курс лекций призван помочь адъюнктам и аспирантам упорядочить систему методов и моделей, применяемых в информационно-аналитической работе органов внутренних дел при решении оперативно-служебных задач.

Подготовленный материал предназначен для адъюнктов и аспирантов, а также может быть полезен слушателям и курсантам учебных заведений МВД России и должностным лицам руководящего состава органов внутренних дел Российской Федерации.

УДК 52-17:316.3
ББК 22.1в6

ISBN 978-5-906942-95-1

Введение

Развитие любой социальной системы, в том числе и органов внутренних дел, происходит в тесной взаимосвязи с окружающей ее средой. Квалифицированному специалисту в сфере управления ОВД необходимо иметь не только четкое представление о внутренних закономерностях функционирования системы, но и возможных влияниях на нее со стороны внешней среды. Такое представление невозможно получить, не прибегая к математическим методам исследования социальных систем. Основой этих методов являются фундаментальные законы теории вероятностей и математической статистики, положения теории игр, марковских процессов, теории автоматов и т. д.

В настоящее время существует большое количество математических моделей различных реальных социальных систем. Эти модели создавались для решения практических задач применительно именно к этим системам и описывают их с определенной точностью. В то же время как сами модели, так и созданные для их реализации информационные технологии представляют собой достаточно разнообразное сочетание различных математических подходов и методов. Это обстоятельство затрудняет типизацию задач анализа социальных систем и оказывает негативное влияние на расходование сил, средств и времени при проведении исследований социальных систем.

Одной из задач курса лекций является проведение систематизации существующих математических подходов и методов к исследованию социальных систем, а также рассмотрение надежных (с точки зрения практики) и малозатратных (с точки зрения времени), и, кроме того, достаточно простых математических моделей, используемых для описания социальных систем.

В общем виде технология описания социальных систем заключается в построении на основании имеющихся данных математических моделей, описывающих поведение исследуемого объекта, процесса или явления, изучении этих моделей с целью их уточнения для придания им большей адекватности описания. Разработанные математические модели используются

в дальнейшем при принятии управленческих решений в практической деятельности.

Актуальность настоящего курса лекций обусловлена необходимостью систематизации знаний и подходов, касающихся современных аспектов построения и модификации математических моделей, используемых для описания социальных систем на основе внедрения компьютерных технологий.

В процессе подготовки квалифицированных кадров органов внутренних дел курс лекций будет способствовать реализации стратегических направлений МВД России, направленных на совершенствование управления органами внутренних дел.

Курс лекций состоит из 6 тем, каждая из которых входит в одноименную учебную дисциплину, преподаваемую на факультете подготовки научных и научно-педагогических кадров Академии управления МВД России.

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

Применение моделирования в научных исследованиях берет свое начало с древнейших времен и, по мере развития науки и техники, постепенно внедрялось во все области научных знаний: астрономию, механику, физику, химию, биологию, а также в общественно-гуманитарные науки.

Методология моделирования, формировавшаяся отдельными науками в XX в., стала приобретать иной смысл. Моделирование стало восприниматься как универсальный метод познания, чему способствовало бурное развитие научно-технического прогресса.

Во второй половине XX в. моделирование стало применяться в соответствии с принципами системного подхода, наиболее значимыми среди которых являлись целостность и обратная связь.

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 науками, вкладывающими в ее содержание свой особый смысл. Отсутствовала единая система понятий и единая терминология. Однако постепенно стала осознаваться роль моделирования как универсального метода научного познания.

Построение моделей базируется на принципах *системного подхода,* при этом главными являются принципы целостности и обратной связи, роль которых в процедуре моделирования является особенно значимой.

Если целостность модели отображает способность воспроизводить механизм функционирования объекта, то при помощи обратной связи эта целостность поддерживается в процессе взаимодействия объекта с факторами окружающей среды.

В социально-экономических исследованиях модель обращена на отображение социальных систем и процессов, благодаря чему можно получить не только полное и целостное представление о лежащих в их основе механизмах, но и позволит заинтересованным сторонам управлять этими системами и процессами, целенаправленно влиять на режим их функционирования и развития.

Вопрос 1. Моделирование как метод социальных исследований

Модель (от лат. *modus, modulus* – «мера», «образ», «способ») – это такой материальный или мысленно представляемый объект, который в процессе исследования замещает объект-оригинал так, что его непосредственное изучение дает новые знания об объекте-оригинале. Модель – это упрощенное представление объекта, используемое для имитации возможных состояний этого объекта. Известный российский ученый Н. Моисеев писал: «Модель можно рассматривать как специальную форму кодирования информации... С помощью моделей из старых знаний могут возникать новые знания. И потому одной из важнейших задач науки является не только систематизация, кодирование известной информации и построение на этой основе системы моделей (теорий), но и создание методов теоретического анализа, т. е. раскодирования той информации, которая потенциально содержится в моделях и приводит к получению нового знания».

Модель в научно-исследовательских программах выполняет три основные функции: *прогностическую, имитационную и проективную.*

Прогностическая функция основана на свойстве модели предсказывать вероятные изменения свойств и параметров исследуемых процессов и явлений с учетом действия различных факторов среды.

Имитационная функция концентрирует внимание исследователя исключительно на искусственном воспроизводстве естественных свойств исследуемого объекта, что является крайне важным при сложном характере объекта и неопределенности проблемной ситуации.

Проективная функция предполагает исследование возможности интродукции в исследуемый объект, явление или процесс предварительно заданных свойств, чья реализация позволит достичь позитивных результатов.

Конструируя модели, исследователь реализует процедуру моделирования. Под *моделированием* понимается процесс построения, изучения и применения моделей. Понятие моделирования тесно связано с такими категориями, как абстракция, аналогия, гипотеза и др.

Процесс моделирования обязательно включает и построение абстракций, и умозаключения по аналогии, и конструирование научных гипотез.

Моделирование является конечным этапом системного подхода. Здесь системный подход получает свое практическое выражение в способности воспроизведения исследуемого объекта во всей совокупности выявленных в ходе анализа связей и отношений.

Главное предназначение процедуры моделирования заключается в способности опосредованного познания с помощью объектов-заместителей. При этом модель выступает как своеобразный инструмент познания, который исследователь ставит между собой и объектом, а также с помощью которого изучает интересующий его объект. Именно эта способность процедуры моделирования определяет специфические формы использования абстракций, аналогий, гипотез, других категорий и методов познания.

Процедура моделирования предполагает проведение предварительной реконструкции объекта исследования, осуществляемой при помощи причинного, корреляционного или факторного анализа.

В ходе причинного анализа выявляются экзогенные и эндогенные параметры модели, вскрываются функциональные зависимости между переменными.

Корреляционный анализ позволяет установить соответствие между составляющими информационное поле динамическими рядами количественных показателей.

Факторный анализ помогает выявить скрытые факторы, способные оказывать влияние на систему взаимосвязанных показателей, локализующих тип и характер описываемых моделью процессов.

Вопрос 2. Классификация моделей

Выделяется несколько оснований для классификации моделей.

По носителю информации модели делятся на *абстрактные* и *материальные*. Абстрактные модели, в свою очередь, могут быть динамическими и статическими. Динамические модели бывают линейными и нелинейными. Среди нелинейных моделей выделяют: *неустойчивые* и *устойчивые*.

Широкую популярность в последнее время приобретает использование исследователями *стохастических* (вероятностных) моделей, противопоставляемых *детерминированным*.

Детерминированной является модель с фиксированным перечнем входных параметров, определяющих свойство и динамику моделируемого объекта. В детерминированных моделях факторы, оказывающие влияние на развитие исследуемой ситуации, четко определены, а их значения легко вычислить.

Образцом детерминированной модели является мячик, иллюстрирующий свойства шара. За исключением свойств упругости данная модель один в один способна имитировать представленный аналог.

Специфической чертой *стохастических* моделей является наличие элемента неопределенности, заключающегося в вероятностном распределении значений факторов и параметров, определяющих развитие ситуации, что предполагает обязательное наличие в качестве одного из параметров модели показателя вероятности. Особенно эффективным представляется использование вероятностных моделей при прогнозировании многофакторных социальных процессов, развивающихся с разной степенью интенсивности.

Так, к примеру, при помощи стохастической модели можно рассчитать перспективы роста города с учетом сложившихся в нем экономических и социальных особенностей. Сначала выделяются блоки модели, составляющие информационную базу для решения поставленной задачи. К ним можно отнести: характеристики земельного фонда (цены на землю, процент незанятых участков), характеристики транспортной сети (размеры дорожной сети, время, затрачиваемое на поездку в центр), характеристику застройки (типы строений, наличие трущоб, благоустроенных кварталов), характеристику сферы обслуживания и благоустройства территории (социально-бытовая инфраструктура и пр.).

Затем территория города разбивается на отдельные зоны, по которым вычисляется вероятность застройки каждой из них за определенный период времени. Полученная модель позволит руководителям городов эффективно производить ценовое зонирование городского пространства, определять оптимальный размер арендной платы на объекты муниципальной собственности, прогнозировать спрос на объекты недвижимости.

По характеру отношения модели к среде выделяют *закрытые* и *открытые* модели.

В *закрытой* модели изменения значений переменных во времени определяются внутренним взаимодействием самих переменных. Закрытая модель может проиллюстрировать поведение системы без ввода в нее внешних переменных.

Характерным свойством открытой модели является обусловленность ее свойств влиянием внешних факторов, составляющих содержание среды объекта моделирования.

В зависимости от цели, закладываемой в содержание модели, выделяют *аналитические* и *имитационные* модели.

Аналитическая модель ориентирована на объяснение связей и отношений в структуре исследуемого объекта на базе его детальной структуризации. Аналитическая модель по своей структуре

является замкнутой когнитивной системой, составляющей об объекте целостное представление.

Чаще всего в социальных исследованиях аналитическая модель представлена как *трендовая модель,* назначение которой состоит в установлении тенденции исследуемого процесса и в прогнозе его развития.

Создание целостного представления об объекте на основе определения характерных для него тенденций развития является одним из способов диагностики свойств этого объекта, факторов воздействия на него. Информация об этих свойствах и факторах служит условием для прогнозирования социальных событий, сопровождающих процесс функционирования исследуемого объекта.

Широкую популярность в истории социальной мысли получила циклическая модель социального развития, предполагающая периодическое повторение определенных фаз развития. Впервые описанная Дж. Вико циклическая модель социального развития была затем трансформирована в различных социальных теориях Нового времени, приобретя особый смысл в философских теориях Г. Гегеля и К. Маркса.

Характерной чертой циклической модели является качественный способ ее теоретического представления. Наиболее распространенной формой циклического представления социальных изменений стала *модель спирали,* в основу механизма которой положен гегелевский закон «отрицания отрицания», выражающий способность исторических событий «повторяться» на качественно новом уровне развития, придавая историческому процессу целеполагающий смысл.

В отличие от аналитической, *имитационная модель* предназначена для получения информации об исследуемом объекте с точки зрения выработки управленческих решений. Для этого с помощью имитационных моделей формируется информационная база о свойствах и структуре объекта с воспроизводством лежащих в их основе связей и отношений.

Полученные данные обобщаются, группируются по блокам с выделением в них ряда контрольных показателей. Значение показателей варьируется, производится оценка возможных промежуточных и конечных решений, после чего определяется последовательность принятия оптимальных решений.

В основе этого метода – теория вычислительных систем, статистика, теория вероятностей, математика. Имитационные модели не накладывают ограничений на исходные данные, выражая собственные свойства и признаки непосредственно на их базе.

Все имитационные модели построены по типу «черного ящика», то есть сама система, ее элементы и структура представлены в виде такого «ящика»: есть какой-то вход, который описывается экзогенными переменными (возникают вне системы под воздействием внешних причин), и выход (описывается выходными переменными), который характеризует результат действия системы.

В имитационном исследовании большое значение имеет этап оценки модели, который включает в себя следующие шаги.

1. Верификация модели (модель ведет себя так, как это было задумано исследователем).

2. Оценка адекватности (проверка соответствия модели реальной системе).

3. Проблемный анализ (формирование статистически значимых выводов на основе данных, полученных в результате экспериментов с моделью).

В соответствии с критерием подвижности среди имитационных моделей выделяют *статические и динамические модели. Статические* имитационные модели нацелены на выявление структуры объекта. Такой способ моделирования особенно эффективен при недостатке информации о содержании обследуемого объекта, его характерных признаках. *Динамические модели* позволяют делать заключения о динамических свойствах объекта, не зависящих от начальных условий.

С точки зрения средств выражения процедура моделирования предполагает использование следующих форм представления объектов.

Словесное описание. Распространено на первых этапах моделирования и предполагает вербальный способ выражения данных.

Графическое представление. Иллюстрирует состояние и динамику основных показателей в виде кривых, чертежей, номограмм.

Блок-схемы, матрицы решений – схематически выраженные последовательности решений проблемных ситуаций.

Математическое описание – переложение логических схем в символическую форму.

Выбор каждой из форм связан с прохождением той или иной стадии в процессе моделирования и целями, стоящими перед исследователем.

По средствам выражения модели бывают *предметные, знаковые* и *математические*. *Предметные* модели отображают содержательные свойства исследуемого объекта, сконцентрированные на пространственно и функционально ограниченном объеме его аналога. *Знаковые* модели ориентированы на отображение структуры объекта, обозначенной системой значений (символов), выражающих сущностные

свойства и признаки данного объекта в условиях изменяющейся среды. *Математические* модели базируются на использовании логико-математических методов, с помощью которых раскрываются закономерности в динамике изменений исследуемого объекта.

С точки зрения назначения моделей, целей, на достижение которых ориентирован их выбор, можно выделить следующие *типы* моделей.

Модели *принятия решений* – модели, имитирующие типовой способ подготовки и реализации управленческого решения.

Модели *компромиссов* – это такие модели, которые описывают способы взвешивания и оценки замен в средствах и целях.

Одно- и *многоцелевые* модели – модели, предназначенные для осуществления выбора между сложными вариантами.

Оптимизационные модели – модели, ориентированные на нахождение локальных оптимумов.

Оценочные модели – модели, служащие способом определения отношения к состоянию исследуемой системы.

Познавательные модели – модели, описывающие способ достижения достоверности в рамках данного метода рассуждений.

Диагностические модели – модели, призванные организовать оптимальный путь нормализации работы системы в случае нарушения ее нормальной работы.

Между понятиями «модель» и «теория» имеются определенные отличия, вызванные различиями в целях исследователя. Если предназначением теории является объяснение процессов, то модель призвана обеспечить управление этими процессами. При этом характер используемых теорий и моделей определяется особенностями лежащих в их основе методологических подходов, рассмотренных нами в первой главе этого издания.

С точки зрения *эволюционного (исторического)* подхода объектом моделирования выступает сеть причинно-следственных связей, событий, выстраиваемых в соответствии с их местом во временном континууме. Знание о причинно-следственных связях между элементами исследуемого процесса позволяет выявить механизм формирования результатов такого процесса, факторы обусловленности его промежуточных этапов.

Основу эволюционного подхода в моделировании составляют *экстраполяционные модели,* используемые в прогнозировании, суть которых заключается в переносе (экстраполяции) прошлых тенденций в будущее.

Ограниченность экстраполяционных моделей определяется предположением о необратимости исторического времени, в соот-

ветствии с которым на каждом временном этапе складывается уникальное сочетание факторов среды, не позволяющее воспроизвести результат, порожденный прошлыми состояниями.

С точки зрения *функционального подхода* объектом моделирования выступает уже функциональная структура исследуемого процесса, включающая в себя статические и динамические свойства объекта исследования, обеспечивающие поддержку определенного качества этого процесса и составляющих его элементов во взаимодействии друг с другом.

Важнейшим признаком функциональной модели является ее равновесный характер, обусловливающий тесную привязку свойств исследуемого процесса и условий его воспроизводства.

Функциональные модели призваны обслуживать конкретную исследовательскую цель, вовлекая в свое содержание те элементы, которые необходимы для достижения этой цели. Перечень функциональных моделей включает в себя и *оценочные,* и *описательные,* и *прогностические,* и *имитационные модели*.

Отличительной особенностью функциональной модели является ее ориентация на функциональную полноту и ролевую организацию входящих в нее элементов, а не выстраивание между последними причинно-следственных зависимостей, как это бывает при эволюционном подходе.

Институциональный подход не предполагает жесткой нормативной структуры и служит только способом упорядочивания используемой информации в соответствии с требованиями текущей целесообразности.

В рамках институционального подхода может быть обнаружено множество точек равновесия в исследуемом объекте, образованных вокруг исходных параметров, составляющих основу когнитивной актуализации объекта.

Главной целью институционального подхода в моделировании является разработка сценариев, с помощью которых уясняется весь спектр проблем, подлежащих решению. Ядро институционального подхода составляют *модели принятия решения* и *компромиссов,* предполагающие наличие нескольких вариантов оптимальных решений.

Всякая модель должна соответствовать необходимым для ее существования свойствам, иллюстрирующим эвристические возможности процедуры моделирования. Что касается свойств модели, то к таковым относятся: *абстрактность* – модель должна обладать некими элементами идеальных символов; *полнота* – модель должна содержать максимально возможное количество

релевантных элементов; *адекватность* – модель должна быть адекватной исследуемому реальному объекту; *робастность* – проектирование в моделируемой системе способности реагировать и исправлять возникающие в ходе моделирования ошибки; *динамизм* – способность в случае необходимости перестройки модели на другой уровень; *релевантность элементов* – соответствие привлекаемых параметров целям и характеру моделируемого объекта.

Многофункциональность процедуры моделирования сопутствует выделению отдельных типов моделей, с помощью которых реализуются различные принципы моделирования.

Вопрос 3. Принципы моделирования

При проектировании моделей необходимо придерживаться некоторых принципов, соблюдение которых позволит получить адекватное и точное отображение исследуемого события или процесса. К их числу 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.

Непосредственно конструированию модели предшествует проведение следующих мероприятий.

1. Формулировка основных целей и задач исследования.

2. Определение границ системы, отделение ее от внешней среды (посредством отделения эндогенных факторов от экзогенных).

3. Составление списка элементов системы (подсистем факторов, переменных и т. д.).

4. Обоснование целостности системы.

5. Анализ взаимосвязей элементов системы.

6. Построение структуры системы.

7. Установление функций системы и ее подсистем.

8. Согласование целей системы и ее подсистем (этот процесс называется субоптимизацией).

9. Уточнение границ системы и каждой подсистемы.

10. Анализ явлений эмерджентности.

11. Объединение людей разных профессий на срок решения проблемы.

Заключение

Моделирование представляет собой инструмент проектирования социально-экономических и политических систем с заданными свойствами. Социальные системы представляют собой определенным образом организованные совокупности социальных элементов, располагающих целями и средствами для их достижения. Социальные системы являются источником происходящих в обществе процессов, служащих предметом научного рассмотрения. Основу социальных систем составляют общественные стереотипы, выраженные в виде типовых ориентаций представителей отдельных социальных групп, связанных общностью интересов и целей. Общественные стереотипы служат способом типизации социальных процессов и, как следствие этого, выражением системной целостности объединенных посредством проявления этого стереотипа сообществ социальных субъектов. Моделирование социально-экономических и политических систем является одним из этапов в реализации научно-исследовательских программ. Способность программы создавать продукт с заданными свойствами на основе анализа процессов и систем выступает главным свидетельством полезности и методологической состоятельности такой программы.

Главное предназначение процедуры моделирования заключается в способности опосредованного познания с помощью объектов-заместителей. При этом модель выступает как своеобразный инструмент познания, который исследователь ставит

между собой и объектом и с помощью которого изучает интересующий его объект. Именно эта способность процедуры моделирования определяет специфические формы использования абстракций, аналогий, гипотез, других категорий и методов познания.

Контрольные вопросы

1. Предпосылки построения моделей.
2. Примеры абстрактных и материальных моделей.
3. Определение стохастической модели.
4. Основные функции моделирования.
5. Классификация моделей с точки зрения их назначения.
6. Основные принципы моделирования.
7. Этапы моделирования.

Лекция 2. Основы математической статистики. Расчет описательных статистик

Математическая (или теоретическая) статистика опирается на методы и понятия теории вероятностей, но решает в каком-то смысле обратные задачи. В теории вероятностей рассматриваются случайные величины с заданным распределением или случайные эксперименты, свойства которых целиком известны. Предмет теории вероятностей – свойства и взаимосвязи этих величин (распределений). Но часто эксперимент представляет собой черный ящик, выдающий лишь некие результаты, по которым требуется сделать вывод о свойствах самого эксперимента. Наблюдатель имеет набор числовых (или их можно сделать числовыми) результатов, полученных повторением одного и того же случайного эксперимента в одинаковых условиях.

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

Например: Если наблюдается одновременно проявление двух (или более) признаков, т. е. имеем набор значений нескольких случайных величин, то что можно сказать об их зависимости? Есть она или нет? А если есть, то какова эта зависимость? Либо: Если имеется набор показателей, характеризующих объекты, формально принадлежащие к одному классу, то можно ли выделить среди них наборы объектов, похожих друг на друга (провести кластеризацию)?

Если имеется динамический ряд, характеризующий изменение во времени одного показателя, то можно ли рассчитать прогнозное значение на следующий период времени?

Это лишь некоторые типичные вопросы, при поиске ответов на которые необходимы методы математической статистики.

Вопрос 1. Основные понятия математической статистики

Статистика (от итал. – *«statio»*) представляет собой как вид практической деятельности, так и отрасль знаний[1].

Статистическое исследование включает в себя четыре основных этапа[2].

---

[1] *Илышев А. М.* Общая теория статистики: учеб. пособие / А. М. Илышев, О. М. Шубат. М. 2013. 432 с.

[2] *Минашкин В. Г., Шмойлова Р. А., Садовникова Н. А., Моисейкина Л. Г., Рыбакова Е. С.* Теория статистики: учебно-методический комплекс. М. 2008. 296 с.

1. *Сбор данных*. На данном этапе осуществляется проверка его полноты и достоверности с использованием методов статистического наблюдения. Окончательные результаты во многом зависят от качества исходных статистических данных.

2. *Предварительная обработка данных*. На данном этапе осуществляется подсчет групповых и общих итогов, расчет некоторых относительных показателей. В результате применения метода группировок осуществляется переход от больших массивов неструктурированных статистических данных к более компактным и удобным для анализа агрегированным данным в таблицах.

3. *Расчет и интерпретация обобщающих статистических показателей*. На данном этапе рассчитываются структурно-динамические показатели, средние показатели, вариации и взаимосвязи изучаемых процессов (явлений). Полученные результаты подвергаются более глубокому анализу.

4. *Моделирование*. На последнем этапе осуществляется моделирование взаимосвязей между социально-экономическими процессами и явлениями, на основе корреляционно-регрессионного анализа строятся уравнения (пространственные, временные, пространственно-временные), отражающие основные тенденции динамики изучаемых показателей.

В настоящее время в системе МВД России сбор, обработка и обобщение статистической информации осуществляются по линии информационных центров. На федеральном уровне центральным органом этой системы является Главный информационно-аналитический центр МВД России (ФКУ «ГИАЦ МВД России»). В субъектах Российской Федерации – республиках, краях, областях и районах – информационные центры Министерств внутренних дел по республикам, управлений, главных управлений внутренних дел по краям, областям Российской Федерации. Непосредственная обработка поступающих из регионов статистических данных осуществляется в ФКУ «ГИАЦ МВД России», который обладает необходимыми мощными информационно-вычислительными ресурсами.

На государственном уровне методологическое руководство работами по сбору, обработке и анализу статистических данных возложено на Федеральную службу государственной статистики. С основными социально-экономическими показателями Российской Федерации можно ознакомиться на сайте Росстата (http://www.gks.ru). В системе ИМТС для этих целей предназначен сайт Центра статистической информации ФКУ «ГИАЦ МВД России» (http://10.5.0.16/csi).

Прежде чем приступить к изучению основных статистических показателей, необходимо познакомиться с основными категориями статистики.

Математическая статистика – это раздел прикладной математики, в котором рассматриваются методы отыскания законов и характеристик случайных величин по результатам наблюдений и экспериментов.

1. Создание методов сбора и группировки обрабатываемого статистического материала, полученного в результате наблюдений за случайными процессами.

2. Разработка методов анализа полученных статистических данных.

3. Получение выводов по данным наблюдений.

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

Математическая статистика опирается на теорию вероятностей и в свою очередь служит основой для разработки методов обработки и анализа статистических результатов в конкретных областях человеческой деятельности.

Основными понятиями математической статистики являются генеральная совокупность и выборка.

Определение. *Генеральная совокупность* – это совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.

Генеральная совокупность может быть *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность составляющих ее объектов.

Не следует смешивать понятие генеральной совокупности с реально существующими совокупностями. Например, на склад поступила продукция некоторого цеха за месяц, что является реально существующей совокупностью, которую нельзя назвать генеральной, поскольку выпуск продукции можно мысленно продолжить сколь угодно долго.

Определение. *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.

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.

Число N объектов генеральной совокупности и число n объектов выборки называют объемами генеральной и выборочной совокупностей соответственно. При этом предполагают, что $N>>n$ (значительно больше).

Полученные различными способами отбора данные образуют выборку, обычно это множество чисел, расположенных в беспорядке. По такой выборке трудно выявить какую-либо закономерность их изменения (*варьирования*).

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 1. 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_i, ...

Изменение значения варианта называется *варьированием*.

Определение. 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 $p_i^* = \left(\frac{n_i}{n}\right)$ 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.

Определение. *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. 1.

Таблица 1

x_i	1	2	3	4	7	
n_i	2	2	3	1	2	$\sum_{i=1}^{5} n_i = 10$
$\frac{n_i}{n}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\sum_{i=1}^{5} p_i^* = 1$

Характеристики дискретного статистического ряда:

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

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

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

Пусть n – объем выборки.

Если $n = 2k$, то есть ряд имеет четное число членов, то $M_e^* = \frac{x_k + x_{k+1}}{2}$. Если $n = 2k+1$, то есть ряд имеет нечетное число членов, то $M_e^* = x_{k+1}$ (в примере 1. $M_e^* = 3$).

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.

Таблица 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: $h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg n}$, 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 границе.

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

Во второю 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.

Вопрос 2. Описательные статистики выборки

Среднее арифметическое – это условная величина. Реально она не существует. Реально существует общая сумма. Поэтому среднее арифметическое это не характеристика одного наблюдения, а 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. сумма 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 центнеров с га и гречиха – по 5–6 центнеров 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 трудоемкость возведения здания и т. д. – 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, которому 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, когда 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 четное число элементов, то 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 n :

x_i	x_1	x_2	x_3	x_4	$\cdots$	x_m
n_i	n_1	n_2	n_3	n_4	$\cdots$	n_m

где m – число вариантов.

Определение. *Выборочным средним* $\overline{x_в}$ называется среднее арифметическое всех значений выборки:

$$\overline{x_в} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i \,.$$

Выборочное среднее можно записать и так: $\bar{x}_в = \frac{1}{n}\sum_{i=1}^{m} x_i p_i^*$, где p_i^* – частость.

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

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

$$D_в = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^2 \cdot n_i \quad \text{или} \quad D_в = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^2 \cdot p_i^* .$$

Выборочное среднее квадратическое выборки определяется формулой:

$$\sigma_в = \sqrt{D_в} .$$

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*:

$$S^2 = \frac{n}{n-1} D_в .$$

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

Приведем краткий обзор характеристик, которые наряду с уже рассмотренными применяются для анализа статистических рядов и являются аналогами соответствующих числовых характеристик случайной величины.

Среднее выборочное и выборочная дисперсия являются частным случаем более общего понятия – *момента* статистического ряда.

Определение. *Начальным выборочным моментом порядка* l называется среднее арифметическое l – х степеней всех значений выборки:

$$\nu_l^* = \frac{1}{n}\sum_{i=1}^{m} x_i^l \cdot n_i \quad \text{или} \quad \nu_l^* = \sum_{i=1}^{m} x_i^l \cdot p_i^* .$$

Из определения следует, что начальный выборочный момент первого порядка: $V_l^* = \frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i = \overline{x_в}$.

Определение. *Центральным выборочным моментом порядка* l называется среднее арифметическое l – х степеней отклонений наблюдаемых значений выборки от выборочного среднего $\overline{x_в}$:

$$\mu_l^* = \frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x_в})^l \cdot n_i \quad \text{или} \quad \mu_l^* = \sum_{i=1}^{m}(x_i - \overline{x_в})^l \cdot p_i .$$

Из определения следует, что *центральный выборочный момент второго порядка*:

$$\mu_2^* = \frac{1}{n}\sum_{i=1}^{m}(\mathrm{x}_i - \overline{x_в})^2 \cdot n_i = D_в = \sigma_в^2 .$$

Определение. *Выборочным коэффициентом асимметрии* называется число A_s^*, определяемое формулой: $A_s^* = \frac{\mu_3^*}{\sigma_в^3}$.

Выборочный коэффициент асимметрии служит для характеристики асимметрии полигона вариационного ряда. Если полигон асимметричен, то одна из ветвей его, начиная с вершины, имеет более пологий «спуск», чем другая.

Если $A_s^* < 0$, то более пологий «спуск» полигона наблюдается слева; если $A_s^* > 0$ – справа. В первом случае асимметрию называют *левосторонней*, а во втором – *правосторонней*.

Определение. *Выборочным коэффициентом эксцесса* или *коэффициентом крутости* называется число E_k^*, определяемое формулой:

$$E_k^* = \frac{\mu_4^*}{\sigma_в^4} - 3 .$$

Выборочный коэффициент эксцесса служит для сравнения на «крутость» выборочного распределения с нормальным распределением.

Коэффициент эксцесса для случайной величины, распределенной по нормальному закону, равен нулю.

Поэтому за стандартное значение выборочного коэффициента эксцесса принимают $E_k^* = 0$.

Если $E_k^* < 0$, то полигон имеет более пологую вершину по сравнению с нормальной кривой; если $E_k^* > 0$, то полигон более крутой по сравнению с нормальной кривой.

Таблица 3

x_i	n_i	$x_i \cdot n_i$	$x_i - \overline{x_в}$	$(x_i - \overline{x_в})^2 \mathrm{n}_i$	$(x_i - \overline{x_в})^3 \mathrm{n}_i$	$(x_i - \overline{x_в})^4 \mathrm{n}_i$
x_1	n_1					
⋮	⋮					
x_m	n_m					

Окончание таблицы 3

	$\sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} x_i n_i$		$\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^2 n_i$	$\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^3 n_i$	$\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^4 n_i$

x_i – середины интервалов; n_i – частоты; $\sum_{i=1}^{m} n_i = n$ – объем выборки;

с помощью суммы $\sum_{i=1}^{m} x_i n_i$ находим $\overline{\quad}_в$;

с помощью суммы $\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^2 n_i$ находим $D_в$ и $\sigma_в = \sqrt{D_в}$;

с помощью суммы $\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^3 n_i$ находим A_s^*;

с помощью суммы $\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^4 n_i$ находим E_k^*.

Вопрос 3. Нормальное распределение случайной величины

Нормальное распределение, также называемое распределением Гаусса, – это распределение вероятностей, которое играет важнейшую роль во многих областях знаний, особенно в физике. Физическая величина подчиняется нормальному распределению, когда она подвержена влиянию огромного числа случайных помех. Ясно, что такая ситуация крайне распространена, поэтому можно сказать, что из всех распределений в природе чаще всего встречается именно нормальное распределение. Не является исключением и социальная сфера, где большинство величин стремится к нормальному распределению. График нормального распределения приведен на рис. 1: по оси *x* отложены значения наблюдаемой величины со средним значением μ и отклонениями от среднего в δ.

В идеальном случае нормального распределения действует так называемое правило трех сигм.

Почти достоверно, что случайная величина (ошибка) не отклоняется от математического ожидания μ больше, чем на 3δ. Именно это предположение и называют правилом трех сигм.

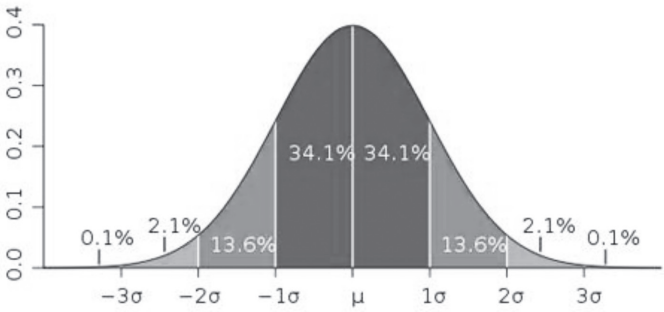

Рис. 1. График плотности вероятности нормального распределения и процент попадания случайной величины на отрезки, равные стандартному отклонению

Вероятности P попадания наблюдаемой величины в интервалы от – 3δ до 3δ следующие:

$$P(-3\delta < x < -2\delta) = 1{,}075\ \%,$$
$$P(-2\delta < x < -1\delta) = 13{,}55\ \%,$$
$$P(-1\delta < x < -\mu) = 34{,}15\ \%,$$
$$P(\mu < x < -1\delta) = 34{,}15\ \%,$$
$$P(1\delta < x < 2\delta) = 13{,}55\ \%,$$
$$P(2\delta < x < 3\delta) = 1{,}075\ \%.$$

Другими словами, с вероятностью 68,3 % отклонение наблюдаемой величины от математического ожидания (среднего значения) не будет превышать одного стандартного отклонения в ту или другую сторону, с вероятностью 95,4 % – двух стандартных отклонений, и с вероятностью 99,7 % – трех. Вероятность того, что наблюдаемая величина окажется за пределами трех стандартных отклонений, составляет 0,3 %.

$$P(x < -3\delta) = 0{,}15\ \%,$$
$$P(3\delta < x) = 0{,}15\ \%.$$

Рассмотренные вероятности свойственны случаю идеального нормального распределения. Распределение эмпирически наблюдаемых величин будет лишь стремиться к такому состоянию.

Заключение

Установление закономерностей, которым подчинены массовые случайные явления, основаны на изучении методами теории вероятностей статистических данных результатов наблюдений.

Первая задача математической статистики – указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.

Вторая задача математической статистики – разработать методы анализа статистических данных в зависимости от целей исследования.

Современная математическая статистика разрабатывает способы определения числа необходимых испытаний до начала исследования (планирования эксперимента), в ходе исследования (последовательный анализ). Ее можно определить как науку о принятии решений в условии неопределенности.

Если кратко, то можно сказать, что задача математической статистики состоит в создании методов сбора и обработки статистических данных для выдвижения обоснованных научных гипотез, их подтверждения или опровержения.

Контрольные вопросы

1. Что такое выборка и генеральная совокупность?
2. Назовите основные описательные статистики.
3. Что такое медиана? Дайте определение.
4. Назовите характеристики разброса случайной величины.
5. Что такое частота (частость) появления случайной величины?
6. Назовите характеристики нормального распределения случайной величины (правило трех сигм).

Лекция 3. Пространственные модели анализа данных

Деятельность органов внутренних дел на современном этапе характеризуется достаточно обширным набором различных показателей, которые отражают криминальную ситуацию, сложившуюся в регионах, динамические характеристики этой ситуации, результаты работы полиции по регистрации, раскрытию и расследованию преступлений, охране общественного порядка и др. Многие из этих показателей взаимообусловлены, отдельные находятся в причинно-следственной связи.

В связи с этим при проведении научных исследований в социальной и правовой сфере необходимо не только иметь четкие представления об основных направлениях развития социума, но и уметь учитывать сложное взаимосвязанное многообразие факторов, оказывающих существенное влияние на изучаемый процесс. Такие исследования невозможно проводить без знания основ теории вероятностей, математического моделирования, многомерных статистических методов, т. е. дисциплин, позволяющих исследователю разобраться в огромном количестве стохастической информации и среди множества различных вероятностных моделей выбрать единственную, наилучшим образом отражающую изучаемый процесс или явление.

Центральное место во всем математико-статистическом инструментарии занимают пространственные методы анализа данных как метод, используемый для получения модели, дающей наилучшую оценку истинного соотношения между исследуемыми переменными.

При построении пространственных моделей приходится сталкиваться и с такой проблемой, как наличие функциональной или тесной корреляционной зависимости между объясняющими переменными, т. е. мультиколлинеарности. Это может привести к получению неустойчивых, не имеющих реального смысла оценок. При изучении социальных процессов может оказаться необходимым включить в модель фактор, имеющий два или более качественных уровней. Это могут быть разного рода качественные признаки, например, образование, пол, профессия, принадлежность к определенному региону. Такого рода переменные в социометрике принято называть фиктивными переменными. Качественные признаки могут существенно влиять на структуру линейных связей между переменными и приводить к скачкообразному изменению параметров регрессионной модели. В этом случае говорят об исследовании

регрессионных моделей с переменной структурой или построении регрессионных моделей по неоднородным данным.

Вопрос 1. Линейные модели парной регрессии

Линейная регрессия (от 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 *X1, X2, ... Xm*: $Y_x = f(x_1, x_2, ..., x_m)$.

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):

$$Y_i = a + b \cdot x_i + e_i,$$

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:

$$\sum_{i=1}^{n} (y_i - y_x)^2 \to \min .$$

Интерпретация параметров модели представлена на рис. 2.

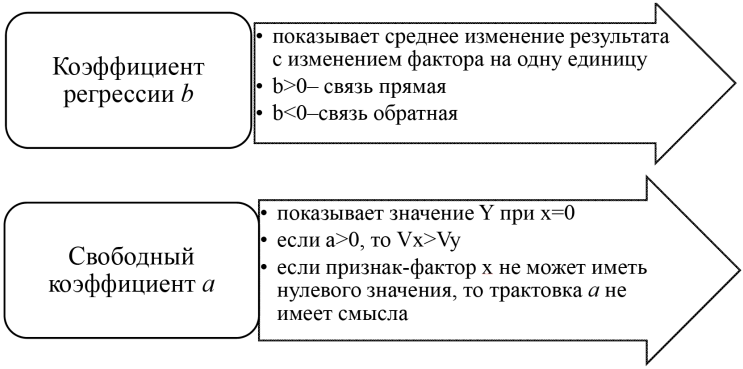

Рис. 2. Интерпретация параметров модели

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 rxy 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.*

В графическом виде сущность метода наименьших квадратов можно представить в следующем виде (рис. 2.1).

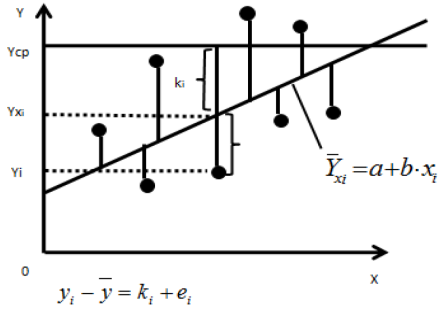

Рис. 2.1. Геометрическая интерпретация

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

Например, возьмем два показателя, характеризующих 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);

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.

Аналитическая функция в линейном виде представлена на рис. 2.2.

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

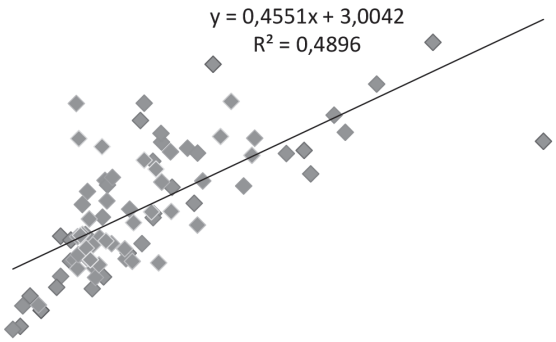

Рис. 2.2. По оси X – количество тяжких
и особо тяжких преступлений, совершенных на бытовой почве
(в расчете на 100 тыс. населения), по оси Y –
число несовершеннолетних, совершивших преступления
(на 1 тыс. несовершеннолетних в возрасте 14–17 лет)

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 интерпретация процесса моделирования и оценка построенного уравнения возможны только для парной регрессии.

Вопрос 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*)[3].

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: $D(\varepsilon_i) = D(\varepsilon_j) \forall i \neq j$.

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

4. Случайная компонента ε является нормально распределенной случайной величиной.

5. 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}:

$$t_{pi} = \frac{b_i}{S_{bi}},$$

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

Далее полученная величина t_{pi} сравнивается с 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.* = $n - m$).

[3] *Бородич С. А.* Эконометрика: учеб. пособие. Мн.: Новое знание, 2004. С. 122–125.

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 *bi* считается значимым, если соответствующее ему *p*-значение меньше *0,05 (pi<0,05)*. Данная пороговая величина говорит о том, что мы определили *bi* с точностью 95 %. В табл. 4 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 *bii* уравнения регрессии:

$$H_0 : b1 = b2 = ... = b_n = 0.$$

Таблица 4

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 $D(\varepsilon)$:

$$D(Y) = D(Yx) + D(\varepsilon),$$

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

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

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

Это означает (см. рис. 2.2), что построенная по исходным данным линия регрессии идеально соответствует регрессионной модели (полностью ее объясняет), параметры уравнения регрессии определены точно и гипотеза H_0 $(b_1 = b_2 = ... = b_n = 0)$ может быть отклонена.

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

$$F_p = \frac{R^2 / n}{\left(1 - R^2\right) / (m - n - 1)},$$

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 ($\alpha = 0{,}05$).

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 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 $\alpha = 0{,}05$: $F_{кр} = 4{,}67$.

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 5

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 табл. 6 приведены результаты дисперсионного анализа, полученного с помощью Microsoft Excel.

Таблица 6

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) отражены основные характеристики двух компонент регрессионной модели: уравнения регрессии (строка – «регрессия») и случайной составляющей (строка – «остаток»).

В столбце **df** приведены значении 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.

Вопрос 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 (*dummy variables*, 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 $(X'X)$ ста-

новится 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.

Заключение

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.

Лекция 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 зачастую 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 рассмотрением 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.

Вопрос 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 описа-

на таким показателем, как количество зарегистрированных тяжких преступлений, и т. п.

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

В табл. 7 приведен пример интервального временного ряда, уровни которого фиксируют количество зарегистрированных грабежей за определенный год в одном из субъектов Российской Федерации. На рис. 3 представлен тот же ряд в графической форме (для наглядности все точки соединены отрезками ломанной).

Таблица 7

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).

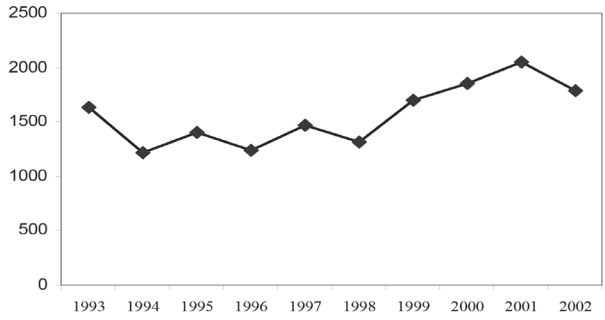

Рис. 3. Графическая форма представления временного ряда

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 информацион-

ные 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: $T_{pi} = \frac{Y_t}{Y_{t-1}} \times 100$;

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

– *средний темп роста* – характеризует интенсивность изменения уровней ряда $\overline{T}_p = \sqrt[t]{\prod_{i=1}^{t} T_{pi}}$.

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* $\boldsymbol{\varepsilon}_t$: $Y_t = R(t) + \boldsymbol{\varepsilon}_t$, 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 специфики дея-

тельности 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.

Вопрос 2. Методы исследования компонент временного ряда

1. Метод последовательных разностей.

Данный метод основывается на представлении временного ряда Y_t в виде регулярной и случайной компонент: $R(t)$ и $\boldsymbol{\varepsilon}_t$.

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:

$$Y_1 = a + b + \boldsymbol{\varepsilon}_1$$
$$Y_2 = a + 2 \cdot b + \boldsymbol{\varepsilon}_2$$
$$...$$
$$Y_n = a + n \cdot b + \boldsymbol{\varepsilon}_n.$$

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

$$\Delta'_1 = Y_2 - Y_1 = b + (\boldsymbol{\varepsilon}_2 - \boldsymbol{\varepsilon}_1)$$
$$\Delta'_2 = Y_3 - Y_2 = b + (\boldsymbol{\varepsilon}_3 - \boldsymbol{\varepsilon}_2)$$
$$...$$
$$\Delta'_{n-1} = Y_n - Y_{n-1} = b + (\boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_{n-1}).$$

Из выражений для первых разностей видно, что они зависят только от разности $\boldsymbol{\varepsilon}_t$, т. е. от величин случайной компоненты в соседних наблюдениях.

Если они незначительно отличаются друг от друга, (т. е. $\forall t : t = 1, 2, ..., n : \boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_{t-1} \approx 0$ и $\Delta_1' \approx \Delta_2' \approx ... \Delta'_{n-1}$), выбор полинома первой степени является правильным.

Таким образом, получаем: $Y_t = a + b \cdot t$.

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: $R(t) = a + bt$;

– 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);

– степенной: $R(t) = a \cdot t^b$;

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

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

– логарифмической: $R(t) = a + b \cdot \ln t$.

Степень соответствия измеряется с помощью рассчитываемого коэффициента R^2. Функция, имеющая большее 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 также величина служит 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. 3.1 (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. 3.1 представлены также график (сплошная линия) и параметры оцениваемой нами линейной функции ($R(t) = 300401 + 45190 \cdot t$; $R^2 = 0{,}965$).

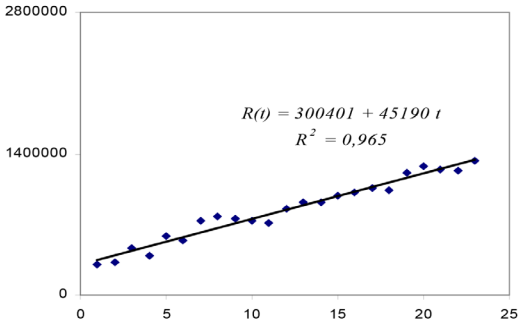

Рис. 3.1. Исходный ряд и параметры выбранной линейной функции

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. 3.2).

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 $t = 1, 2, \ldots 23$ вычислим остатки для линейной и параболической функции соответственно и итоги занесем в столбцы электронной таблицы:

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

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

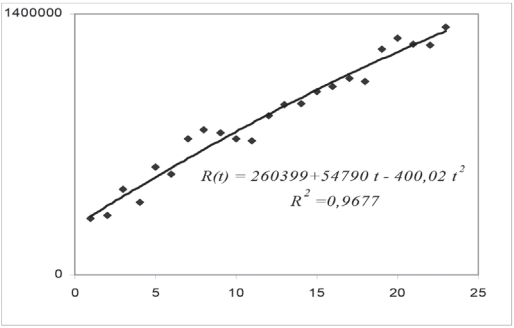

Рис. 3.2. Исходный ряд и параметры выбранной параболической функции

На рисунках 3.3 и 3.4 приведены гистограммы плотности распределения значений остатков для линейной модели и модели, представленной полиномом второй степени соответственно.

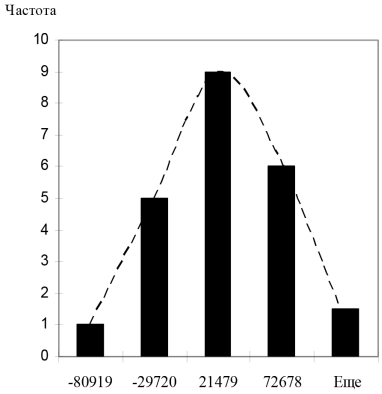

Рис. 3.3. Гистограмма остатков линейной функции

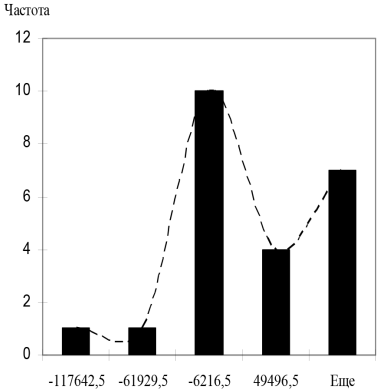

Рис. 3.4. Гистограмма остатков параболической функции

Из рисунков видно, что в качестве временной модели следует использовать линейную функцию. График плотности распределения ее остатков больше напоминает вид нормального распределения (рис. 3.5), по сравнению с графиком гистограммы остатков параболической функции. Это означает, что в линейной модели значения случайной компоненты носят действительно более слу-

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

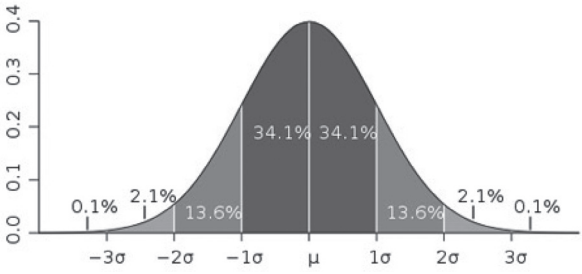

Рис. 3.5. График плотности вероятности нормального распределения и процент попадания случайной величины на отрезки, равные стандартному отклонению

Вопрос 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 б*о́*льшие 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:

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

где S_t – 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:

$$S_t = S_{t-1} + \alpha (Y_t - S_{t-1}).$$

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 сезонный характер регистрируемой преступности (рис. 3.6 – 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 сглаживание» 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:

$$\varepsilon_t = Y_t - S_t.$$

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.

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

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

В качестве примера на рисунках 3.7 и 3.8 приведены гистограммы остатков для $\alpha = 0{,}2$ и $\alpha = 0{,}5$ соответственно.

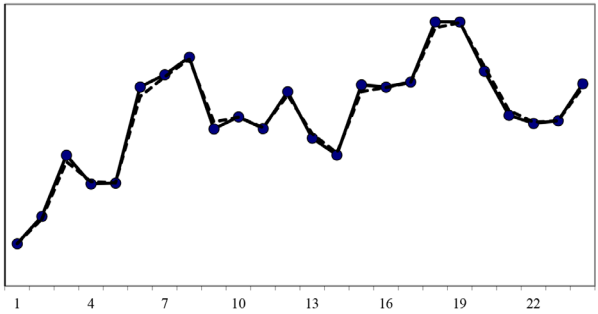

Рис. 3.6. Исходный временной ряд – сплошная линия; сглаженный временной ряд ($\alpha = 0{,}2$) – пунктирная линия

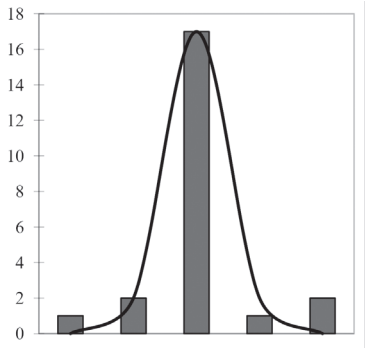

Рис. 3.7. Гистограмма остатков $\alpha = 0{,}2$.

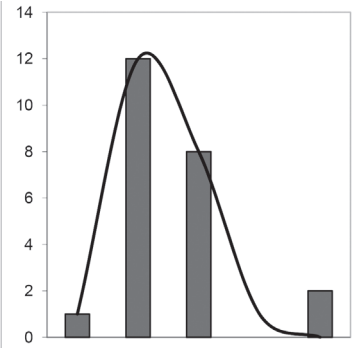

Рис. 3.8. Гистограмма остатков $\alpha = 0{,}5$.

График сглаженного временного ряда с использованием параметра сглаживания $\alpha = 0{,}2$ представлен на рисунке 4.7 пунктирной ломаной линией. Как видно из рисунка, он практически совпадает с первоначальным графиком.

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

Заключение

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. Основные компоненты временного ряда.
2. Особенности использования метода скользящего среднего для сглаживания временного ряда.
3. Моделирование тренда, циклической компоненты.
4. Определение и формы представления временного ряда, выявление тенденции и оценка параметров линии тренда.
5. Аналитические функции, используемые для моделирования тенденций временного ряда.
6. Методы прогнозирования временных рядов.

Лекция 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 РФ № 1246 от 30 декабря 2007 г. 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 пунктах 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.

Вопрос 1. Технологии первичной обработки данных анкетных опросов

Первичная компьютерная обработка результатов опроса есть последовательная «набивка» каждого заполненного документа (анкеты, бланки 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.

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. 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

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

Вопрос 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 величин 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.

$$i = \frac{C - H}{C + H}.$$

C – число пар, у которых знаки отклонений значений от их средних совпадают.

H – 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:

$$\tau = \frac{2s}{n(n-1)},$$

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:

$$\tau = \frac{S}{\sqrt{\left[\frac{n(n-1)}{2} - U_x\right]\left[\frac{n(n-1)}{2} - U_y\right]}},$$

$$U_x = \frac{\sum t(t-1)}{2},$$

$$U_y = \frac{\sum t(t-1)}{2},$$

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 (X_1, Y_1), ..., (X_n, Y_n). Каждому показателю X и Y присваивается ранг. Ранги значений X расположены в естественном порядке $i=1, 2, \ldots, n$. Ранг Y записывается как R_i и соответствует рангу той пары (X, Y), для которой ранг X равен i. На основе полученных рангов X_i и Y_i рассчитываются их разности d и вычисляется коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{6\sum d^2}{\mathrm{n}(\mathrm{n}^2 - 1)}.$$

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)

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

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

m – число групп, которые ранжируются.

n – число переменных.

R_{ij} – ранг i-фактора у j-единицы.

Значимость:

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

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

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

В случае наличия связанных рангов:

$$W = \frac{12S}{m^2(n^3 - n) - m\sum_{j=1}^{m}(t^2{}_j - t_j)},$$

$$\chi^2 = \frac{12S}{mn(n+1) - \frac{\sum_{j=1}^{m}(t^3{}_j - t_j)}{n-1}}.$$

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.

Вопрос 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:

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

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 (а) 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, C – 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 $p \le 0{,}05$ (принятой в статистических исследованиях) и $n = 2$ критическое значение $\chi^2 = 5{,}99$.

Полученное эмпирическое значение больше критического – различия частот достоверны ($\chi^2 = 9{,}64$; $p \le 0{,}05$), 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	ДПС	высшее	частично	от 1 до 3 лет
2	26–30	до 5	ППСП	высшее	не устраивает	до 1 года
3	20–25	от 5–10	СУ	высшее	частично	от 3 до 5 лет
4	25–30	от 11–15	ЭБиПК	среднее	не устраивает	до 1 года
5	20–25	до 5	УР	высшее	не устраивает	до 1 года
6	41-45	свыше 20	ППСП	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 1до 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 m_i.

Построим матрицу теоретических (ожидаемых) частот m'_i, 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»:

$$\chi_{\Phi}{}^2 = \sum_{i=1}^{n} \frac{(m_i - m'_i)^2}{m'_i}.$$

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

$$\text{Рассчитаем: } \chi_{ф}^{2} = \sum_{i=1}^{n} \frac{(m_i - m_i')^2}{m_i'} \approx 12{,}87.$$

Найдем критическое значение $\chi_{кр}^{2}$. Для этого зададим вероятность ошибки равную 0,05 и найдем число степеней свободы n = (R – 1) * (C – 1) = (3–1)*(3–1) = 4.

$$\chi_{кр}^{2} \approx 9{,}49.$$

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.

Заключение

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 имеющуюся 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.

Лекция 6. Моделирование криминальных процессов в социальных системах

Методы математического моделирования криминальных процессов имеют общий объект исследования – преступность как социальную систему-организацию. Методы и модели анализа криминальных процессов имеют собственный предмет – количественные взаимосвязи и закономерности функционирования и развития объекта исследования. В дисциплине «Управление в социальных и экономических системах» используемые математические методы сами становятся объектом исследования.

Ввиду отсутствия возможности экспериментального изучения реальных социальных систем моделирование является практически единственным инструментом исследования при помощи вспомогательных объектов, называемых моделями.

Попытки формализации сложных социальных явлений и процессов при помощи математических методов и моделей начались в конце 60-х – начале 70-х гг. Модель, описывающая преступность, впервые была предложена в работе Г. Беккера:

$$O_j = O_j\,(p_j, f_j, u_j),$$

где O – количество правонарушений, которые должен совершить преступник, p – вероятность его задержания, f – мера наказания в случае его поимки, u – переменная, включающая факторы, оказывающие влияние на решение совершить преступление. Ввиду того, что выбор преступника происходит в условиях неопределенности, ожидаемая полезность от преступления описывалась как:

$$EU_j = p_j U_j(Y_j - f_j) + (1-p_j)\,U_j\,(Y_j),$$

где Y_j – доход от совершенного преступления. В уравнении представлена не только ожидаемая полезность, зависящая от дохода, но и вероятность его задержания p_j, против успешного преступления $(1 - p_j)$, и мера наказания f_j.

Совершенное преступление оказывает влияние не только на преступника, но и на общество в целом, функция убытков общества г. Беккером представлена в виде:

$$L = D(O) + C(p, O) + bfpO,$$

где D – ущерб от преступления, C – стоимость задержания и поимки преступника, $bfpO$ – затраты общества на правоохранительную систему, b – коэффициент, отражающий стоимость правоохранительной системы, штрафы и другие «мягкие» меры наказания приводят к значению данного коэффициента близким к 0, тогда как лишение свободы и другие более жесткие меры наказания приводят к $b>1$. В данном уравнении такая социальная функция государства, как защита обще-

ства, личности и человека от противоправных и иных посягательств, представлена переменными f и p. Регулируя их, государство старается минимизировать влияние преступлений на общество и уменьшить риск совершения новых правонарушений. Другим инструментом является изменение переменной u, оказывающей влияние на факторы, влияющие на выбор человека совершить преступление.

Таким образом, задача снижения степени общественной опасности преступления зависит от мер предпринимаемых государством. Так, уменьшая возможный доход от преступного поведения, увеличивая вероятность поимки преступника и регулируя строгость наказания, согласно данной теории, можно добиться снижения уровня преступности.

Описанная модель имеет существенный недостаток: она не учитывает временной фактор. В работе И. Эрлиха исследовалось влияние уровня дохода и распределения доходов на преступность. Функция ожидаемой полезности была адаптирована с учетом временного фактора:

$$EU_j = p_j U_j (Y_j, t_j) + (1 - p_j) U_j (Y_j, t_j),$$

где $U_j (Y_j, t_j)$ – функция от полученного дохода и времени, расходуемого индивидом на занятие преступной деятельностью. Было доказано, что более высокие семейные доходы связаны с насильственными преступлениями, а также что безработица имеет низкую статистическую связь по отношению к уровню преступности.

Таким образом, в соответствии с классическими теориями, увеличение вероятности ареста или суровости наказания приводит к уменьшению ожидаемой полезности от противоправной деятельности и, как следствие, к снижению уровня преступности.

Основной идеей рассмотренных моделей, основывающихся на экономическом подходе, является наличие неравенства в доходах. В регрессионном анализе для оценки влияния институциональных изменений часто используется фиктивные переменные. В исследовании Дж. Лотта и Б. Мустарда были использованы панельные данные для оценки системы уравнений регрессии. Зависимыми переменными в каждом уравнении являлись различные виды преступлений (убийства, изнасилования, грабежи, кражи и т. д.), в качестве регрессоров выступали количество арестов (A_{jt}), различные социально-экономические факторы (X_{jt}) и фиктивная переменная, отражающая нормативно-правовые факторы (H_{jt}):

$$C_{jt} = \alpha + \gamma H_{jt} + \beta A_{jt} + \delta X_{jt} + \varepsilon_{jt}.$$

Проблема исследования факторов преступности при помощи методов математического моделирования социально-правовых процессов рассматривалась в трудах других отечественных ученых. Основными факторами, выделяемыми различными исследователя-

ми в разное время, являются имущественное расслоение общества, уровень безработицы, уровень образования, соотношение городского и сельского населения и др.

Вопрос 1. Преступность как объект моделирования

На современном этапе развития общественных отношений криминальные процессы, происходящие в стране, оказывают значительное влияние на уровень социально-экономического развития. Задачи анализа данных о преступности зачастую решаются описательными методами, оцениваются ее структурно-динамические характеристики. Применение математического моделирования в значительной мере расширяет инструментарий аналитика. На следующем рисунке представлена структурная схема исследования преступности (рис. 4).

Рис. 4. Схема исследования преступности

Рассмотрим процесс построения математических моделей исследования социально-экономических систем, состоящий из четырех этапов.

1. Формирование гипотез и разработка концептуальной модели.

2. Разработка модели или системы моделей в рамках разработанной концептуальной модели.

3. Анализ результатов моделирования и сравнение с фактическими данными.

4. Формирование новой гипотезы (корректировка модели).

Условная схема оперативной обстановки представлена на рис. 4.1.

Рис. 4.1. Оперативная обстановка

Результативность ОВД можно представить в виде производственной функции, где количество раскрытых преступлений Y_i на территории i определяется уровнем преступности Q, социально-экономическими факторами X_i и кадровыми ресурсами полиции R_i:

$$Y_i = f_i(Q, X_i, R_i).$$

Традиционная экономическая теория предполагает, что производители стремятся максимизировать прибыль (J) с учетом ограничений, налагаемых производственной функцией. В формальном виде данное соотношение можно записать как: $maxJ = pY - cQ - wR$, при $Y = F(Q, R)$, где Y – количество произведенной продукции, p – цена продукции, Q – капитал, c – цена капитала, R – труд, w – ставка заработной платы, $Y = F(Q, R)$ – производственная функция.

Для аппроксимации производственной функции используем опыт экономико-математического моделирования, в котором наибольшее применение получила функция Кобба-Дугласа:

$$Y_i = AQ_i^{\alpha_i} R_i^{\beta_i}, \alpha_i \beta_i \in [0,1], i = 1,...,N,$$

где α_i и β_i – оцениваемые параметры модели.

Данное уравнение (3) можно представить в линейном виде:

$$\ln\left(\frac{Y}{R}\right)_i = \ln A + \alpha \ln\left(\frac{Q}{R}\right)_i + \beta \ln X_i^j + \ln \varepsilon_i,$$

где Y – число раскрытых преступлений, Q – число совершенных преступлений, R – численность сотрудников полиции, X – набор социально-экономических факторов.

Вопрос 2. Модели криминальных процессов

Одними из первых моделей, разработанных отечественными исследователями, были математические модели, предложенные О.А. Гавриловым и В.А. Колемаевым. Они оценивали влияние демографических процессов на состояние преступности, в частности зависимость между численностью населения и количеством осужденных. Отсутствие достаточного статистического материала по Советскому Союзу привело к тому, что в основу моделей легли данные о преступности в Польской Народной Республике.

Полученная зависимость описывалась линейной однофакторной моделью:

$$x = \alpha_0 + \alpha \cdot X + \varepsilon,$$

где x – количество осужденных за год; X – численность населения на данный год по воеводствам (в тыс. человек); α – показатель влияния численности населения на число преступных проявлений; α_0, ε – параметры, характеризующие совокупное влияние прочих факторов.

На основе анализа статистических данных были получены уравнения, связывающие количество осужденных преступников с численностью населения в 1962 и 1963 гг.:

$$\begin{cases} x_{1962} = 2661 + 7{,}9 \cdot X \\ x_{1963} = 2856 + 6{,}8 \cdot X. \end{cases}$$

Эти уравнения показывают, что при увеличении населения на тысячу человек количество осужденных преступников возросло бы в среднем на 7,9 человек в 1962 г. и на 6,8 человек в 1963 г.

По данным за 1968 г. о количестве городского и сельского населения в 22 воеводствах ПНР и числе осужденных за конкретный один вид преступления – кражи, была достроена и обобщенная-двухфакторная модель:

$$x = 4588 + 7{,}4 \cdot Y + 0{,}4 \cdot Z,$$

где Y – численность городского населения (в тыс. человек); Z – численность сельского населения (в тыс. человек).

Уравнение свидетельствует о том, что увеличение городского населения на тысячу человек приводит к увеличению числа краж (осужденных за кражи) в среднем на 7,4, а увеличение сельского населения – на 0,4. Коэффициент детерминации, характеризующий обоснованность модели по расчетам авторов, составил 0,68.

В следующей таблице представлены некоторые результаты исследования криминальных процессов за рубежом и в нашей стране.

Таблица 8

Авторы (Год)	Зависимые переменные	Независимые переменные	Основные результаты	Применяемые методы
Отечественная литература				
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 растут с неравенством в распределении доходов	Метод наименьших квадратов
Зарубежная литература				
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	Корреляционный анализ
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	МНК, ОММ, динамические модели панельных данных
Stephen Machin and Costas Meghir. Crime and Economic Incentives (2004)	Преступления против собственности	Заработная плата, доля населения в возрасте 15–24 лет, срок наказания	Уровень заработной платы оказывает наибольшее влияние на уровень преступности	МНК. Модели с лаговыми зависимыми переменными
Eric Neumayer. Inequality and Violent Crime: Evidence from Data on Robbery and	Число грабежей на 1 млн населения	Индекс Джини, ВВП на душу населения, темпы роста ВВП, безработица,	Наибольшее влияние неравенства в доходах	ОММ. Модели с лаговыми переменными. Модели с фиксированными

Окончание таблицы 8

Авторы (Год)	Зависимые переменные	Независимые переменные	Основные результаты	Применяемые методы
Violent Theft (2005)		урбанизация, мужчины в возрасте 15–64		и случайными эффектами
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	Модель авторегрессии и распределенного лага (ARDL)
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 лет, владельцы оружия	Три фактора оказались значимыми во всех регрессиях: плотность населения, неустойчивость семьи и соотношение между женщинами и мужчинами	Двухэтапная регрессия МНК. Взвешенный МНК

Моделирование является эффективным методом исследования не только социальных систем. С его помощью можно и социальное явление представить в виде системы, чтобы затем, конкретизируя ее состав и внутреннюю организацию, определить особенности протекания, взаимодействия с внешней средой и, в конечном итоге, выработать адекватные подходы к управлению.

Следуя методологии системного подхода, моделирование любого объекта должно осуществляться комплексно, во взаимосвязи с основными компонентами, принимающими участие в тех же процессах, в которых задействован сам объект.

В рассматриваемой нами предметной области задача моделирования управления органами внутренних дел неотделима от задачи моделирования преступности, поскольку ОВД и преступность как

социальное явление выступают главными составляющими процесса борьбы с преступностью.

Следует подчеркнуть, что эффективность моделирования во многом определяется правильным выбором типа моделей, который, в свою очередь, зависит от особенностей объекта.

Вопрос 3. Технология построения моделей криминальных процессов

Классические методы решения задач исследования преступности (корреляционно-регрессионный анализ, кластерный анализ, анализ временных рядов и др.) часто приводят к несопоставимым для многих субъектов Российской Федерации результатам. Связано это в первую очередь с неоднородностью данных и различиями протекающих в регионах социально-экономических процессов, которые приводят к расхождениям в оценке деятельности ОВД. Значимость этой проблемы вызвана неравномерностью показателей преступности, что приводит к росту негативных социальных явлений, особенно остро проявляющихся в период кризиса.

Иерархическая система организации деятельности ОВД предполагает построение специфических моделей анализа данных на различных уровнях. В ходе исследования автором разработана специальная технология построения моделей анализа результатов деятельности территориальных органов, состоящая из пяти последовательных этапов (рис. 5). На первом этапе строится модель по панельным данным на федеральном уровне. Второй этап заключается в проверке адекватности модели, для чего на основе методов кластерного анализа производится разбиение регионов на группы по их основным административно-территориальным характеристикам. На данном этапе формулируется задача классификации данных, которая решается при помощи моделей типологизации. На следующем этапе производится адаптация построенной модели для отдельных регионов и ее улучшение. Далее в модель включаются фиктивные переменные, отражающие природно-географические особенности регионов. На последнем этапе осуществляется проверка пригодности модели для построения прогнозов. Очевидно, что данный механизм может быть использован и для анализа данных о состоянии преступности, а также для решения задач описания преступности и прогнозирования тенденций ее развития.

Необходимость возврата на предыдущие этапы может быть обусловлена непригодностью модели для описания преступности

в исследуемых регионах, а также для включения в модель факторов, отражающих региональные особенности. Таким образом, обратные связи на каждом функциональном блоке отражают деятельность ЛПР по формированию корректирующих воздействий на существующую систему управления.

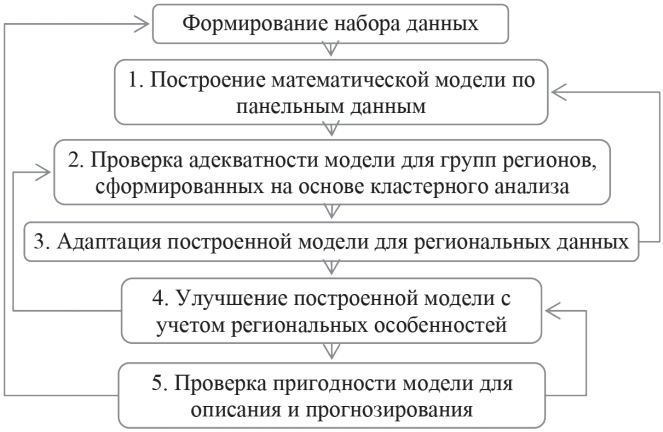

Рис. 5. Структурная схема технологии анализа данных

Эффективное решение задач обеспечения правопорядка невозможно без комплексной оценки действующих факторов внешней среды, особенно социально-экономических факторов, действующих в регионах. Очевидно, что мониторинг состояния преступности в общественных местах является одним из основных компонентов анализа криминогенной обстановки, который основывается на необходимости применения математических методов и моделей в повседневной оперативно-служебной деятельности ОВД. Предварительным этапом построения моделей является специальная процедура идентификации, которая включает в себя методы анализа временных рядов.

При анализе данных о состоянии преступности, на основе эконометрических методов, замечено, что построенные модели отличаются достаточно низкими прогнозными качествами и недостаточно полно описывают статистические взаимосвязи. Очевидно, что данные недостатки связаны с некоторыми ограничениями и ошибками, допускаемыми при построении уравнений регрессии. По нашему мнению, наиболее существенными из них являются: недостаточно глубоко проработанная теория исследования преступности как сложно-формализуемого социального явления, а также несовершенство современной системы сбора и обработки данных о состоянии преступности.

Предложенная теоретическая модель является нелинейной по оцениваемым параметрам. Однако такие модели не всегда удобно представить в логарифмическом виде и следует исследовать модель на другие функциональные формы. Следовательно, можно сказать, что одним из условий построения качественной модели является исследование ее на спецификацию (функциональную форму). Прямой способ проверить нелинейность – добавить в модель некоторые степени переменных, например, квадраты, и проверить статистическую зависимость этих переменных при помощи *F*-значения.

Для исследования спецификации применим *RESET*-тест, который был предложен Дж. Б. Рамсеем в 1969 г. Вначале находятся оцененные значения из первоначальной модели регрессии и генерируются новые регрессоры, которые являются функциями от значений $\hat{y} = x'\hat{\beta} + \varepsilon$. Например, $w = \left[(x'\hat{\beta})^2, (x'\hat{\beta})^3, ..., (x'\hat{\beta})^p\right]$, затем оценивается модель $y = x'\beta + w'\gamma + \varepsilon$ и проверяется гипотеза $H_0 : \gamma = 0$ против $H_a : \gamma \neq 0$. Оценку данной гипотезы произведем на основе статистики Вальда. Если полученное значение меньше критического, то функциональная форма модели принимается. Очевидно, что для практических целей достаточно применение *p* равным 2 или 3.

Корректная спецификация уравнения регрессии является базовым принципом построения качественной модели и отражает правильное соотношение между используемыми показателями. Очевидно, что неправильный выбор функциональной формы уравнения либо избыточный (недостаточный) набор объясняющих переменных приводит к низкому качеству оценивания моделей. Рассмотренный этап спецификации может включаться в процедуру идентификации модели. Однако при решении практических задач целесообразно вынести данную процедуру в отдельный шаг технологического алгоритма построения моделей анализа данных (рис. 5.1).

Рассмотрим более подробно каждый этап приведенного на рисунке алгоритма.

1. На основе корреляционно-регрессионного анализа осуществляется идентификация модели, включающая в себя формулировку гипотез о существовании статистической зависимости между данными.

2. Оцениваются структурные изменения на основе теста Г. Чоу.

3. Осуществляется тестирование модели на нелинейность при помощи *RESET*-теста Рамсея. Строятся нелинейные модели (логарифмическая, степенная).

4. Методом наименьших квадратов производится оценка: коэффициентов модели $b_0 ... b_k$ и дисперсии ошибок $\hat{\sigma}^2$.

5. На основе *P*-значения (или *t*-статистики) определяется уровень значимости коэффициентов модели.

6. Оценивается уровень значимости модели в целом (на основе *F*-статистики).

7. Исследуется модель на выполнение условий Гаусса-Маркова.

8. Осуществляется выбор наилучшей модели на основе информационных критериев: *AIC*, *BIC* и *HQ*.

Рис. 5.1. Блок-схема алгоритма построения моделей анализа данных

Контрольные вопросы

1. Подходы к моделированию преступности.
2. Нормативные модели.
3. Дескриптивные модели.
4. Модель Г. Беккера.
5. Модель О. Гаврилова и В. Колемаева.
6. Методика исследования преступности.

Заключение

Успешность применения математических моделей в аналитической работе во многом зависит от правильного выбора вида модели. Для обоснования необходимости разработки и использования моделей того или иного вида требуется провести исследование с позиций системного подхода. В нем в обязательном порядке должны быть отражены два основных направления, изучение состояния преступности, ее динамики, структурных изменений и изучение современных особенностей управления органами внутренних дел. Причем каждое из направлений следует реализовывать с точки зрения перспектив моделирования, делая акцент на наиболее актуальных проблемах.

Предложенный курс лекций не претендует на полноту описания затрагиваемых социальных процессов. Вместе с тем предложенные методы и модели являются удобным инструментом как в научном исследовании, так и в практической деятельности аналитических подразделений.

Список литературы

Андерсон Т. Введение в многомерный статистический анализ. М.: Физматгиз, 1963. 500 с.

Блувштейн Ю.Д. Криминология и математика. М.: Юридическая литература, 1974. 176 с.

Бородин С.В. Борьба с преступностью: теоретическая модель комплексной программы. М., 1990. 72 с.

Бурков В.Н. Основы математической теории активных систем. М., 1977. 255 с.

Горошко И.В. Использование многомерных статистических методов анализа и прогнозирования результатов деятельности ОВД на региональном уровне: методические рекомендации / И. В. Горошко, В. А. Апульцин, М. П. Дубинин и др. М., 2015. 44 с.

Горошко И. В. Математическое моделирование в управлении органами внутренних дел. М., 2000. 145 с.

Горошко И.В. Правовые основы информационных технологий / Горошко И. В., Лебедев В. Н., Петрова В. Ю. М., 2015. 68 с.

Горошко И. В., Сичкарук А. В., Флока А. Б. Методы и модели анализа данных в правоохранительной деятельности. М., 2007. 224 с.

Девятко И.Ф. Методы социологического исследования / И. Ф. Девятко. Екатеринбург, 1998. 208 с.

Доугерти К. Введение в эконометрику: пер. с англ. М., 1999. 402 с.

Информационные технологии в управлении органами внутренних дел: учебник / В. В. Баранов и др. под ред. И. В. Горошко. М., 2015. 155 с.

Кравченко Ю.А. Работа полицейского в MS Excel 2013: практ. пособие / Ю. А. Кравченко М., 2016. 320 с.

Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник / Н. Ш. Кремер, 2001. 551 с.

Лавриненко В.Н. Исследование социально-экономических и политических процессов: учеб. пособие / В. Н. Лавриненко, Л. М. Путилова. М., 2004. 184 с.

Лебедев В.В. Математическое моделирование социально-экономических процессов / В. В. Лебедев. М., 1997. 229 с.

Минаев В.А. Кадровые ресурсы органов внутренних дел: современные подходы к управлению. М.: Академия управления МВД СССР, 1991. 163 с.

Моисеев Н.Н. Математические задачи системного анализа. М., 1981. 488 с.

Тихомиров Н. П., Дорохина Е.Ю. Эконометрика: учебник. М., 2002. 640 с.

Новиков Д.А. Теория управления организационными системами. М., 2005. 581 с.

Носко В.П. Эконометрика для начинающих. М., 2005. 379 с.

Орлов А.И. Теория принятия решений: учеб. пособие / А.И. Орлов М., 2004. 656 с.

Рой О.М. Исследование социально-экономических и политических процессов: учебник для вузов / О.М. Рой. СПб., 2004. 364 с.

Социальная статистика / под ред. И.И. Елисеевой. М., 2003. 480 с.

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

Франчук В.И. Основы построения организационных систем. М., 1991. 111 с.

Ядов В.А. Стратегия социологического исследования: Описание, объяснение, понимание социальной реальности / В.А. Ядов. М., 2001. 316 с.

Fishburn P.C. Utility Theory for Decision-Making. N. Y., 1970. 234 p.

Pyle D.J. The economics of crime and law enforcement. London, 1983.

William H. Greene. Econometric Analysis. 7th ed. Boston, 2012. 1231 p.

Wooldridge J.M. Introductory Econometrics: A modern approach. Michigan, 2009. 865 p.

Оглавление

Учебное издание

Математические методы
исследования социальных систем

Курс лекций

Редактор: *М.А. Фильчагина*
Верстальщик: *А.А. Мельникова*

Подписано в печать. Формат 60 х 84 $^{1}/_{16}$.
Усл. печ. л. 4,65. Уч-изд. л. 3,21. Тираж 77 экз. Заказ № 04у.

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

ISBN 978-5-906942-95-1