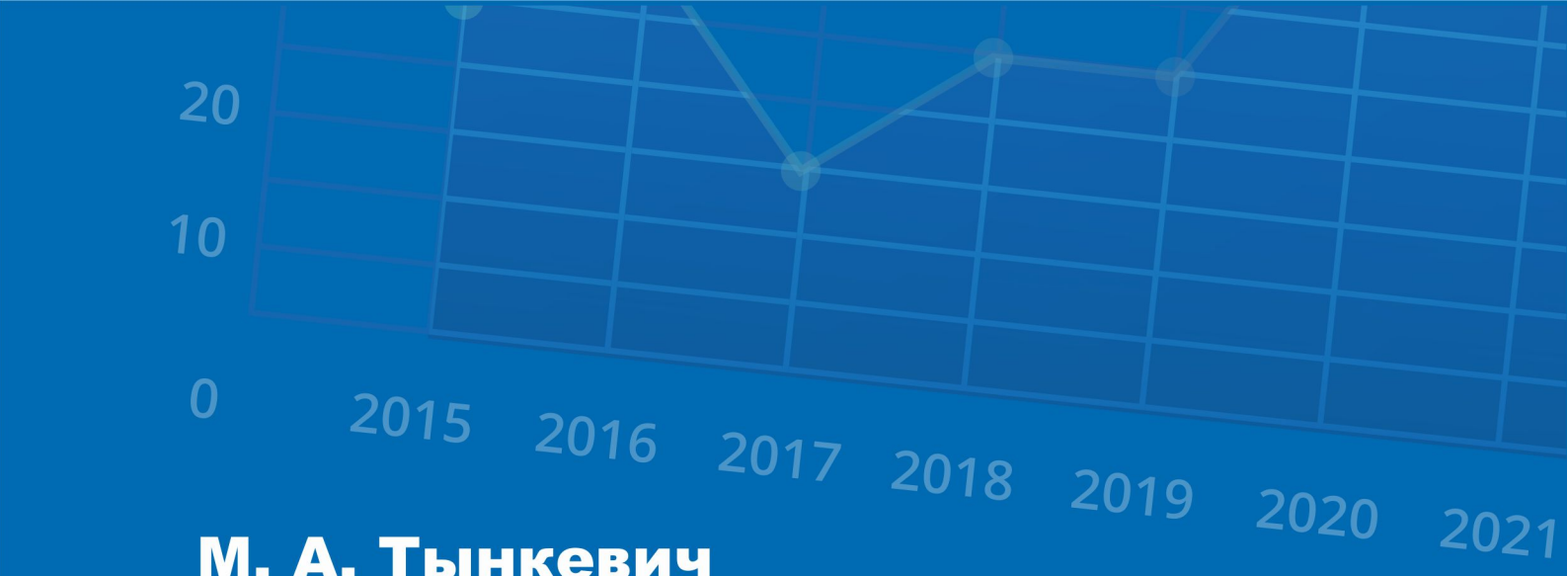
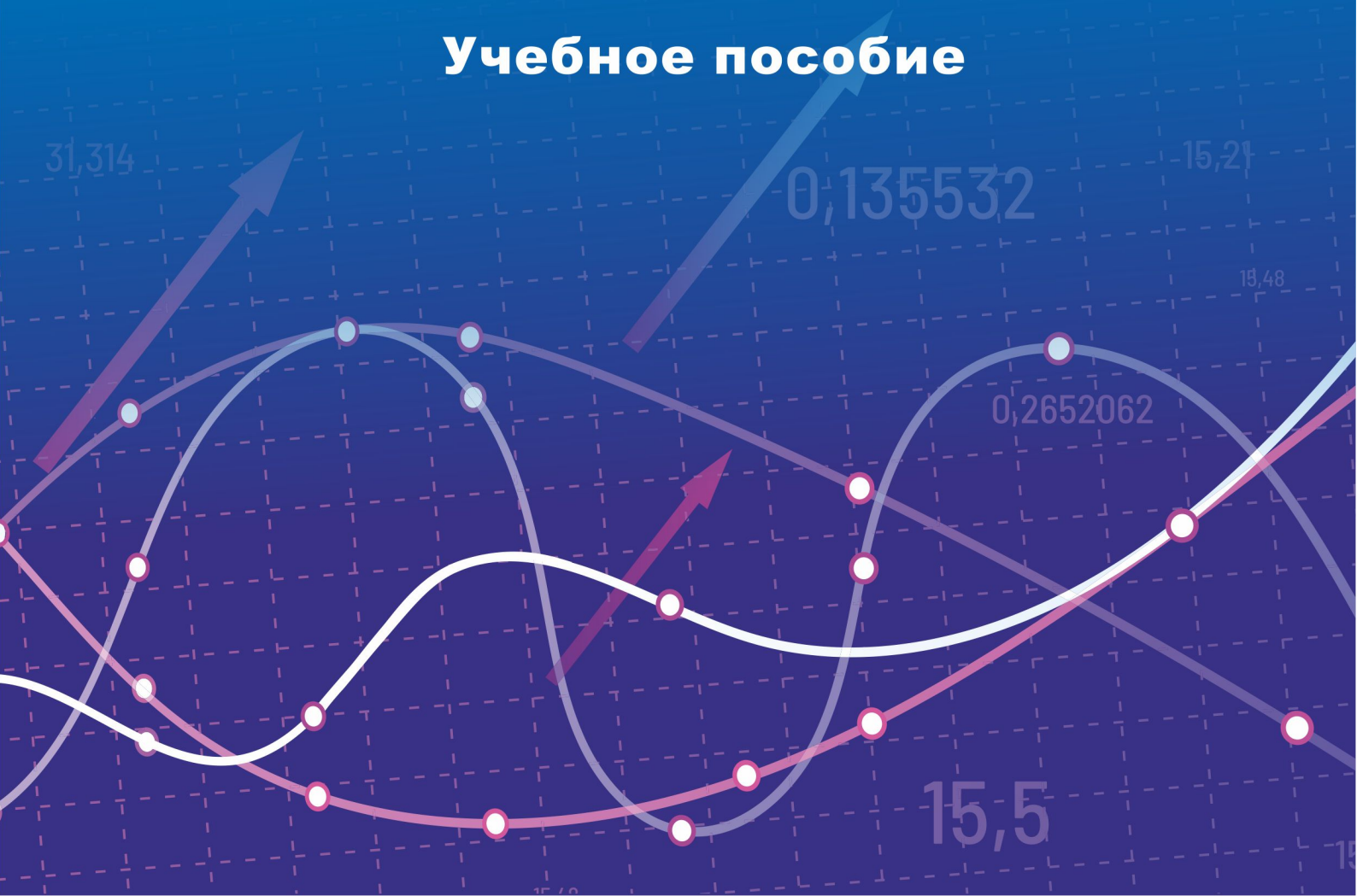


М. А. Тынкевич
А. Г. Пимонов
Я. В. Славолюбова

ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ (ТЕОРИЯ И ПРАКТИКА)

Учебное пособие



МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Кузбасский государственный технический университет
имени Т. Ф. Горбачева»

М. А. Тынкевич А. Г. Пимонов Я. В. Славолюбова

ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ (ТЕОРИЯ И ПРАКТИКА)

Учебное пособие

Кемерово 2021

УДК 519.23:004.942(075.8)

РЕЦЕНЗЕНТЫ

А. М. Гудов, доктор технических наук, доцент, директор Института фундаментальных наук федерального государственного бюджетного образовательного учреждения высшего образования «Кемеровский государственный университет»

Кафедра вычислительной математики и компьютерного моделирования федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет» (заведующий кафедрой *А. В. Старченко*, доктор физико-математических наук, профессор)

Тынкевич Моисей Аронович

Введение в статистический анализ данных (теория и практика) : учебное пособие / М. А. Тынкевич, А. Г. Пимонов, Я. В. Славолюбова ; Министерство науки и высшего образования Российской Федерации, Кузбасский государственный технический университет имени Т. Ф. Горбачева. – Кемерово, 2021. – 158 с. – ISBN 978-5-00137-246-2. – Текст : непосредственный.

Учебное пособие содержит теоретические основы дисциплины «Статистический анализ данных», дает представление о задачах и методах статистического анализа как в экономических исследованиях, так и в других областях науки и техники. Уточняются классификация случайных величин, их основные характеристики и способы получения таковых для эмпирического распределения, приводится информация о наиболее часто используемых распределениях и основных критериях согласия, рассмотрены методы корреляционного и регрессионного анализа, представляющие наибольший интерес в практических приложениях, и возможности, предоставляемые табличным процессором MS Excel и системой MatLab, кластерный анализ и анализ временных рядов. В пособии представлены практические задания и методические материалы для их выполнения.

Пособие предназначено для обучающихся по направлениям подготовки 09.03.03, 09.04.03 и 09.06.01, изучающих дисциплины «Статистический анализ данных», «Системы статистического анализа данных», «Статистический анализ результатов вычислительных экспериментов».

Печатается по решению редакционно-издательского совета Кузбасского государственного технического университета имени Т. Ф. Горбачева.

УДК 519.23:004.942(075.8)

Оглавление

Светлой памяти старшего товарища, коллеги, Учителя посвящается

ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Как много жизни, полной пыла, страстей и мысли, глядит на нас со статистических таблиц.
И. Ильф, Е. Петров

Каждый отдельно взятый человек – неразрешимая головоломка, но обо всех людях вместе можно говорить с математической точностью. Люди меняются – проценты остаются.
Артур Конан Дойл

Что скрывается под термином статистика? Обратившись к поисковым системам Интернета, вы получите многотысячный список файлов, где упоминается это сочетание букв.

Всезнающая Википедия под этим термином (лат. *Status* – состояние дел) понимает отрасль знаний, занимающуюся общими вопросами «сбора, измерения и анализа массовых *статистических* (количественных или качественных) данных». Такое определение называют рекурсивным (определение объекта прямо или опосредованно через себя; объект является частью самого себя), и оно вполне приемлемо, если читатель осознает смысл понятия «статистические данные». Там же узнаем, что этот термин ввел немецкий ученый Готфрид Ахенвалль (1746 г.) вместо преподававшегося в университетах Германии курса «Государствоведение»(?!).

Математическая энциклопедия [8] говорит, что «статистика – термин, употребляемый в математической статистике для названия функции от результатов наблюдений».

Другое определение: «статистика (в узком смысле) – измеримая числовая функция от выборки, не зависящая от неизвестных параметров распределения», что не требует опровержения, но у непосвященного порождает комплекс неполноценности.

Согласно мнению известного философа-материалиста Бенедикта (Баруха) Спинозы (1632–1677), «всякое определение есть ограничение». Подчас лучше в пограничных науках вообще не давать категорических определений, что и делают по отношению к таким аксиоматическим понятиям, как *множество*, *точка* и пр.

По И. Канту (1724–1804), «**в каждой естественной науке заключено столько истины, сколько в ней математики**».

Это лестное для математиков заявление не отвергается физиками, привыкшими создавать или использовать математические модели физических явлений, избранными представителями иных естественных наук: химии, эпидемиологии, биологии и геологии. Даже отдельные представители «неестественных наук» (социологи, специалисты по рекламе и т. п.) вынуждены прибегать к математическому моделированию (юмористы в свое время делили науки на *точные*, *естественные*, *сверхъестественные* и *противоестественные*).

Прикладная статистика не существует без разумного, *целенаправленного* и *систематизированного сбора* и хранения информации. Статистик не может уподобляться Плюшкину с его «веревочкой, которая в хозяйстве пригодится».

Понятие статистики неотделимо от требования **массовости**. Курьезы типа двухголовых телят, рекордного удоя конкретной коровы или успехов в выращивании цитрусовых за полярным кругом для нее не представляют интереса.

Другое требование к статистике связано с ее **репрезентативностью** – соответствием характеристик конкретной выборки характеристикам так называемой генеральной совокупности в целом.

Едва ли разумно смешивать в одной базе статистических данных потребность в обуви в странах Западной Европы и на островах Полинезии, спрос на бриллианты у обитателей Рублевки и Енисейска или потребление мандаринов в Абхазии и Магадане. Все подобные различия очевидны и носят отнюдь не случайный характер, хотя и для кого-то любопытны. В итоге подобной свалки находят «среднюю температуру по больнице» и «среднюю посещаемость театров по стране», из телепередач и социальных сетей доверчивый зритель узнает о росте или ухудшении благосостояния народа (в зависимости от пожеланий политолога).

Истоки статистики теряются в глубине веков. Несомненно, строительство египетских пирамид сопровождалось учетом численности и потребления работающих. В Древнем Китае статистический учет проводился при переписи населения, а в Древнем Риме вели учет поступления зерна из Египта и численности легионов. С появлением подобия государственных учреждений ведется учет имущества граждан для налогообложения.

Эти данные не всегда нуждались в длительном хранении, и сегодня они служат предметом интереса лишь для историков. Но сама *проблема носителей информации* оставалась (резать на камне долго и дорого, а дерево способно гореть или гнить). Жизнь папирусных свитков **Александрийской библиотеки** и берестяных новгородских грамот по разным причинам была недолгой. Только шумеры, обитавшие в третьем

тысячелетии до н. э. в междуречье Тигра и Евфрата, оставили нам достаточно большую хозяйственную информацию, поскольку использовали глиняные таблички.

Пергамент (материал из сыромятной кожи животных, пришедший из малоазиатского Пергама в Грецию) был дорог, в средние века немногие носители грамотности в тиши монастырей счищали с него труды Аристотеля, Евклида и прочих «еретиков» и заменяли «писаниями святых отцов». Даже с появлением «дешевой» бумаги проблема сохранности информации не нашла идеального решения – увы, рукописи горят и вместимость библиотек не безгранична.

Появление ЭВМ и последовавшее фантастическое расширение емкости их памяти в сочетании с аналогичным ростом быстродействия породили гигантские базы данных, хотя и не устранили проблемы их сохранности и доступности. Несколько уменьшились темпы роста потребления бумаги (вырубки лесов) в сфере производства, управления и планирования, но поток директивных методических указаний в образовании, медицине и т. д. на бумажных и электронных носителях не иссякает. Как говорили в старину, знающий преподает, знающий и умеющий работает, не знающий и не умеющий пишет для остальных методические указания.

Человеческий фактор в статистике был и остается помехой для последующих исследований. Марк Твен приписывает английскому премьеру Бенджамину Дизраэли крылатое выражение «Существуют три вида лжи: ложь, наглая ложь и статистика».

Из безобидной статистики размеров выловленных рыб, по словам рыболовов, едва ли можно судить о водных богатствах морей. Преподаватель с удивлением узнает, что реформа образования способствовала улучшению качества образования и повышению интеллектуального уровня учащихся. Искусный водитель грузовика корректирует показания счетчика, дорожный строитель докладывает о выполнении плана на 100 %. Глава администрации в рапорте об успехах региона под его чутким руководством приписывает или теряет нуль. Причины таких явлений многообразны, равно как и их последствия.

Даже при отсутствии целенаправленных искажений в статистических данных возникают естественные ошибки из-за ограниченной точности измерительного инструмента. Нелепо измерять расстояние от Земли до Альфа Центавра с точностью до километра, а численность населения страны с точностью до человека. Преподаватель технического вуза при 4-балльной системе оценок ставит «неуд» лентяю в надежде, что принудит его к познанию (нарушение прав человека?), и юноше, решающему уравнение $5 \times x = 13$ в виде $x = 8$.

Математический аппарат статистического анализа выводов далеко выходит за пределы арифметики.

Его основой стала теория вероятностей, родившаяся в XVIII веке фактически по запросам одержимых разгадкой таинства игры. Постепенно возникла математическая статистика как самостоятельная научная дисциплина, востребованная для экономического анализа и завоевавшая сферу контроля качества массовой продукции в промышленности.

Традиционное единство физиков и математиков породило теорию случайных процессов и *статистическую физику*. Навеки в статистической физике сохранены имена великих физиков – статистики Бозе – Эйнштейна, Максвелла – Больцмана, Ферми – Дирака и др.

ЭВМ с их быстродействующими датчиками случайных чисел сделали реальным применение методов статистических испытаний (Монте-Карло) для *математического моделирования* технологических процессов и массового обслуживания.

Методы статистического анализа проникли в геологию, метеорологию, астрофизику и биологические науки.

Статистический анализ традиционно используется в социологии и политологии, но трактовка полученных выводов здесь часто настолько субъективна, что вспоминаются слова английского философа Томаса Гоббса: «Если бы геометрические аксиомы задевали интересы людей, они бы опровергались».

Аппарат специальных ответвлений статистического анализа связан со спецификой информации, и специалист по распознаванию образов едва ли найдет общий язык с исследователем потоков быстрых нейтронов для защиты реакторов. Даже в определения самих понятий математик и экономист вносят субъективные подходы. По Л. В. Канторовичу, «*эконометрика* – достаточно обширная область математики (математическое программирование, комбинаторный анализ, теория вероятностей и математическая статистика, теория графов и др.), для которой экономика является полигоном для проявления своих возможностей». Экономист считает эконометрику областью экономики, лишь допускающей применение аппарата математической статистики.

Каковы бы ни были сферы применения статистического анализа, их объединяет общность фундаментальных понятий: *случайная величина*, *распределение вероятностей*, *статистическая гипотеза*, *корреляция*, *случайный процесс*.

Статистический анализ имеет двоякое назначение. Он дает *осведомляющую* информацию о природе явлений, неподвластных регулированию, и *управляющую* информацию, допускающую регулирование и используемую сегодня в реальной практике жизни. Установив из ста-

тистики испытаний некоторого изделия, что некая добавка не улучшает его характеристики, попросту можем исключить ее из технологической цепочки. Познание статистики космических излучений способствует улучшению защиты космического корабля, но регулировать их мы не в силах.

Умение моделировать случайные величины с заданными характеристиками и имитировать соответствующие случайные процессы минимизирует потребность в физическом эксперименте (в экономике и социологии реальный эксперимент не всегда возможен).

Возможно, что выпускнику вуза – специалисту по информационным технологиям суждено всю жизнь заниматься прозаической защитой информации и сопровождением баз данных в каком-то банке или медицинском учреждении, бухгалтерским учетом или менеджментом в супермаркете (по Маяковскому, «все работы хороши, выбирай на вкус») и встречаться лишь с бытовыми случайностями, не задумываясь о законах их распределения. Тем не менее, столкнувшись с исследовательской работой в банковской сфере, рекламой памперсов или кандидата в мэры многомиллионного города, контролем качества хлебобулочных изделий, установкой светофоров в городе и т. п., на задворках памяти можно найти подсказку к каким-то полезным решениям.

Кануло в Лету время, когда встречались так называемые *эрудиты* – специалисты в нескольких областях науки и техники. Джон фон Нейман (1903–1957) создал теорию автоматов (идеологию компьютерной техники) и теорию игр, генерировал идеи методов исследования операций, сделал значимый вклад в квантовую физику и функциональный анализ. Блез Паскаль (1623–1662) известен работами по теории вероятностей и проективной геометрии, создал первый арифмометр, сформулировал основной закон гидростатики. Уровень развития современной науки ныне таков, что профессионализм такого рода едва ли реален. Еще 150 лет назад известный Козьма Прутков заявлял, что «никто не может объять необъятное».

В не отличающемся деликатностью обществе специалистов или считающих себя таковыми, часто звучит оскорбительное слово *дилетан*т в адрес человека, для которого работа в смежной области знания просто доставляет удовольствие. В ответ тот же Козьма Прутков говорил, что односторонний «специалист подобен флюсу». Для дилетантов характерно желание подойти к проблеме со стороны. Именно дилетантами получены выдающиеся результаты в смежных областях наук. Насколько приятнее и полезнее общение с человеком, сочетающим профессионализм в сфере информационных технологий с дилетантизмом в истории, географии, живописи и т. п.

<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>И сотворил Бог
Растения</td><td>И сотворил Бог
Животных</td></tr>
<tr><td>«Плодитесь
и размножайтесь»
…
И, ужаснувшись
начавшемуся хаосу, создал
СТАТИСТИКУ,
и для ведения оной
по своему образу и подобию
создал
ЧЕЛОВЕКА</td><td> 
Похоже, что мне суждено
расхлебывать эту кашу</td></tr>
</table>

Сотворение мира и человека
(рисунки Ж. Эффеля, комментарии старшего из авторов)

«Мыслители» наших дней убеждают молодежь в ненужности каких-то знаний – «все можно найти в Интернете» (вообще «день поступления в вуз считать днем его окончания»). Они не понимают, что Интернет не даст приемлемого ответа на запрос от человека, не имеющего

базовых знаний. Приходится констатировать, что за последние годы компьютер, созданный как помощник в интеллектуальной жизни человека, способствовал развитию интеллекта одних и интеллектуальному среднестатистическому убожеству других.

Предлагаемое учебное пособие не ставит цель сделать читателя крупным специалистом в океане статистического анализа. Решение этой задачи непосильно даже для многотомных монографий и справочников, основанных на многолетнем опыте работы поколений исследователей в прикладных сферах. Как и в большинстве учебных руководств, сначала рекомендуется прочесть «по диагонали», ограничившись знакомством с идеями и понятиями, не вдаваясь в детали, и лишь затем приступить к более внимательному чтению.

Прилагаемый практикум выделяет необходимый минимум познаний о методах статистического анализа и базу осмысления результатов возможного обращения к популярным пакетам.

Естественно, от читателя потребуется знание математики в пределах средней школы уровня 60-х годов (понятие о функции вообще и поведении элементарных функций, знание нескольких букв древнегреческого и латинского алфавитов, понимание термина *решить уравнение*, азбучные истины о дифференцировании и интегрировании).

P. S. В систематизации изложенного материала принимали участие в своих выпускных дипломных работах разработчики пакета статистической обработки экономической информации (СтЭк) студенты КузГТУ Е. И. Латышева (2004 г.), О. С. Болотова (2004 г.) и Ю. С. Кирина (2006 г.), за что авторы приносят им искреннюю благодарность.

Глава 1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

1.1. Терминология и классификация

Вдумайтесь в смысл термина *случайная величина*. Упоминание о случайности здесь едва ли нуждается в комментарии, но понятие *величины* в русском языке так или иначе ассоциируется с некой метрикой – количественной оценкой. В этом смысле оно применимо в оценке высоты дома и полета вертолета, потребляемых калорий или уровня воды в водохранилище, но едва ли применимо к национальной принадлежности, многообразию цветов, фруктов и темпераменту человека. Более близок к сути дела англоязычный эквивалент *random variable* (случайная переменная), ограничивая суть явления лишь факторами случайности и изменчивости, и ниже речь пойдет о случайных величинах именно в этом смысле.

С точки зрения аппарата статистического анализа имеет смысл различать три основных типа (шкалы) таких случайных величин:

1) *номинальные* (шкала наименований) – для качественной классификации таких данных, как адрес, пол, национальность и т. п., допускающих только вопрос «сколько таких?»;

2) *ранговые* (порядковые, относительные) – для характеристики степени обладания некоторым качеством (отлично – хорошо – плохо, социальное положение, система приоритетов), не отвечая на вопрос «*насколько* лучше?»;

3) *абсолютные* – численные оценки с наличием точки отсчета и масштаба.

Различают *дискретные* и *непрерывные* случайные величины.

Едва ли в природе можно найти что-то непрерывное. Всякая прямая линия на бумаге под микроскопом выглядит как множество точек, расположенных на прямой отнюдь не идеально. Солнечная энергия, рентгеновское и радиоизлучение, естественные процессы в природе имеют квантовую природу. Оценкой труда выступает денежная компенсация, исчисляемая в России с точностью до копеек (ушли в прошлое анекдотические квитанции на оплату 0,00 руб., возникающие как плод некачественного программирования). Непрерывные величины выступают в нашем сознании как идеализация дискретности, обеспечивая получение каких-то универсальных оценок с минимальными затратами труда.

Очевидно, что в компьютерный век вместо хранения больших таблиц, отражающих функциональную связь между какими-то объектами, разумнее подобрать аналитическое представление функции, определяющей изучаемую связь с достаточной точностью, чем уже 200 лет зани-

мается прикладная математика, создавая методы аппроксимации данных с различными методами оценки ее качества.

Уточним понятия *генеральной совокупности* и *выборки* из нее.

Генеральная совокупность понимается как символ многообразия возможных проявлений случайной величины (продолжительность разговоров по телефону, вулканическая активность или допустимые комбинации цифр в лотерее 6 из 36). То есть это множество (оно может быть и бесконечным) всех возможных значений, принимаемых случайной величиной из заданного распределения.

Случайная выборка – совокупность конечного множества элементов из некой генеральной совокупности, каждый из которых имеет равные шансы быть отобранным и попасть в выборку.

Конкретная выборка ограниченного объема дает возможность оценить свойства всей генеральной совокупности. Она должна быть репрезентативной, чтобы эти оценки можно было принимать или отвергать с какой-то достаточно высокой степенью доверия.

1.2. Основные характеристики случайных величин

Ограничимся здесь рассмотрением только абсолютных величин, относительно простых с точки зрения статистического анализа, и договоримся о терминологии.

Если случайная величина X принимает дискретные значения x_i ($i = 1, n$) (например, $x_1 \le x_2 \le ... \le x_n$), то вероятность выбора конкретного значения $P(X = x_i) = p_i$ подчинена условиям

$$\sum_{i=1}^{n} p_i = 1;\ p_i \ge 0\ (i = 1, 2, \ldots, n) \qquad (1.1а)$$

(никаких других исходов быть не может).

Если X принимает любые значения x в диапазоне $[\alpha, \beta]$ с вероятностями $p(x)$, то естественно заменить суммирование интегрированием и представить условия в виде

$$\int_{\alpha}^{\beta} p(x)dx = 1;\ p(x) \ge 0,\ x \in [\alpha, \beta]. \qquad (1.1б)$$

В том и ином случаях представляет интерес *распределение вероятностей* (шансов на получение значений случайной величины), и возникает понятие так называемой *функции плотности распределения вероятностей*, представимой в виде табличной или аналитически заданной функции $p(x)$, которую в литературе иногда обозначают как $f(x)$.

Соответственно определяется и монотонно возрастающая *функция распределения вероятностей* (вероятность того, что значение случайной величины не превышает какого-то порога):

$$F_k = P(X \le x_k) = \sum_{i=1}^{k} p_i \; ; \; F(X) = P(X \le x) = \int_{\alpha}^{x} p(y)dy \qquad (1.2)$$

Приведенные ниже рис. 1.1 и 1.2 иллюстрируют эти понятия для непрерывной и дискретной случайных величин.

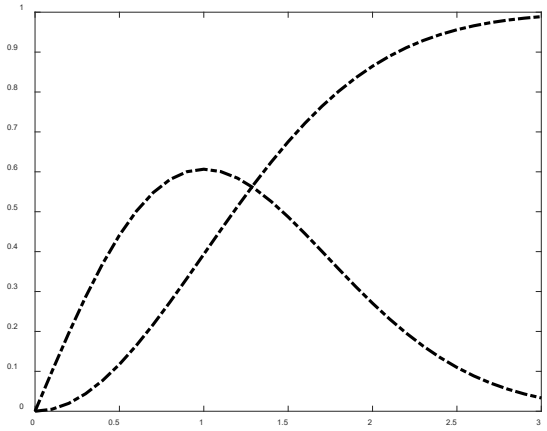


Рис. 1.1. Плотность и функция непрерывного распределения

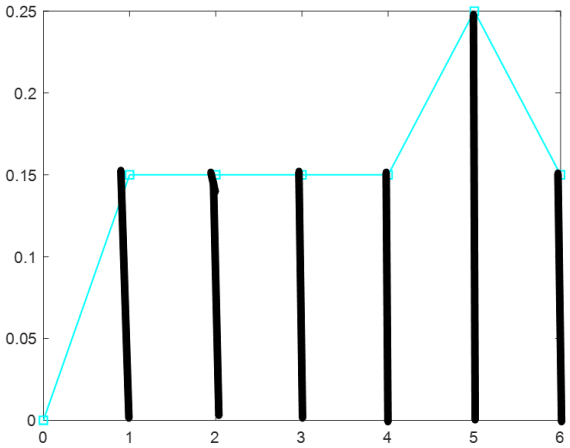


Рис. 1.2. Плотность дискретного распределения (фальшивая кость)

Одними из основных характеристик распределения являются *математическое ожидание* (момент первого порядка, среднее взвешенное значение, ожидаемое значение и т. п.)

$$Mx = \mu = \sum_{i=1}^{n} x_i p_i \; , \; Mx = \mu = \int_{\alpha}^{\beta} xp(x)dx \qquad (1.3)$$

и так называемые *центральные моменты* более высоких порядков (относительно среднего)

$$M_k x = \sum_{i=1}^{n} (x_i - \mu)^k p_i \; , \; M_k x = \int_{\alpha}^{\beta} (x - \mu)^k p(x)dx \, , \qquad (1.4)$$

используемые при поиске некоторых полезных свойств распределения.

Центральный момент второго порядка определяет *дисперсию* распределения – характеристику суммарного разброса значений случайной величины относительно среднего значения

$$Dx = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p_i \; , \; Dx = \sigma^2 = \int_{\alpha}^{\beta} (x - \mu)^2 p(x)dx. \qquad (1.5)$$

Чаще вместо дисперсии используют величину

$$\sigma = \sqrt{Dx}, \tag{1.6}$$

называемую *среднеквадратическим* или *стандартным отклонением*.

Если смысл математического ожидания интуитивно понятен, то значение σ без учета диапазона значений случайной величины не говорит о чем-либо интересном. При стрельбе на расстояние порядка 1000 м отклонению от цели σ = 1 можно лишь позавидовать, но при хирургическом вмешательстве подобная ошибка трагична.

Значения случайных величин x_i могут иметь порядок $10^{-17} \div 10^{19}$. Соответственно, при компьютерной обработке даже простой арифметический реальный расчет (типа возведения в квадрат) может сопровождаться прерываниями типа *overflow* (переполнение) и машинными нулями, не говоря о потере точности.

Знание значений μ и σ позволяет создавать более удобные в анализе *стандартизованные* (*центрированные и нормированные*) значения

$$Z = \frac{x - \mu}{\sigma} \tag{1.7}$$

с нулевым математическим ожиданием и единичным среднеквадратическим отклонением.

По чисто вычислительным соображениям работа со стандартизованными величинами минимизирует погрешность вычислений, уменьшает шансы на упомянутые выше потерю значности при умножении малых или переполнение при умножении больших значений.

Иногда используется *стандартная ошибка среднего*

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}. \tag{1.8}$$

Момент третьего порядка определяет *асимметрию* A_x распределения (левую или правую) – меру несимметричности распределения относительно среднего (идеал $A_x = 0$ для нормального распределения):

$$A_x = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^3 p_i, \; A_x = \int_{\alpha}^{\beta} \left(\frac{x - \mu}{\sigma} \right)^3 p(x) dx. \tag{1.9}$$

Момент четвертого порядка определяет *эксцесс* E_x распределения – меру сглаженности (остроты пика плотности распределения) – в виде

$$E_x = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^4 p_i, \; E_x = \int_{\alpha}^{\beta} \left(\frac{x - \mu}{\sigma} \right)^4 p(x) dx. \tag{1.10}$$

При ненулевых μ используют *коэффициент вариации* по Пирсону

$$V_x = \frac{\mu}{\sigma} \qquad (1.11)$$

как безразмерный показатель вариабельности случайной величины (иногда используют обратное представление).

1.3. Понятие о нормальном распределении

Самым распространенным явлением в деятельности человека являются случайные ошибки.

Если мастер на стройке предложит своим подчиненным с помощью подручных средств измерить ширину проема двери и отправить ему результаты измерения, то обнаружит отклонения от проекта. Обычно отклонения незначительны и зависят от возможностей измерительного инструмента, но могут обнаружиться и существенные, определяемые темпераментом измерителя, его близорукостью и даже днем получения зарплаты. Стрельба из некоторого «изделия» сопровождается *эллипсом рассеивания* относительно цели. На рынке вам взвесят арбуз с тем или иным отклонением от истины. Напряжение в сети электроснабжения колеблется относительно стандарта и т. д.

Появление подобных отклонений естественно, и не случайно их распределение вероятностей называют *нормальным*.

Идеал этого распределения описывается формулами

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \; F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx. \qquad (1.12)$$

Графики плотности и функции нормального распределения представлены на рис. 1.3 и 1.4 соответственно.

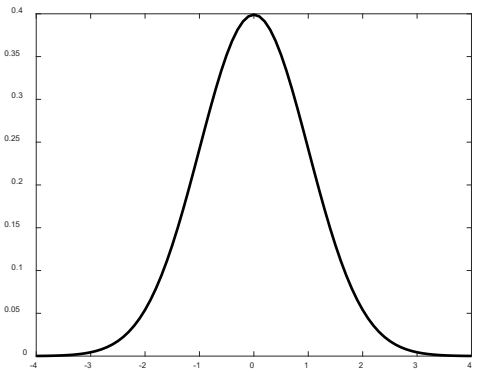


Рис. 1.3. Плотность нормального распределения (μ = 0, σ = 1)

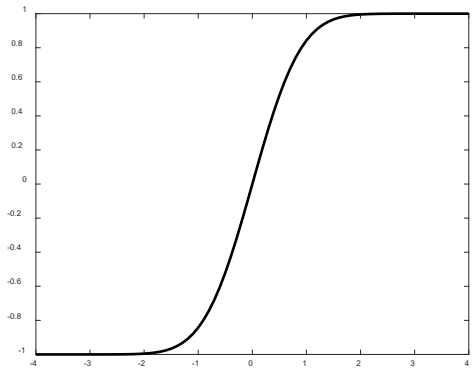


Рис. 1.4. Функция нормального распределения (μ = 0, σ = 1)

Следует обратить внимание на симметричность плотности распределения (вероятности отклонения от среднего в обе стороны равны), уменьшение вероятности больших отклонений и ничтожную вероятность отклонений по модулю более чем на 3σ. Единственный недостаток нормального распределения в том, что функция распределения представлена *неберущимся интегралом*, но с помощью преобразования (1.7) и учета симметрии

$$F(x) = 0{,}5 + \Phi(x) = 0{,}5 + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}x^2}\, dx.$$

Фигурирующий здесь интеграл, называемый функцией Гаусса или интегралом ошибок, имеет отличные аппроксимации и включен в библиотеки стандартных функций множества систем программирования.

Что касается эталонных (идеальных) асимметрии и эксцесса, то они равны соответственно 0 и 3 (к этим характеристикам вернемся при последующем знакомстве с распределением Лапласа – Шарлье).

Большинство распределений в классической эконометрике и вообще в статистическом анализе так или иначе связано с нормальным. В предположении асимптотической нормальности строятся все критерии оценки параметров распределений и близости их характеристик. Большинство физических законов получено в результате усреднения итогов массовых экспериментов так, чтобы сумма квадратов отклонений экспериментов от ожидаемого идеала была минимальна (метод наименьших квадратов).

1.4. Характеристики эмпирических распределений

Хорошо, если для объекта исследования известны закон распределения вероятностей или хотя бы математическое ожидание и стандартное отклонение, но для реальных процессов это нереально.

Приходится строить *эмпирические распределения* на основе выборки, состоящей из N (*объем выборки*) элементов с неизвестными (непредсказуемыми) вероятностями их появления.

По *принципу недостаточного основания* вероятность появления очередного значения принимается равной $1/N$ и вычисление оценок для выборочного распределения сводится к поиску:

$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,,\ \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2\,,$$

$$A_x = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3;\ E_x = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4. \qquad (1.13)$$

В статистике (особенно в случае малых выборок) существенно понятие *числа степеней свободы*.

По аналогии с теоретической механикой предполагаем, что исходная выборка обладает $f = N$ степенями свободы (своеобразное облако случайных точек). Зафиксировав μ, уменьшаем число степеней свободы $f = N - 1$ (привязываем центр диапазона к указанному месту). С учетом найденной оценки σ уменьшаем f, и все последующие оценки, базирующиеся на μ и σ, ищем при $f = N - 2$.

Соответственно получаем так называемые *несмещенные оценки*:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2,$$

$$A_x = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3,\ E_x = \frac{N(N+1)}{(N-1)(N-2)(N-3)}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4. \qquad (1.13a)$$

При больших N эти корректуры не представляют интереса.

1.5. Медиана, мода и квантили

Мода (*Mo*) соответствует такому значению случайной величины, при котором функция плотности распределения вероятностей достигает максимума. Большинство известных распределений является *унимодальными* (функция плотности распределения имеет единственную точку максимума). Поиск моды в этом случае не составляет особых сложностей.

Если максимумов два, распределение называют *бимодальным* (рис. 1.5), в общем случае (при нескольких максимумах) – *полимодальным*.

Медиана (*Me*) делит распределение на две равновероятные половины:

$$\int_{\alpha}^{Me} p(s)ds = \int_{Me}^{\beta} p(s)\,ds. \qquad (1.14)$$

По определению медианы (1.14) упрощается до равенства

$$\int_{\alpha}^{Me} p(s)ds = 0{,}5. \qquad (1.14a)$$

Для нормального распределения очевидно совпадение математического ожидания, моды и медианы.

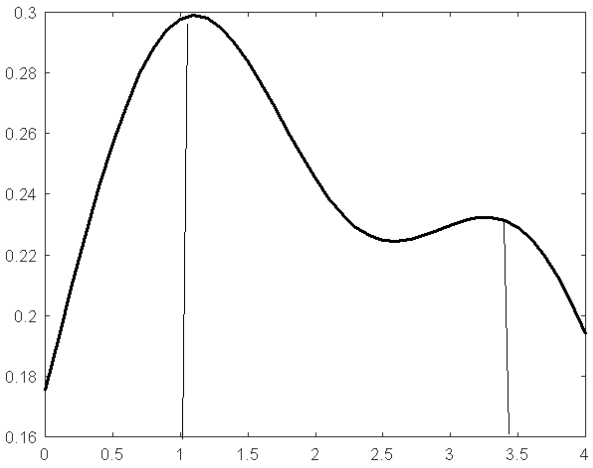


Рис. 1.5. Бимодальное распределение

Если обратиться к распределению

$$p(x) = \lambda\, e^{-\lambda x},$$

где $\lambda = 1 / \mu$, $x > 0$, интегрированием (1.14а) получаем

$$Me = \ln(2) / \lambda = 0{,}693\ \mu.$$

Для дискретных распределений достаточно упорядочить выборку и приять медиану, равной срединному значению $x_{[N/2]+1}$ при нечетном N или полусумме $(x_{[N/2]} + x_{[N/2]+1}) / 2$ – при четном.

При проверке статистических гипотез важнейшей является характеристика распределения, связанная с *оценкой степени доверия* и называемая *квантилью* (рис. 1.6). Так для случайной величины с функцией распределения $F(x)$ квантилью порядка α $(0 < \alpha < 1)$ называется максимальное по модулю число K_α такое, что $F(K_\alpha) \le \alpha$.

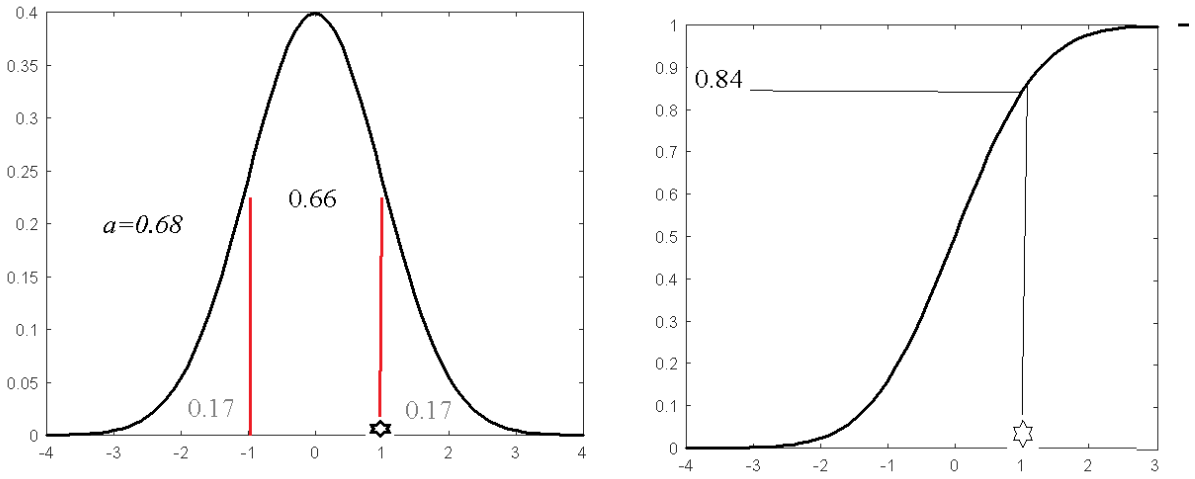


Рис. 1.6. Двусторонняя квантиль для нормального распределения $N(0,1)$

В случае непрерывной функции распределения поиск квантили сводится к решению уравнения $F(x) = \alpha$:

$$\int_{-\infty}^{K_\alpha} p(x)dx = \alpha \text{ или } \int_{K_\alpha}^{\infty} p(x)dx = 1-\alpha. \qquad (1.15)$$

В общем случае, если вероятность некоторого явления не превышает α, то нет оснований его отрицать. В противном случае с вероятностью ошибки $1 - \alpha$ считаем, что это явление неприемлемо.

Для распределений, подобных нормальному, где область значений случайной величины не ограничена, область недоверия распадается на две равные части (рис. 1.6), и говорят о двусторонней квантиле. Очевидно, что квантиль порядка 0,5 совпадает с *медианой* распределения.

Так, если известны математическое ожидание потребления картофеля жителями Кузбасса и соответствующее стандартное отклонение, то нормировкой имеющихся данных получаем классическое нормальное распределение ($\mu = 0$, $\sigma = 1$).

Задавшись вероятностью уверенности $\alpha = 0{,}95$, находим соответствующую квантиль:

$$\int_{-\infty}^{K_{0,95}} p(x)dx = 0{,}95 \quad \text{или} \quad \int_{K_{0,95}}^{\infty} p(x)dx = 1 - 0{,}95 = 0{,}05,$$

чтобы с 5%-й вероятностью быть уверенным в разумности создаваемого запаса $\mu + \sigma K_\alpha$ (существует 2,5%-я вероятность нехватки запаса и такая же вероятность его избытка).

У читателя может создаться впечатление, что подобные задачи решаются элементарно (щелкнул мышкой, и, как у гоголевского Пузатого Пацюка, «вареник» сам в рот заскочит) или наоборот требуют фантастических усилий. Ни то, ни другое. Естественно, нужна достаточная статистика, на базе которой находят оценки математического ожидания и стандартного отклонения и строят эмпирическое распределение. Оценив параметры, сопоставляют найденному эмпирическому распределению «подозреваемые» среди известных непрерывных распределений. Выбрав по определенному критерию ближайшее теоретическое, соответствующее эмпирическому, можем найти упомянутую квантиль и дать рекомендации для реального планирования. С этой относительно простой процедурой читатель познакомится ниже в главе 3, посвященной проверке статистических гипотез.

1.6. Закон больших чисел и объем выборки

Ясно, что чем больше объем статистической выборки, тем надежнее наши суждения об изучаемом явлении, но неограниченное его увеличение не всегда возможно. Оценка объема выборки в статистическом анализе базируется на *законе больших чисел*.

В современной интернет-литературе, переполненной плагиатом (за малым исключением), гуляет анекдотическое утверждение, что этот закон сформулирован швейцарским математиком Якобом Бернулли (1655–1705) в 1713 году. К огорчению для мистиков, твердящих о «жизни после смерти», это произошло при жизни автора, не успевшего подготовить публикацию к печати. В 1713 году она была издана посмертно его братом Николаем под названием «Искусство предположений».

Якоб Бернулли, один из творцов теории вероятностей, показал, что при числе независимых испытаний N и числе «успехов» M вероятность P одиночного успеха можно оценить из условия

$$P\left(\left|\frac{M}{N} - p\right| \le \varepsilon\right) > 1 - \eta$$

при достаточно больших N, точности ε и вероятности ошибки $\eta > 0$.

Отсюда следует, что N достаточно выбрать из условия

$$N > \frac{1+\varepsilon}{\varepsilon^2} \lg\frac{1}{\eta} + \frac{1}{\varepsilon}.$$

Так при выборе $\varepsilon = 0{,}01$ и $\eta = 0{,}05$ это неравенство дает оценку $N \approx 13\,240$.

Более простая, но завышенная оценка $N > \frac{1}{\varepsilon^2 \eta}$ получена П. Л. Чебышёвым (1846 г.). Для нашего примера $N = 200000$.

Обе оценки полезны при моделировании случайных процессов и в численном анализе при решении некоторых многомерных задач.

1.7. Описательная статистика в MS Excel

В MS Excel для статистического анализа данных имеется надстройка «Пакет анализа» (Data Analysis) и широкий набор статистических функций рабочего листа.

Рассмотрим получение оценок основных параметров распределения (описательной статистики) на примере производства картофеля (рис. 1.7) в Российской Федерации (миллионы тонн).

Выполнив команду `Данные – Анализ данных – Описательная статистика`, при уровне значимости $\alpha = 0{,}05$ заполним параметры диалогового окна (рис. 1.8).

В результате получим первые два столбца табл. 1.1, в третьем столбце нами указаны принятые для статистик обозначения.

	A	B
1	**Год**	**Производство картофеля, млн. тонн**
2	1990	31
3	1991	34
4	1992	38
5	1993	38
6	1994	34
7	1995	40
8	1996	38
9	1997	35
10	1998	29
11	1999	28
12	2000	30
13	2001	30
14	2002	27
15	2003	29
16	2004	28
17	2005	28

Рис. 1.7. Фрагмент выборки

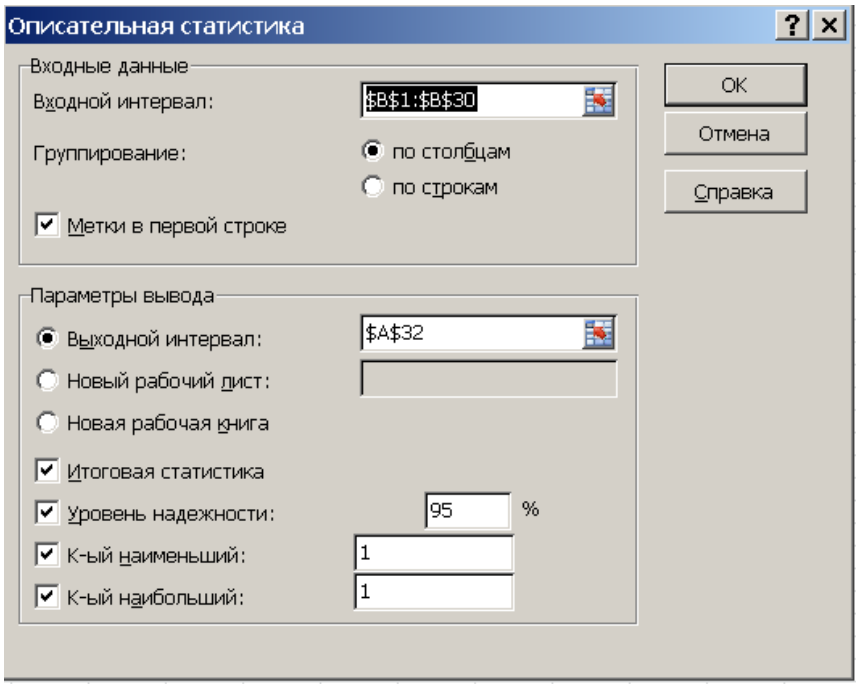


Рис. 1.8. Диалоговое окно описательной статистики

Таблица 1.1. Описательная статистика

Производство картофеля, млн т			**Принятые обозначения**
Среднее	28,93103448		$\bar{x} = \sum x_i n_i / n$
Стандартная ошибка	0,986274068		$\sigma_{\bar{x}} = \sigma / \sqrt{n}$
Медиана	28		M_e
Мода	28		M_o
Стандартное откло-нение	5,311248404		σ
Дисперсия выборки	28,20935961		$\sigma^2 = \sum (x_i - \bar{x})^2 n_i / (n-1)$
Эксцесс	– 0,275200691		E_x
Асимметричность	0,472992416		A_k
Интервал	21		$W = x_{\max} - x_{\min}$
Минимум	19		$x_{\min}$
Максимум	40		$x_{\max}$
Сумма	839		$\sum x_i$
Счет	29		$n = \sum n_i$
Наибольший (1)	40		–
Наименьший (1)	19		–
Уровень надежности (95,0 %)	2,020290846		$\Delta = t_{\alpha;n-1}\sigma_{\bar{x}}$

Контрольные вопросы

1) Какие типы шкал, используемых для измерения случайных величин, вам известны?

2) Для какого типа наблюдаемых случайных величин используется номинальная шкала?

3) В чем основное отличие порядковой шкалы измерений от номинальной?

4) Приведите примеры интервальной, относительной и абсолютной шкал. Какова специфика каждой из них?

5) Чем отличаются друг от друга «дискретные» и «непрерывные» случайные величины?

6) Каким образом связаны функции плотности и распределения вероятностей случайной величины?

7) В чем отличие «обычных» моментов и центральных моментов k-го порядка распределения случайной величины?

8) Каким образом математическое ожидание характеризует среднее значение случайной величины?

9) Мерой чего является дисперсия случайной величины?

10) Что характеризуют центральные моменты третьего и четвертого порядков?

11) В чем заключается принцип перехода от вычисления параметров распределения случайной величины к вычислению оценок для выборочного распределения?

12) Что такое число степеней свободы?

13) В чем отличие смещенных от несмещенных выборочных оценок параметров распределения случайной величины?

14) Каково определение моды и медианы для непрерывных и дискретных случайных величин?

15) Что такое квантиль?

16) Каким образом связаны квантиль и симметричность распределения?

17) Как используется закон больших чисел для оценки объема выборки?

18) Каким образом строится эмпирическое распределение случайной величины?

19) Что представляет собой нормальное распределение случайной величины?

20) Что представляют собой описательные статистики?

21) Какими средствами располагает MS Excel для вычисления описательных статистик?

Глава 2. РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

Распределение вероятностей – одно из основных понятий теории вероятностей и математической статистики. В приложениях используются как дискретные распределения (биномиальное, полиномиальное, Пуассона, геометрическое, Паскаля, Пойа и т. п.), так и непрерывные, среди которых наиболее популярно нормальное распределение, составляющее базу *теории ошибок*.

Как мы указывали выше, всякое распределение характеризуется *функцией плотности распределения* $p(x)$ и связанной с ней монотонно возрастающей *функцией распределения* $F(x) = P(z \le x)$ (вероятность того, что случайная величина z не превышает x, $0 \le F(x) \le 1$).

$$F(x) = \int_{-\infty}^{x} p(z)dz, \quad \int_{-\infty}^{\infty} p(z)dz = 1, \quad p(x) \ge 0.$$

С теоретической и прикладной точек зрения важно знать, к какому типу распределения можно отнести найденные случайные реализации некоторого явления. Это знание поможет построить имитационную модель явления со всеми вытекающими последствиями (возможностью прогноза, минимизацией количества последующих экспериментов и минимизацией затрат). Знание типа распределения приводит к пониманию правомерности применения тех или иных методов для исследования явления (можно получить абсурдные выводы, если распределение далеко от нормального).

2.1. Популярные непрерывные распределения

2.1.1. Равномерное распределение

Непрерывная случайная величина x называется равномерно распределенной на отрезке $[a, b]$, если ее плотность (рис. 2.1) определяется функцией

$$p(x) = \begin{cases} 1/(b-a), & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}.$$

Соответственно функция распределения

$$F(x) = \begin{cases} 0, & x < a \\ (x-a)/(b-a), & a \le x \le b \\ 1, & x > b \end{cases}.$$

Без особого труда можно оценить математическое ожидание

$$\mu = \int_{-\infty}^{\infty} x\, p(x)dx = \int_{a}^{b} x\, p(x)dx = \int_{a}^{b} x\,(b-a)^{-1}dx = (a+b)/2, \qquad (2.1)$$

совпадающее с медианой;
дисперсию

$$\sigma^2 = \int_{a}^{b} (x-\mu)^2\, p(x)dx = (b-a)^2/12 \qquad (2.1а)$$

и стандартное отклонение

$$\sigma = \sqrt{\sigma^2} = (b-a)/(2\sqrt{3}). \qquad (2.1б)$$

Асимметрия и эксцесс соответственно равны 0 и –6 / 5.

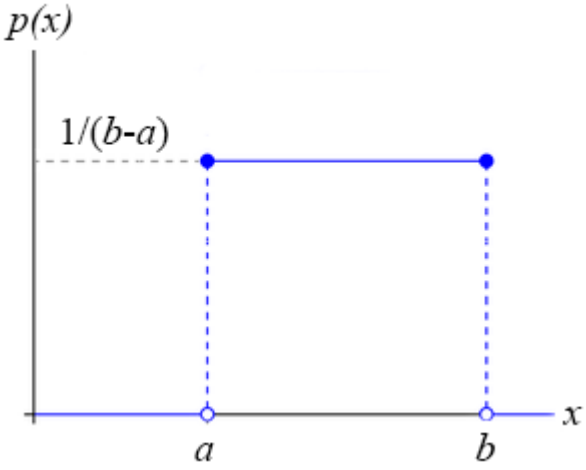


Рис. 2.1. Плотность равномерного распределения

Задача моделирования таких случайных величин, столь необходимых для имитации поведения какой-либо системы со случайными исходами и решения задач оптимального управления, возникла практически с появлением ЭВМ. С помощью датчиков равномерно распределенных случайных чисел можно непредвзято выбрать состав присяжных, разместить участников ЕГЭ в аудитории, найти значение 20-мерного интеграла.

Эти датчики служат базой применения методов Монте-Карло (методов случайных испытаний) при решении сложных задач оптимального планирования и управления. Создавать эффективные физические датчики было дорого, искали чисто компьютерные реализации *псевдослучайных* величин, отвечающие системе тестов на случайность.

Еще в 1946 году основоположник теории автоматов (автоматических вычислительных машин) Джон фон Нейман предложил выбирать некое *n*-значное число (лучше среди *простых* чисел), возводить его в квадрат, выделять *n* цифр в середине результата как следующее число, снова возводить в квадрат и т. д. Увы, этот изящный прием приводил к распределению создаваемых значений более близкому к нормальному.

В 1949 году американский математик Д. Г. Лемер предложил вместо квадрирования умножать очередное *n*-значное число на постоянный удачно подобранный множитель *L* и в итоге брать последние *n* знаков:

$$x_{k+1} = (x_k\, L) \bmod n,\ k = 0, 1, 2, \ldots .$$

Удачно выбранные начальное значение и множитель дают достаточно длинный цикл. Генераторы псевдослучайных чисел, основанные на идее Лемера (линейный конгруэнтный метод), обычно приведенные к интервалу (0, 1), созданы практически для всех систем программирования и служат основой для построения случайных величин с любой плотностью распределения $p(x)$. В библиотеках стандартных программ они носят имена *rand*, *random* (от английского *randomize*) и т. п.

Разумеется, кроме проверки на цикличность все датчики подвергаются проверке на равномерность и другим тестам. Такие программные датчики случайных чисел проверяются на равномерность пар и троек очередных последовательных значений, обеспечивая равномерность распределений в привычных для нас двумерном и трехмерном пространствах.

2.1.2. Моделирование случайных величин с известным законом распределения

Для преобразования случайных чисел x_i, равномерно распределенных в (0, 1), в аналогичные числа z_i из интервала (a, b) достаточно выполнить преобразование $z_i = a + x_i\,(b - a)$.

Если найдены параметры μ и σ, соответствующее равномерное распределение можно построить на основе оценок границ:

$$a = \mu - \sigma\sqrt{3}, b = \mu + \sigma\sqrt{3},$$

получаемых решением системы уравнений (2.1), (2.1б).

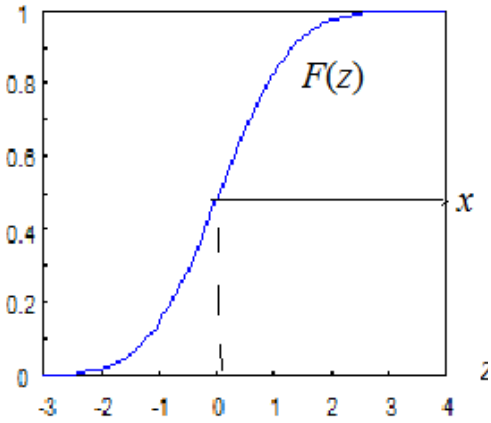


Рис. 2.2. Моделирование случайных величин с заданным распределением

Задача преобразования равномерного распределения к другим существенно усложняется.

Если случайная величина x равномерно распределена на (0, 1), то искомая непрерывная случайная величина z получается с помощью преобразования (рис. 2.2) $z = F^{-1}(x)$, где $F^{-1}(x)$ – функция, *обратная* к функции распределения вероятностей генерируемой случайной величины (метод обратной функции).

Другими словами, задача сводится к поиску (см. рис. 2.2) величины z, являющейся решением уравнения $F(z) = x$.

С учетом $F(z)=\int_{-\infty}^{z} p(z)dz,\ \frac{dF(z)}{dz}=p(z)$ иногда это уравнение решается без труда. Так, для показательного распределения $p(z)=\lambda e^{-\lambda z}, z>0,\ F(z)=1-e^{-\lambda z}$. Найдя обратную функцию, получаем $z = -\ln(1-x)/\lambda$.

Однако, имеются интересные для практики статистические распределения, для которых придется решать относительно z уравнение

$$x=\int_{-\infty}^{z} p(y)dy,$$

что требует от исследователя знакомства с простейшими методами численного интегрирования и численного решения уравнений.

2.1.3. Нормальное распределение

Как было отмечено выше, нормальное распределение представляет наибольший интерес для практики и построения методов статистического анализа. *Нормальное распределение* (Гальтон, 1889 г.), иногда называемое *гауссовским* или *Z-распределением*, обозначается $N(\mu, \sigma^2)$ и определяется как

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},\ F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}dx. \qquad (2.2)$$

Выполнив нормировку данных $z=\frac{x-\mu}{\sigma}$, получаем нормальное распределение $N(0, 1)$ с нулевым математическим ожиданием и единичной дисперсией:

$$p(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2};\ F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz. \qquad (2.3)$$

Асимметрия этого распределения равна нулю, эксцесс равен трем.

В отличие от других распределений, нормальное выступает как своеобразный идеал: идеальная симметрия, мода и медиана совпадают с математическим ожиданием. Нормальное распределение многократно табулировано (табл. 2.1), имеет отличные аппроксимации; практически во всех системах программирования имеются соответствующие функции, позволяющие для заданного x найти $F(x)$ и для заданного значения $F(x)$ найти соответствующее x.

Таблица 2.1. Значения функции нормального распределения

x	$F(x)$	x	$F(x)$	x	$F(x)$	x	$F(x)$
0,0	0,0000	1,0	0,3413	2,0	0,4772	3,0	0,4987
0,1	0,0398	1,1	0,3643	2,1	0,4821	3,01	0,4987
0,2	0,0793	1,2	0,3849	2,2	0,4861	3,02	0,4987
0,3	0,1179	1,3	0,4032	2,3	0,4893	3,03	0,4988
0,4	0,1554	1,4	0,4192	2,4	0,4918	3,04	0,4988
0,5	0,1915	1,5	0,4332	2,5	0,4938	3,05	0,4989
0,6	0,2257	1,6	0,4452	**2,6**	**0,4953**	3,06	0,4989
0,7	0,2580	1,7	0,4554	2,7	0,4965	3,07	0,4989
0,8	0,2881	1,8	0,4641	2,8	0,4974	3,08	0,4990
0,9	0,3159	1,9	0,4713	2,9	0,4981	3,09	0,4990

Например, вероятность того, что случайная величина с распределением $N(0, 1)$ по модулю не превысит 2,6, составляет $\alpha = 0{,}99$ (см. табл. 2.1), то есть 99%-я двусторонняя квантиль $K_{0,99} = 2{,}6$. Если вспомнить правило трех сигм, то вероятность выхода за этот предел составляет $1 - 2 \cdot 0{,}4987 \approx 0{,}0026$, то есть примерно 0,26 %.

2.1.4. Распределение Лапласа – Шарлье

Название этого экзотического по происхождению распределения (рис. 2.3) связано с именем Карла Шарлье (1862–1934), шведского астронома, занимавшегося вопросами небесной механики и звездной астрономии и применившего методы математической статистики к изучению пространственного распределения звезд в Галактике и истинных движений звезд в окрестностях Солнца. Его исходные представления через сферические функции доступны пониманию лишь избранных, и здесь мы ограничимся его аппроксимацией

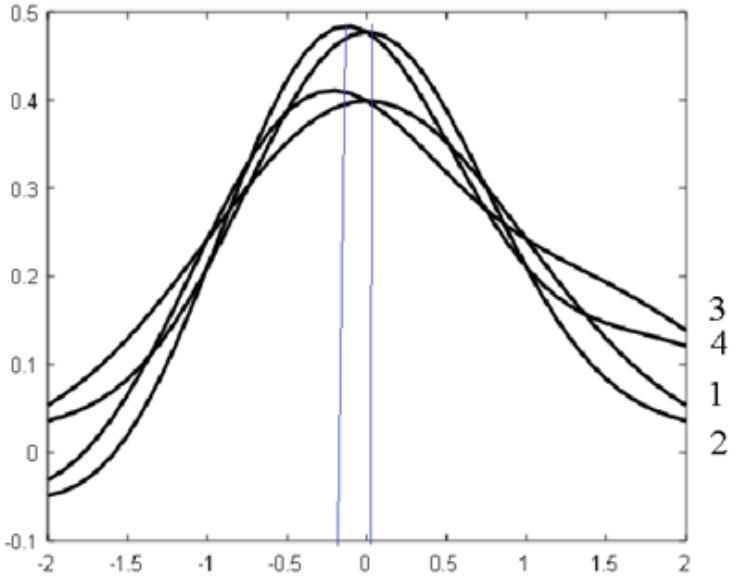


Рис. 2.3. Распределение Лапласа – Шарлье при различных асимметрии и эксцессе
(1 – $A_x = 0$, $E_x = 0$; 2 – $A_x = 0$, $E_x = 1{,}5$; 3 – $A_x = 1{,}5$, $E_x = 0$; $A_x = 1{,}5$, $E_x = 1{,}5$)

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-R^2}{2}}\left[1-\frac{A_x}{6}(3R-R^3)+\frac{E_x}{24}(R^4-6R^2+3)\right],\ R=\frac{x-\mu}{\sigma},$$

где $|A_x|<3$ и E_x – коэффициенты асимметрии и эксцесса ($A_x>0$ задает смещение влево, $E_x>0$ увеличивает остроту пика распределения). Эта аппроксимация достаточно точна при $|R|<\sqrt{3}$. Как представлено на рис. 2.4 [23], полученном при тех же параметрах, нарушение этого требования дает абсурдные решения (отрицательную вероятность).

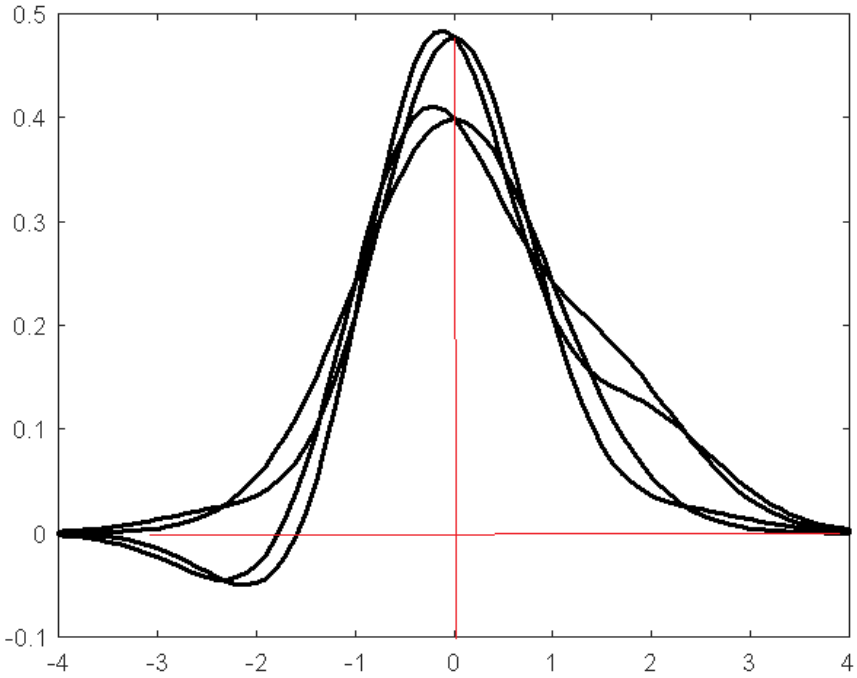


Рис. 2.4. Распределение Лапласа – Шарлье при нарушении правила трех сигм

2.1.5. Логарифмически нормальное распределение

Случайная величина x имеет логнормальное (логарифмически нормальное) распределение с параметрами μ и σ, если $y=\ln(x)$ имеет нормальное распределение с параметрами $\mu=\mu\ln(x)$, $\sigma=\sigma\ln(x)$. Естественно, случайная величина с логнормальным распределением принимает только положительные значения.

Плотность распределения $p(x)$ и функция распределения $F(x)$ приведены на рис. 2.5 и имеют следующий вид:

$$p(x)=\frac{1}{x\,\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2},\ F(x)=\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right),\ x>0.$$

Функция этого распределения (как и для нормального) не представима в элементарных функциях (неберущийся интеграл). Распределение

отличается от нормального лишь представлением случайных величин в логарифмической шкале.

Математическое ожидание $Mx = e^{\mu+\frac{\sigma^2}{2}}$.

Стандартное отклонение $Sx = \sqrt{Dx} = e^{\mu+\frac{\sigma^2}{2}}\sqrt{e^{\sigma^2}-1}$.

Медиана $Me = \frac{\mu^2}{\mu^2+\sigma^2}$. Мода $Mo = \frac{\mu^2}{\mu^2+\sigma^2} \Big/ \frac{Mx^2+Sx^2}{Mx^2}$.

Асимметрия $Ax = e^{\sigma^2+2}\sqrt{e^{\sigma^2}-1}$.

Эксцесс $Ex = \exp(4\,\sigma^2) + 2\exp(3\,\sigma^2) + 3\exp(4\,\sigma^2) - 3$.

К этому распределению прибегают при очень больших и очень малых значениях x. Так, промежуток $[1, 10^6]$ логарифмированием сводится к диапазону от 0 до 17,3.

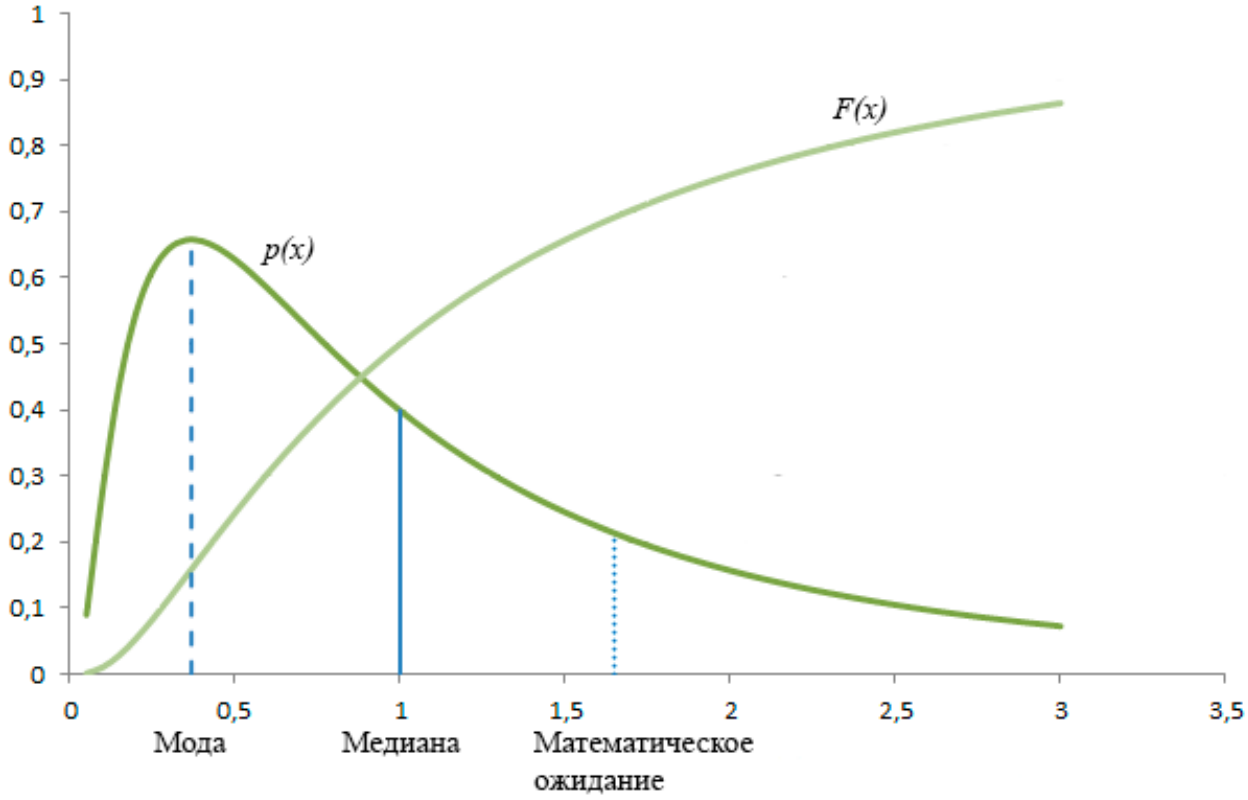


Рис. 2.5. Плотность и функция логнормального распределения ($\mu = 0$, $\sigma = 0{,}7$)

2.1.6. Распределение Лапласа

Это распределение (рис. 2.6), иногда называемое *двусторонним показательным*, имеет функцию плотности $p(x)$ и функцию распределения вероятностей $F(x)$ следующего вида:

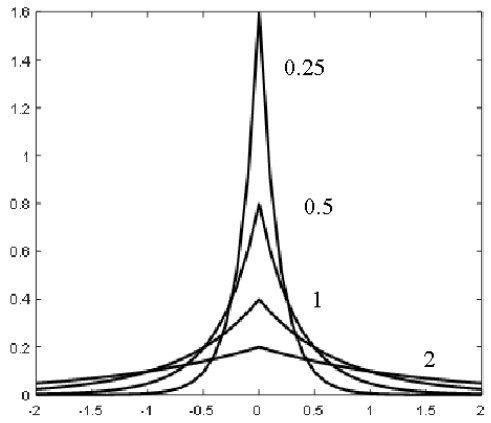


Рис. 2.6. Распределение Лапласа при различных значениях μ

$$p(x) = \frac{1}{\sigma\sqrt{2}} e^{-\sqrt{2}\left|\frac{x-\mu}{\sigma}\right|}, -\infty < x < \infty,$$

$$F(x) = \begin{cases} \frac{1}{2} e^{\sqrt{2}\frac{x-\mu}{\sigma}}, x < \mu \\ 1 - \frac{1}{2} e^{\sqrt{2}\frac{\mu-x}{\sigma}}, x > \mu \end{cases}.$$

Мода и медиана распределения Лапласа совпадают с μ, асимметрия отсутствует, а эксцесс $Ex = 3$.

2.1.7. Треугольное распределение

Треугольное распределение, называемое еще распределением Симпсона (рис. 2.7), имеет следующую функцию плотности

$$p(x) = \begin{bmatrix} \frac{2(x-a)}{(b-a)(c-a)}, & a \le x \le c \\ \frac{2(b-x)}{(b-a)(b-c)}, & c \le x \le b \\ 0, & x \not\subset [a,b] \end{bmatrix}$$

и функцию распределения

$$F(x) = \begin{bmatrix} \frac{(x-a)^2}{(b-a)(c-a)}, & a \le x \le c \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)}, & \\ 0, x \le a] & \\ 1, x \ge b & \end{bmatrix}$$

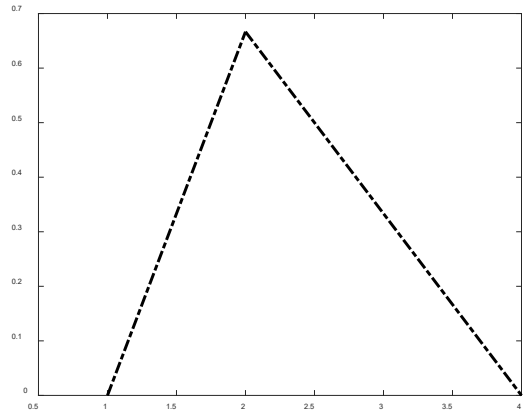


Рис. 2.7. Плотность распределения Симпсона ($a = 1$, $b = 4$, $c = 2$)

Математическое ожидание $\mu = (a + b + c) / 3$. Дисперсия $\sigma^2 = (a^2 + b^2 + c^2 – a\,b – a\,c – b\,c)$.

Треугольное распределение обычно используется как экспериментальное при визуальной оценке моды получаемого эмпирического распределения (для оценки параметра c).

2.1.8. Экспоненциальное распределение

В отличие от описанных выше двусторонних распределений ошибок (отклонений от гипотетического стандарта), используемых в приложениях типа контроля качества продукции, экспоненциальное распределение, иногда называемое *показательным*, чаще используется для выявления тенденций и параметров функционирования реального процесса.

Это распределение стало популярным при решении задач теории массового обслуживания, в частности для имитации времени между очередными поступлениями заявок в систему. Плотность (рис. 2.8) и функция показательного распределения представляются следующим образом: $p(x) = \lambda\, e^{\lambda x}$, $F(x) = 1 – e^{-\lambda x}$, $x > 0$.

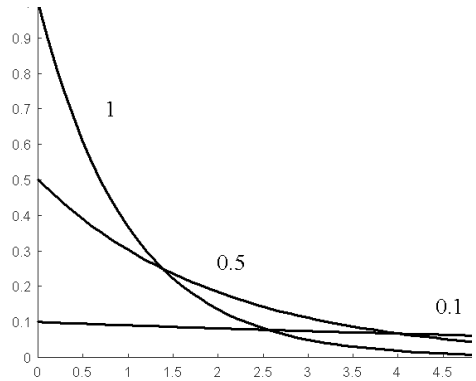


Рис. 2.8. Плотность экспоненциального распределения при разных значениях λ

Математическое ожидание и стандартное отклонение равны $Mx = Sx = 1 / \lambda$.

Медиана $Me = \ln(2) / \lambda$, мода $Mo = 0$, асимметрия $Ax = 2$, эксцесс $Ex = 6$.

Параметр λ в задачах теории массового обслуживания можно понимать как среднее число событий (заявок, отказов, звонков) в единицу времени.

Случайные величины, распределенные экспоненциально, легко моделировать на базе датчиков случайных равномерно распределенных в интервале (0, 1) чисел (см. п. 2.1.1).

2.1.9. Распределение Рэлея

Распределение названо именем лауреата Нобелевской премии по физике Д. У. Стретта, барона Рэлея (1842–1919), который ввел его в 1880 году в связи с задачей сложения гармонических колебаний со спиральными фазами, и оно до сих пор используется в теории связи.

Плотность и функция распределения:

$$p(x)=\frac{x}{\sigma^2}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2},\ F(x)=1-e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2},\ x\in[0,\infty).$$

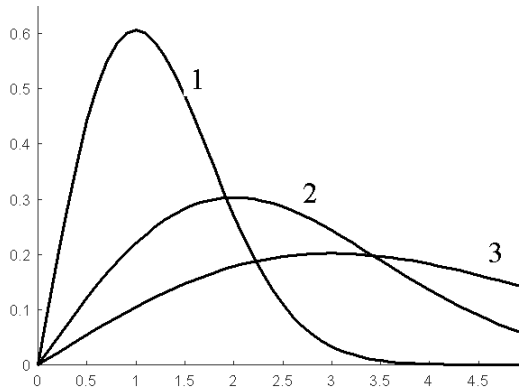


Рис. 2.9. Плотность распределения Рэлея при разных значениях σ

Математическое ожидание и дисперсия равны соответственно

$$Me=\sqrt{\frac{\pi}{2}}\,\sigma \text{ и } Dx=-\frac{4-\pi}{2}\sigma^4.$$

Максимум плотности этого распределения равен $\frac{1}{\sigma}\sqrt{e}$ и достигается при $x = \sigma$.

Случай $\sigma = 1$ выделяется как *распределение квадратного корня* случайной величины, имеющей χ^2-распределение (хи-квадрат распределение) с двумя степенями свободы (см. п. 2.3.1).

2.1.10. Распределение Максвелла

Это распределение, являющееся аналогом распределения Рэлея в трехмерном пространстве, было предложено в 1859 году знаменитым физиком Джеймсом К. Ма́ксвеллом (1831–1879) при создании кинетической теории газов (распределение молекул газа или большого числа частиц по скоростям движения в трехмерном пространстве). Его плотность и функция распределения имеют следующий вид:

$$p(x)=\sqrt{\frac{2}{\pi}}\,\frac{x^2}{\beta^3}e^{-\frac{1}{2}(\frac{x}{\beta})^2},\ F(x)=-\sqrt{\frac{2}{\pi}}\,\frac{x}{\beta}e^{-\frac{1}{2}\frac{x^2}{\beta^2}}-1+2\Phi(x/\beta),$$

где $\beta=\sigma\sqrt{\frac{\pi}{3\pi-8}}$, $\Phi(x)$ – функция Лапласа, $x > 0$.

Математическое ожидание $Mx=\beta\sqrt{\frac{8}{\pi}}$, дисперсия $Dx=\frac{3\pi-8}{\pi}\beta^2$, мода $Mo=\beta\sqrt{2}$.

2.1.11. Гамма-распределение

Это распределение популярно в практике оценки надежности оборудования (срок службы изделия, время наработки на отказ и т. п.), в медицине, логистике.

Функция плотности распределения (рис. 2.10) имеет вид

$$p(x) = x^{k-1}\frac{\exp(-x/s)}{\Gamma(k)s^k},\ x \ge 0.$$

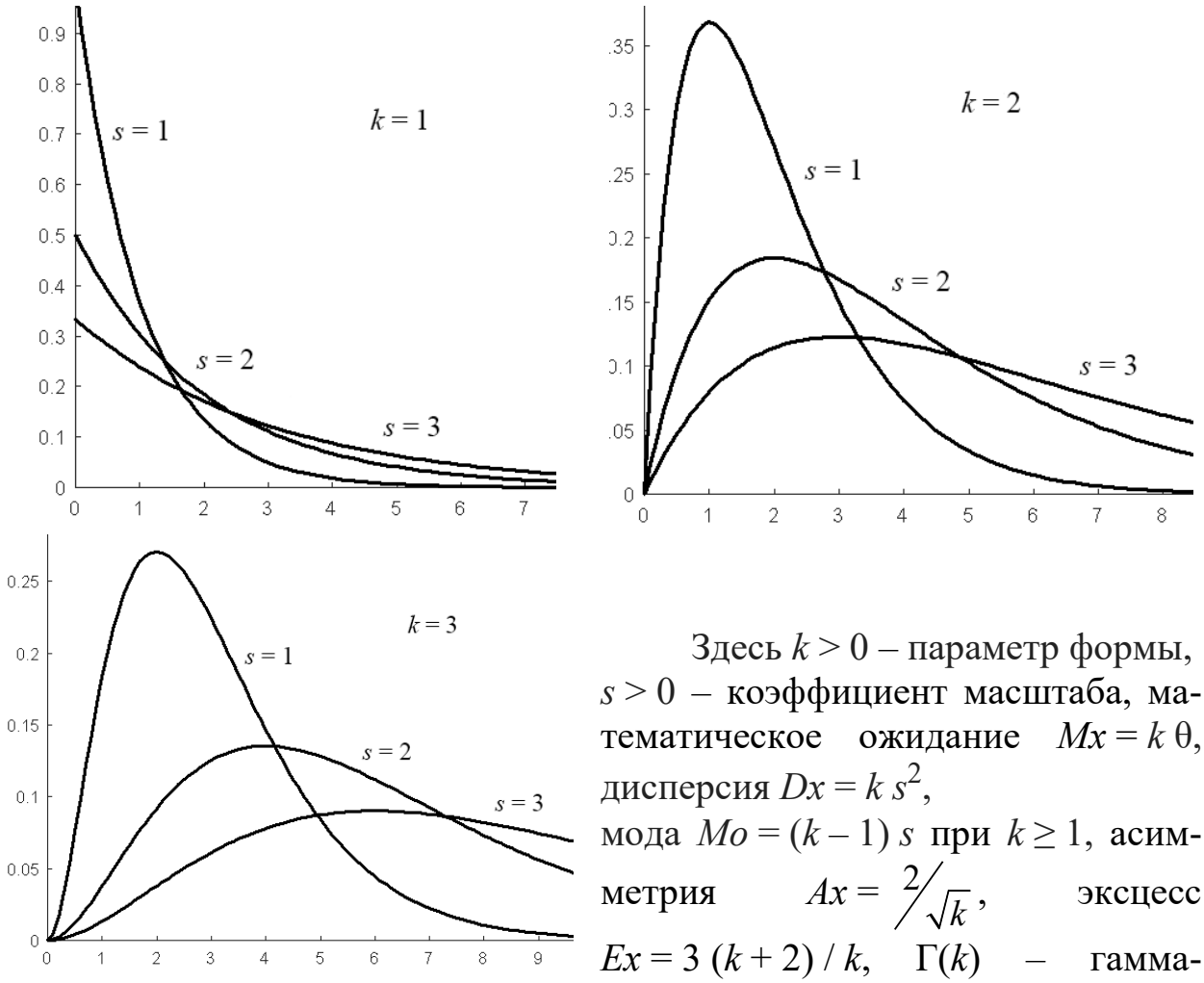


Рис. 2.10. Плотность гамма-распределения

Здесь $k > 0$ – параметр формы, $s > 0$ – коэффициент масштаба, математическое ожидание $Mx = k\,\theta$, дисперсия $Dx = k\,s^2$, мода $Mo = (k-1)\,s$ при $k \ge 1$, асимметрия $Ax = 2/\sqrt{k}$, эксцесс $Ex = 3\,(k+2)/k$, $\Gamma(k)$ – гамма-функция.

С гамма-функцией $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$ приходится иметь дело при рассмотрении многих других распределений. Отметим некоторые ее свойства.

При положительном целочисленном аргументе ($k > 0$) значения гамма-функции совпадают с факториалом, причем

$$\Gamma(1) = \Gamma(2) = 1,\ \Gamma(k+1) = k!$$

При любом k

$\Gamma(0{,}5) = \sqrt{\pi}$, $\Gamma(k+1) = k\,\Gamma(k)$, $\Gamma(x \to 0) \to \infty$, $\Gamma(x \to \infty) \to \infty$.

Для вычисления значений гамма-функции при отрицательных значениях аргумента можно воспользоваться так называемой формулой дополнения Эйлера

$$\Gamma(x)\,\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}.$$

В частном случае при $k = 1$ гамма-распределение совпадает с *экспоненциальным распределением*.

При целочисленных k гамма-распределение называют *распределением Эрланга* (по имени датского ученого К. А. Эрланга (1878–1929), изучавшего в 1908–1922 годах актуальные для того времени задачи функционирования телефонных сетей.

2.1.12. Распределение Вейбулла

Как утверждается в литературе, это распределение названо в честь шведского инженера В. Вейбулла (1887–1979), использовавшего его в 1951 году в практике машиностроения при описании экспериментально полученных разбросов усталостной прочности стали и пределов ее упругости.

В 1927 году Фреше было предложено обратное распределение Вейбулла, получившее название распределения Фреше. Это распределение часто используется в гидрологических приложениях.

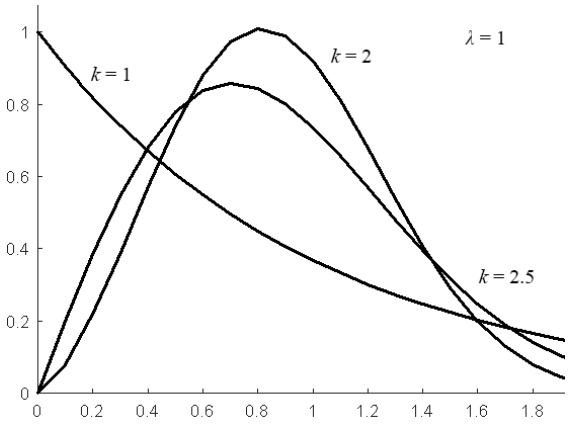


Рис. 2.11. Плотность распределения Вейбулла

Распределение Вейбулла широко применяется при расчете показателей надежности, в частности при исследовании предела упругости ряда металлов, прочности и усталостной долговечности деталей (подшипники качения, напряженные оси и валы и др.). Примечательно, что оно является универсальным, превращаясь при некоторых сочетаниях параметров в нормальное, экспоненциальное и другие.

Функции плотности и распределения выглядят так:

$$p(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}},\quad F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^{k}},\quad x > 0.$$

Характеристики распределения (табл. 2.2):

математическое ожидание $Mx = \lambda\,\Gamma\left(1+\frac{1}{k}\right)$,

дисперсия $Dx = \lambda^2\,S(k)$, где $S(k) = \Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)$,

медиана $Me = \lambda\left[\ln(2)\right]^{1/k}$, мода $Mo = \lambda\left(\frac{k-1}{k}\right)^{1/k}$,

асимметрия $Ax = \frac{1}{S(k)^{3/2}}\left[\Gamma\left(1+\frac{3}{k}\right) - 3\,\Gamma\left(1+\frac{2}{k}\right)\Gamma\left(1+\frac{1}{k}\right) + 2\,\Gamma^3\left(1+\frac{1}{k}\right)\right]$,

эксцесс $Ex = \frac{1}{S(k)^2}\left[\begin{array}{l}\Gamma\left(1+\frac{4}{k}\right) - 4\,\Gamma\left(1+\frac{3}{k}\right)\Gamma\left(1+\frac{1}{k}\right) + \\ + 6\,\Gamma\left(1+\frac{2}{k}\right)\Gamma^2\left(1+\frac{1}{k}\right) - 3\,\Gamma^4\left(1+\frac{1}{k}\right)\end{array}\right]$.

Табл. 2.2. Значения асимметрии и эксцесса при различных k

k	0,5	1	1,5	2,0	2,5
$S(k)$	20,0000	1,0000	0,3757	0,2146	0,1441
Ax	6,6188	2,0000	1,0720	0,6311	0,3586
Ex	89,1600	9,0000	0,3788	–8,3965	–21,0125

Если за случайную величину x принять за наработку до отказа, то получается распределение Вейбулла, где интенсивность отказов при различных значениях k пропорциональна времени:

$k < 1$ – уменьшается со временем;

$k = 1$ – не меняется со временем;

$k > 1$ – увеличивается со временем.

При $k = 1$ распределение Вейбулла трансформируется в экспоненциальное (показательное).

2.1.13. Логистическое распределение

Логистическое распределение ныне известно в связи с многочисленными (не всегда обоснованными, как утверждает ряд авторов) попытками его применения для описания разнообразных законов развития в социологии, экономике, биологии. Внешне оно мало отличается от нормального распределения, а в некоторых случаях оказывается удобнее его. Это распределение часто используется в статистическом анализе при исследовании медико-биологических объектов.

Распределение определяется двумя параметрами (рис. 2.12): μ – математическое ожидание и σ – стандартное отклонение. Плотность и функция распределения приведены ниже:

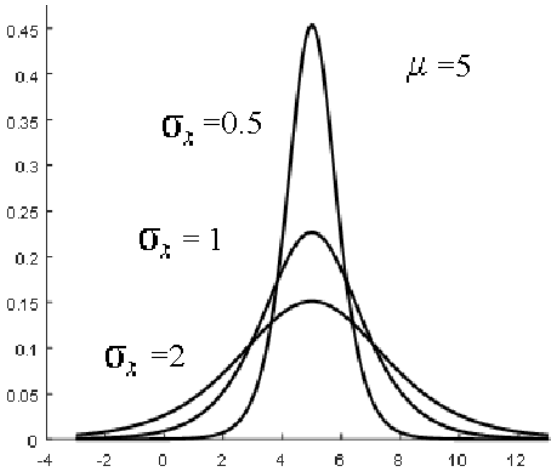


Рис. 2.12. Логистическое распределение

$$p(x)=\frac{1}{k}\frac{e^{-z}}{(1+e^{-z})^2},\ F(x)=\frac{1}{1+e^{-z}},$$

где $k=\frac{\sigma\sqrt{3}}{\pi},\ z=\frac{x-\mu}{k}$.

Для генерации случайных чисел x из этого распределения методом обратной функции используется моделирующая формула $x=\mu+k\ln\left(\frac{r}{1-r}\right)$, где r – случайная величина, равномерно распределенная в интервале (0, 1).

2.1.14. Степенное распределение

Нормальное распределение и родственные с ним строятся на предположении, что события с большим эффектом происходят достаточно редко (их вероятность мала). Существуют явления, где события с очень большим эффектом могут происходить достаточно часто (одно и то же явление можно трактовать двояко: например, с позиций длительности телефонных разговоров или простоя средств связи, спрос на жизненно необходимые товары при кризисах и катастрофах и пр.).

Самым простым среди подобных законов распределения является так называемое степенное распределение. Здесь плотность и функция распределения (рис. 2.13) выглядят следующим образом:

$$p(x)=c\,x^{c-1},\ F(x)=x^c,\ 0<c<1.$$

Математическое ожидание и дисперсия:

$$Mx=\frac{c}{c+1},\ Dx=\frac{c}{(c+1)^2(c+2)}.$$

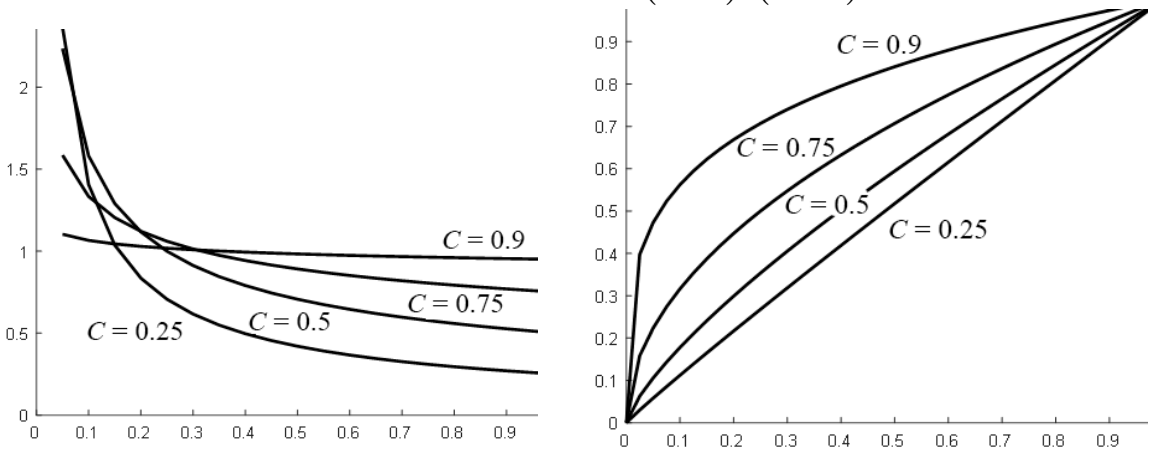


Рис. 2.13. Плотность и функция степенного распределения

Значения $p(x)$ не всегда соответствуют требованию к вероятностям $p(x) \le 1$, носят относительный характер, но соблюдается требование $F(x) \le 1$. Генерация случайных чисел x этого распределения производится посредством простой обратной функции $x = r^{1/c}$, где r – равномерно распределенная в (0, 1) случайная величина.

2.1.15. Распределение Парето

Это распределение (рис. 2.14) связано с понятием оптимальности по Парето при *многокритериальной оптимизации*. Оптимальность по Парето означает, что нельзя улучшить значение одного критерия, не ухудшая значение хотя бы одного из оставшихся. Область значений параметра, где имеет место такое явление, называют областью компромисса, переговорным множеством или множеством Парето.

Данное распределение из упомянутого выше семейства степенных стало популярным с появлением работ В. Парето (1848–1923) о *распределении доходов* и сегодня популярно при исследовании эффективности принимаемых решений в экономике, лингвистике, социологии, эконометрике и др.

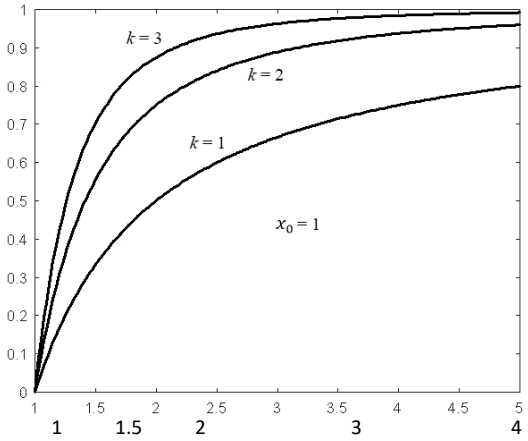


Рис. 2.14. Распределение Парето

Плотность и функция распределения (рис. 2.14):

$$p(x) = \frac{kx_0^k}{x^{k+1}},\ F(x) = 1 - (\frac{x_0}{x})^k,\ x \ge x_0.$$

Математическое ожидание $Mx = \frac{x_0}{k-1}$,

дисперсия $Dx = \frac{x_0^2 k}{(k-1)^2(k-2)}$, медиана

$Me = x_0 \sqrt[k]{2}$, мода $Mo = x_0$.

Моделирование распределения Парето представимо соотношением $x = (1/r)^{1/k}$, где r – случайная величина, равномерно распределенная на (0, 1).

Обратившись к Интернету, получаем весьма большое число ссылок на распределение Парето. Чем вызван такой интерес к этому частному случаю степенного распределения?

Справедливое распределение доходов всегда было темой, на которой оттачивали копья все: от Кампанеллы с его утопическим «Городом Солнца» и Рабочего Интернационала до конкурентов в политических баталиях и представителей желтой прессы.

Никогда в обозримом прошлом этой справедливости не было. Известному «птенцу гнезда Петрова» А. Д. Меншикову приписывают суж-

дение «Чем более владею, тем более рука горит».* Фараонам, жрецам, князьям, энергичным финансистам, не всегда создавшим стартовый капитал по библейским заповедям, «звездам» эстрады (справедливо или нет – другой вопрос)) принадлежала бóльшая доля богатств. Изредка высоко оценивался вклад в науку и технологии.

В. Парето как экономист, имеющий физико-математическое образование, обнаружил, что распределение богатств во Франции далеко от нормального и имеет вид степенного распределения (80 % богатства принадлежат 20 % населения). Более того, он обнаружил, что распределение богатства в обществе обычно подчиняется определенному закону: *с удвоением размера контролируемой собственности/богатства, количество людей, достигших соответствующего уровня сокращается в геометрической прогрессии, причем с примерно постоянным множителем*. Парето пришел к выводу, что «неравенство распределения богатства в обществе – нечто вроде естественного закона природы, эффект которого можно сгладить, но невозможно устранить в денежной системе».

Оказалось даже, что небольшое количество наиболее жизнеспособных стручков производит большую часть гороха. Подобный факт оказался характерным даже для бытовых ситуаций.

«Большинство людей на планете 80 % времени носят только 20 % одежды, имеющейся у них в гардеробе. 80 % книг, прочитанных человеком за свою жизнь, не дали ему полезной информации. Лишь 20 % прочитанной литературы дали человеку 80 % всех новых знаний и умений, которые он успешно применяет в повседневной жизни. В мире преступности 20 % осужденных и заключенных совершили 80 % всех преступлений» (https://ru.wikipedia.org/wiki/Закон_Парето).

В основу соответствующих выводов положен *закон оптимальности по Парето* – состояние некоторой системы, при котором значение каждого частного показателя, характеризующего систему, не может быть улучшено без ухудшения других.

Значение функции распределения $F(x^*)$ определяет вероятность того, что $x \le x^*$. Соответственно, x^* – число людей, обеспечивающих соответствующий процент всего благосостояния, x^* – число фермеров, производящих этот процент урожая всеми фермерскими хозяйствами и т. д.

Резкое отклонение от линии Парето (см. рис. 2.14) свидетельствует о кризисах, землетрясениях, революциях и других нерядовых явлениях, повлиять на которые не в нашей власти.

---

* Тынянов Ю. Н. Сочинения. Т. 1. Восковая персона. 1959. С. 361.

Критерий оптимальности по Парето используют, чтобы оценить факт влияния предложенных экономических реформ на общий уровень благосостояния, хотя это и не говорит о том, что надо делать. От того, что Петр I повесил за мздоимство сибирского губернатора, число руководителей, «не отличавшихся высокой нравственностью» в стране не уменьшилось. Революционные потрясения оказывают определенное влияние, но с течением времени все возвращается на круги своя.

Закон Парето (20/80) подвержен многочисленной критике, заявляющей, что его не следует рассматривать как непреложный закон природы с конкретно заданными числовыми параметрами, поскольку в реально существующих многофакторных системах их свойства описываются совокупностью, а не одним параметром и т. п.

2.2. Дискретные распределения вероятностей

Если непрерывные распределения являются некой идеализацией реальной жизни, то дискретные более понятны для обыденности, имеют превосходную физическую интерпретацию, заложенную в азбуку теории вероятностей.

Они возникли, в первую очередь, на потребу одержимым любителям игр типа рулетки, не требующих высокого интеллекта. Сколько раз подряд ставить на красное? А не фальшива ли эта кость, или карты крапленые? Играли (и играют) все желающие выбраться из нищеты, проигрывающие последнее, нажитое «непосильным трудом» и ищущие развлечений или большей наживы за зеленым сукном.

2.2.1. Биномиальное распределение и распределение Бернулли

Пусть проводится серия из n независимых испытаний, заканчивающихся либо *успехом*, либо *неуспехом* и в каждом испытании (опыте) вероятность успеха p, а вероятность неуспеха $q = 1 - p$. С такими опытами можно связать случайную величину k, значение которой равно числу успехов в серии из n испытаний. Ее распределение называется биномиальным (рис. 2.15) и определяется формулой Бернулли

$$P(x = k \mid n, p) = C_n^k p^k (1-p)^{n-k}, \; 0 < p < 1, \; k = 0, 1, \ldots, n.$$

В случае непрерывных распределений поиск первых моментов распределений обычно не составлял труда. Для дискретных распределений от аппарата интегрального исчисления приходится отказаться и прибегнуть к аппарату дискретной математики (суммы и ряды). Даже поиск математического ожидания требует вычисления суммы

$$\sum_{k=0}^{n} k \cdot C_n^k p^k (1-p)^{n-k}.$$

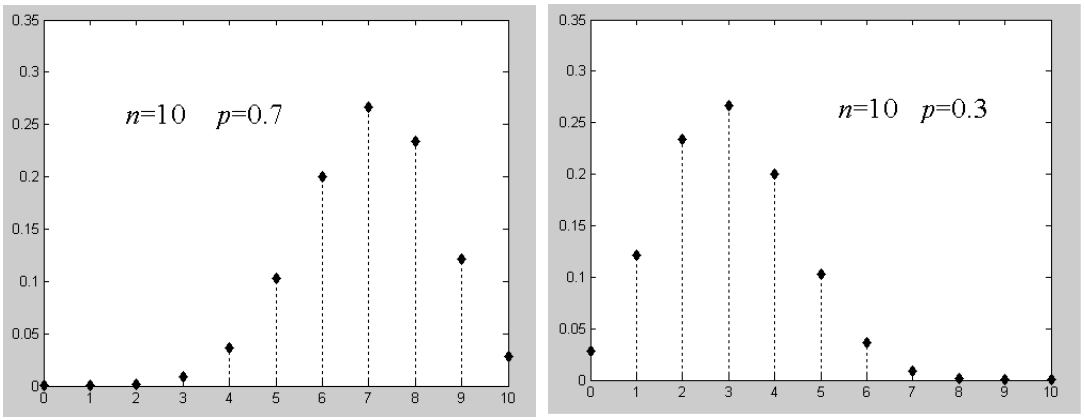


Рис. 2.15. Биномиальное распределение

Обратившись к литературе, мы обнаруживаем, что для рассматриваемого распределения:

$$Mx = n\,p,\ Dx = n\,p(1-p),\ Mo = n\,p,$$

$$Ax = {1-2p}\Big/{\sqrt{np(1-p)}},\ Ex = \frac{1-6p(1-p)}{np(1-p)} + 3\,.$$

В случае $n = 1$ (k равно 0 или 1) распределение называют *бернуллиевым* по имени упомянутого выше Я. Бернулли, изучавшего схему простейшего статистического эксперимента с двумя исходами:

$$P(x = k \mid p) = p^k (1-p)^{1-k},\ k = 0,\ 1.$$

Не следует забывать, что $21! \approx 5{,}1091 \cdot 10^{19}$, $170! \approx 7{,}2574 \cdot 10^{306}$ на грани переполнения (`overflow`) и колоссальной абсолютной погрешности. При большом n и малом p ($p < 0{,}1$) справедливо соотношение:

$$C_n^k p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda},$$

где $\lambda = n\,p$, то есть биномиальное распределение аппроксимируется распределением Пуассона (см. п. 2.2.3), что значительно сокращает время выполнения вычислительной процедуры.

Кстати, в эпоху арифмометра старались минимизировать число операций умножения и при больших n прибегали к асимптотике Дж. Стирлинга $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ $n! = \sqrt{2\pi n} \cdot (n/e)^n$ с точностью до множителя порядка малости $e^{O(n)}$, где $O(n) < 1 / (12\,n)$. Но попробуйте найти n^n при $n = 100$!

2.2.2. Полиномиальное распределение

В отличие от биномиального, полиномиальное (мультиномиальное) распределение связано с появлением одного из $m > 2$ взаимоисклю-

чающих событий при повторных независимых испытаниях (например, в игре с костью имеем дело с 6 исходами – событиями).

Классическая схема интерпретации: имеется урна с шарами m различных цветов $A_1, A_2, \ldots, A_m$, где шары каждого цвета составляют известную долю среди всех шаров. По схеме выбора с возвращением из урны извлекается N шаров. Пусть при каждом испытании вероятность появления события A_i равна p_i $(1 \le i \le m)$.

Вероятность совместного распределения количества n_i появления событий в серии из N испытаний определяется как

$$P(n_i / i = 1, m \mid N, p) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m} .$$

Каждое из событий (произошло или нет) имеет биномиальное распределение с математическим ожиданием $Mx_i = N p_i$ и дисперсией $Dx_i = N p_i (1 - p_i)$.

2.2.3. Распределение Пуассона

Это одно из популярнейших дискретных распределений, сообщение о котором впервые опубликовано в 1837 году знаменитым французским математиком С. Пуассоном (1781–1840) как предельный случай закона больших чисел (рис. 2.16):

$$P(x = k) = p_k = \frac{\lambda^k}{k!} e^{-\lambda}, (k = 0, 1, 2, \ldots), \lambda = n p.$$

Математическое ожидание и дисперсия таких случайных величин одинаковы:

$$Mx = Dx = \lambda.$$

В роли конкретного значения k может выступать число событий за фиксированный промежуток времени в предположении независимости этого числа от момента времени (число вызовов, поступивших на телефонную станцию за время t, число выявленных дефектных изделий, количество студентов, опоздавших на лекцию и т. д.). Череду таких событий называют пуассоновским потоком событий. Распределение Пуассона часто используется для описания появления редких событий. Обычно предполагается, что для пуассоновского потока событий характерна *ординарность* (невозможно одновременное появление двух и более событий), *стационарность* (независимость наступления определенного числа событий за определенное время от начала его отсчета), *отсутствие последействия* (вероятность поступления определенного числа событий за отрезок времени не зависит от числа ранее наступивших событий). В последнее столетие распределение Пуассона приобрело особую популярность как база теории массового обслуживания.

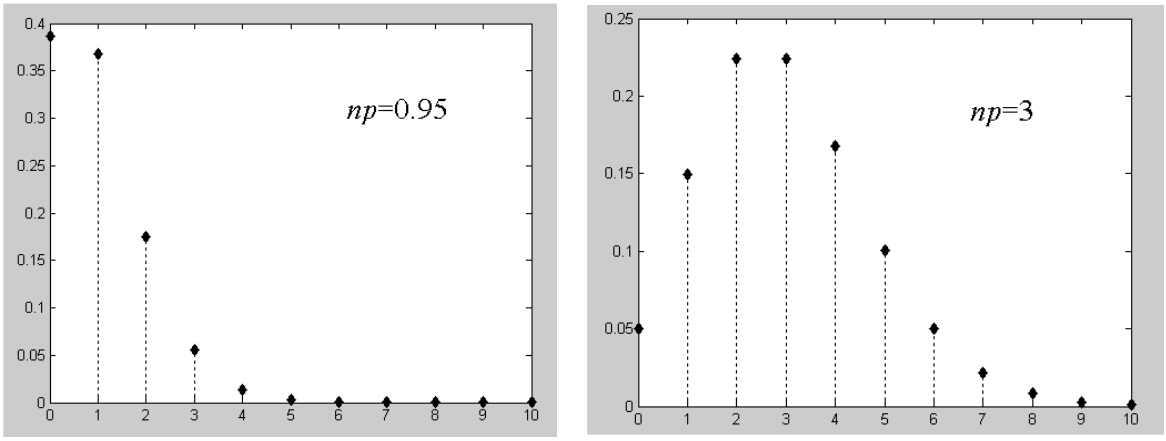


Рис. 2.16. Распределение Пуассона

Распределение Пуассона дает хорошую аппроксимацию биномиального распределения для больших значений n и малых p, если $\lambda = n\,p$ невелико (схема редких событий).

Многие авторы иллюстрируют этот факт примером из знаменитой книги [1]. Здесь выясняется вероятность p_k того, что в группе из 500 случайно выбранных людей k из них родились именно 1 января. При оценках в схеме Бернулли можно принять $n = 500$ и $p = 1 / 365$, для пуассоновского потока $\lambda = n\,p = 500 / 365 \approx 1{,}37$.

Плотности этих распределений совпадают с точностью 0,00005, но вычислительные затраты несопоставимы (поиск посредством оценок факториалов для биномиального распределения почти безнадежен, попробуйте вычислить 500! = ?!).

2.2.4. Геометрическое распределение

Со схемой испытаний Бернулли можно связать еще одну случайную величину k – число испытаний *до первого успеха* (последовательность $k - 1$ неудач).

Ее распределение называется геометрическим (рис. 2.17) и определяется формулой

$$p_k = P(x = k) = (1 - p)^{k-1}\,p,\ 0 < p < 1,\ k = 1, 2, \ldots,$$
$$Mx = 1 / p,\ Dx = (1 - p) / p^2.$$

Геометрическое распределение реализуется последовательным моделированием с одновременным подсчетом количества генераций случайной величины, равномерно распределенной в (0, 1), до тех пор, пока не обнаружится равномерно распределенная случайная величина, меньшая p.

Заметим, что нетрудно разрешить и обратную задачу моделирования величины $k = 1 + \left[\dfrac{\ln(r) - \ln(p)}{\ln(1 - p)}\right]$, где r – случайная величина, равно-

мерно распределенная в (0, 1), а квадратными скобками обозначена целая часть числа.

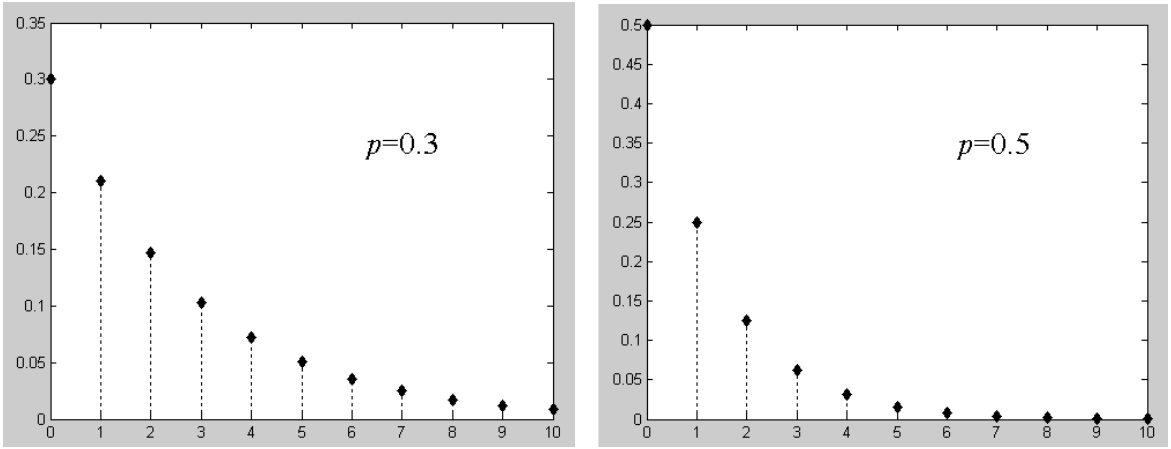


Рис. 2.17. Геометрическое распределение

Геометрическое распределение достаточно популярно, в частности, при разработке математических методов контроля качества промышленной продукции.

2.2.5. Отрицательное биномиальное распределение

В последовательности испытаний Бернулли с вероятностями «успеха» p и «неудачи» $q = 1 - p$ число k неудач до получения r-го успеха определяет так называемое отрицательное биномиальное распределение, плотность которого

$$P(x = k) = p(k \mid r, p) = p_k = C_{r+k-1}^{k} p^r (1-p)^k, k = 0, 1, 2, \ldots,$$

где $0 \le k < \infty$ и $r > 0$ – целые числа.

Заметим, что приведенная функция в различных источниках имеет различное представление с учетом тождества

$$C_{r+k-1}^{k} = C_{r+k-1}^{r-1}.$$

Математическое ожидание случайных чисел из отрицательного биномиального распределения с заданными параметрами p и r $Mx = r\,(1-p)\,/\,p$, дисперсия $Dx = r\,(1-p)\,/\,p^2$.

Если ограничиться только целыми значениями $r > 0$, то вышеприведенная интерпретация становится вполне естественной и отрицательное биномиальное распределение называют *распределением Паскаля.*

2.2.6. Распределение Паскаля

Плотность распределения Паскаля (рис. 2.18)

$$P(x = k) = p(k \mid r, p) = p_k = C_{r+k-1}^{k} p^r \left(1-p\right)^k,\ r \ge 0,\ k < \infty$$ – целые.

Математическое ожидание $Mx = r\,(1-p)\,/\,p$.

Дисперсия $Dx = r\,(1 - p) / p^2$.

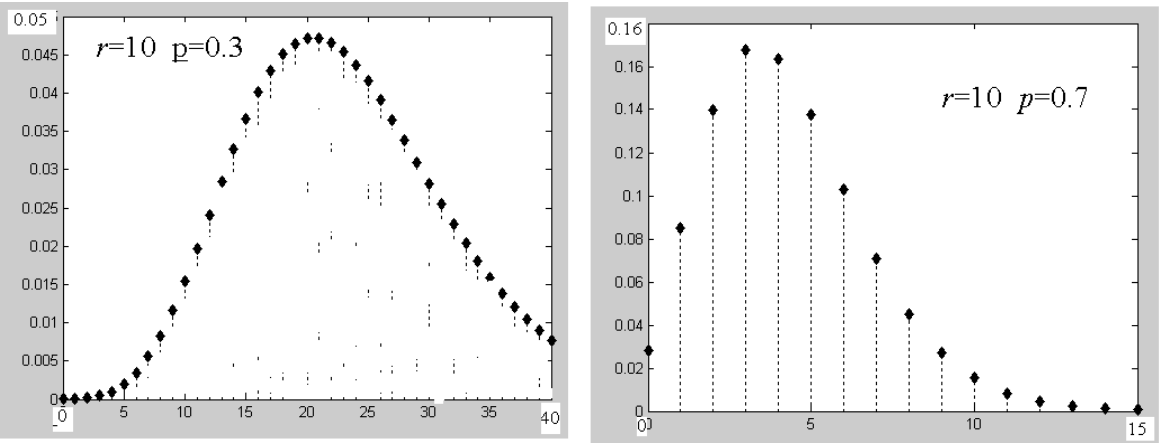


Рис. 2.18. Распределение Паскаля

Для генерации случайных чисел такого рода берем последовательность независимых случайных чисел, равномерно распределенных на (0, 1), и последовательно накапливаем количество чисел, меньших p, и количество чисел, больших p. В момент, когда количество чисел, меньших p, станет равным r, количество чисел, больших p, даст случайное число k, подчиняющееся распределению Паскаля.

В частном случае при $r = 1$ распределение Паскаля превращается в геометрическое – распределение числа испытаний (неуспехов), предшествующих первому успеху в последовательности Бернулли.

2.2.7. Гипергеометрическое распределение

Случайная величина x имеет *гипергеометрическое распределение* (рис. 2.19) с параметрами m, n и s, где m, n, $s > 0$ – целочисленные, $s \le m + n$, если

$$P(X = k) = p_k(x) = \frac{C_m^k C_n^{s-k}}{C_{m+n}^s},\ x = 0, 1, 2,\ ..., s\,.$$

Математическое ожидание и дисперсия гипергеометрического распределения

$$Mx = \frac{m}{m+n}s\,,\ Dx = \frac{mn}{(m+n)^2}\,\frac{s(m+n-s)}{m+n-1}\,.$$

Это распределение достаточно просто интерпретируется традиционной классической схемой «урны». В урне имеются m белых и n черных шаров. Из нее извлекаются (*случайно без возврата*) s шаров. Количество k извлеченных среди них белых шаров выступает как случайная величина с вышеприведенной плотностью распределения.

Естественно, что эта величина имеет ненулевую вероятность лишь при $\max(0, s - n) \le k \le \min(m, s)$.

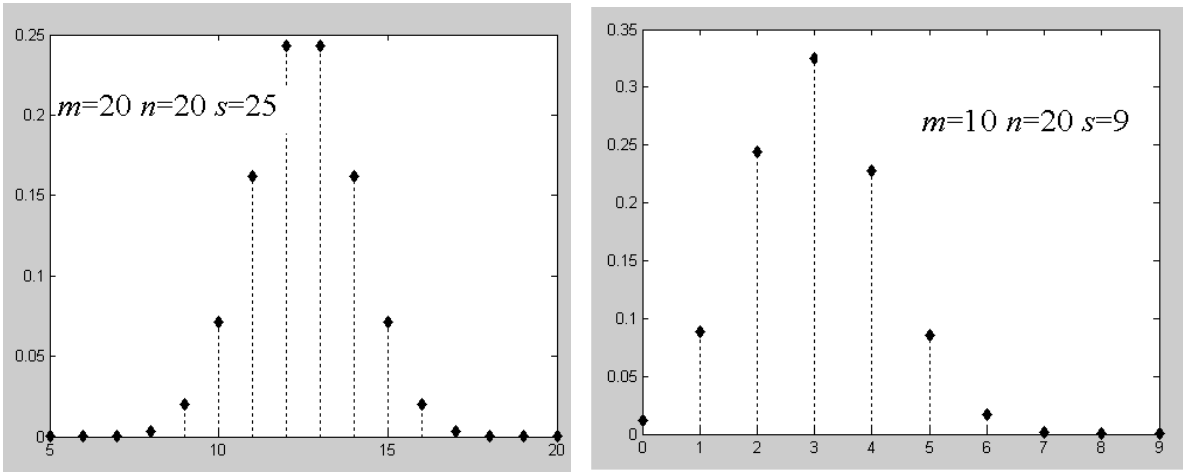


Рис. 2.19. Гипергеометрическое распределение

Гипергеометрическое распределение широко используется при статистическом контроле качества продукции и в аналогичных задачах выборочных обследований.

2.2.8. Распределение Маркова – Пойа

Впервые это распределение появляется в работах выдающегося русского математика академика А. А. Маркова (1856–1922), который опубликовал результаты анализа этого распределения в 1917 году в «Известиях Петербургской академии наук». Но из-за военно-политических бурь тех дней эта публикация не была замечена зарубежными читателями, и в 1923 году появляется работа Д. Пойа и Ф. Эггенбергера, где вводится такое же распределение.

В зарубежной литературе данное распределение принято называть именем одного из его авторов – венгерского математика Д. Пойа, известного в нашей стране, в частности, своей изумительной книгой «Математика и правдоподобные рассуждения». – М.: Наука, 1975. – 464 с., хотя более справедливо название *распределение Маркова – Пойа* (рис. 2.20).

Это распределение интерпретируется следующим образом. Из урны, содержащей первоначально a белых и b черных шаров, выбирается наугад (равновероятно) один шар, фиксируется его цвет, и шар возвращается в урну с одновременным добавлением c новых шаров того же цвета. Затем из урны, содержащей теперь $a + b + c$ шаров, снова случайно извлекается шар и т. д. в течение n шагов.

Математическое ожидание и дисперсия соответственно:

$$Mx = np = n\frac{a}{a+b},\ Dx = \frac{nab(a+cn)}{(a+b)^2(a+b+c)}.$$

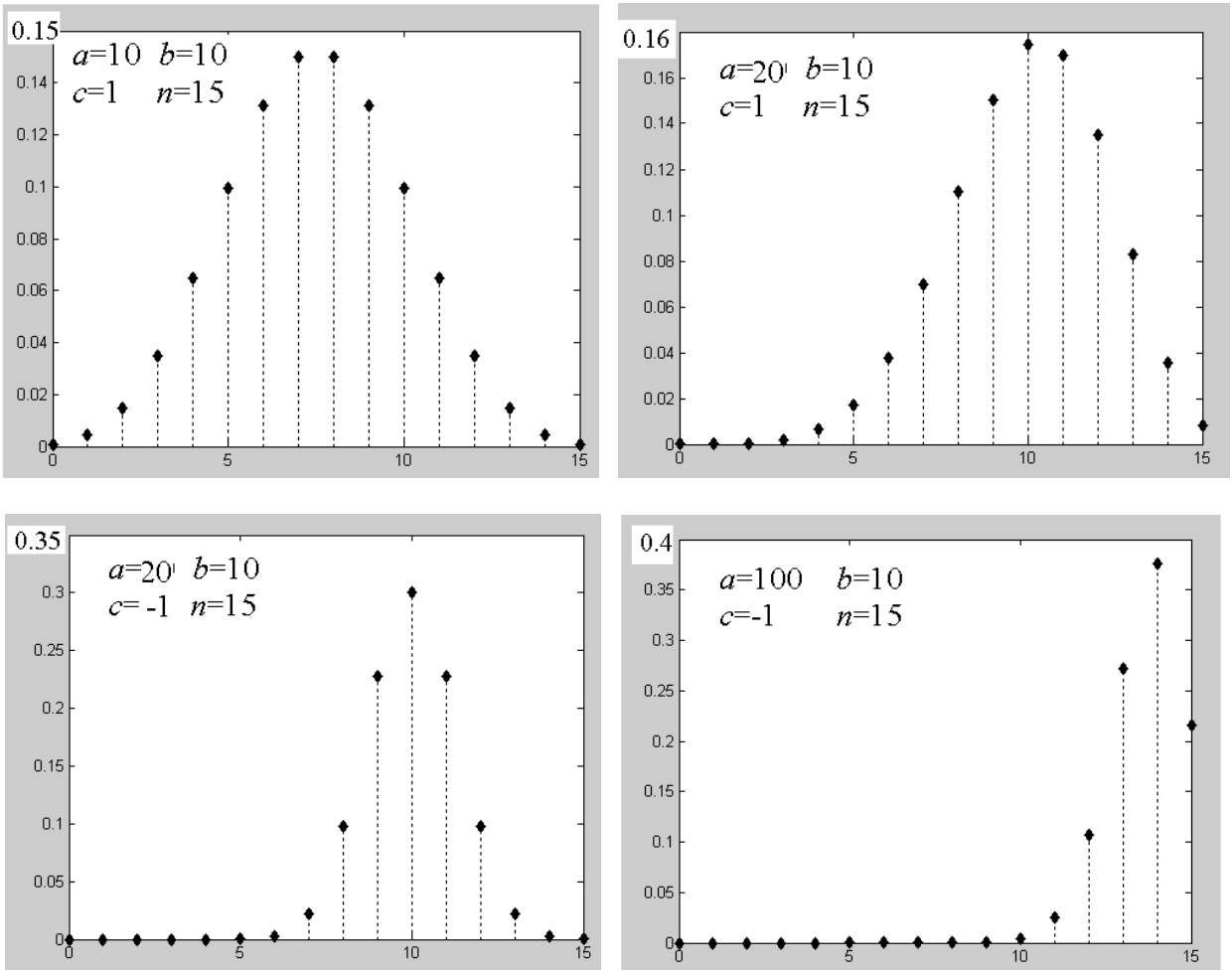


Рис. 2.20. Распределение Маркова – Пойа

Если $c = 0$ (новые шары не добавляются), имеем случайный выбор с возвращением, определяемый биномиальным распределением.

Если $c = -1$ (извлеченный шар в урну не возвращается и процесс извлечения шаров кончается через $n = a + b$ шагов из-за отсутствия шаров в урне), сталкиваемся с предельным случаем гипергеометрического распределения $p(x = a \mid a, b, a + b)$.

При $c > 0$ начинается своеобразная «эпидемия»: если извлекается «больной» шар, при возврате прихватывает с собой еще c зараженных шаров, и при следующем извлечении вероятность извлечь такой шар возрастает. В такой схеме вероятность выбора x белых шаров при n извлечениях ($c \geq 0$, $n > 0$) имеет вид

$$P(x \mid a,b,c,n) = \frac{C_n^x \cdot \prod_{j=0}^{x-1}(a+jc) \cdot \prod_{j=0}^{n-x-1}(b+jc)}{\prod_{j=0}^{n-1}(a+b+jc)}, x = 0,1,2,\ldots .$$

2.3. Распределения особого назначения

В отличие от рассмотренных выше распределений естественного происхождения, здесь упомянем несколько распределений, созданных для проверки статистических гипотез.

2.3.1. Хи-квадрат распределение

Это популярное распределение (рис. 2.21) с k степенями свободы возникает как **распределение** суммы квадратов **независимых стандартизованных нормальных случайных величин** и имеет следующие плотность и функцию распределения:

$$p(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}, \; F(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^{\frac{x}{2}} t^{\frac{k}{2}-1} e^{-t} dt .$$

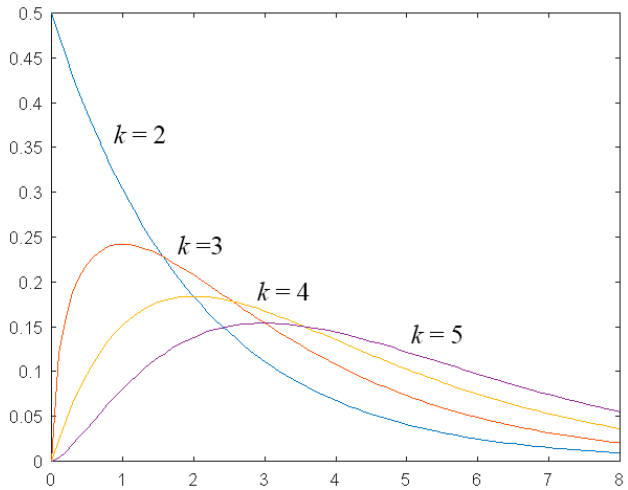


Рис. 2.21. Распределение Пирсона

Математическое ожидание χ^2 (хи-квадрат) распределения $Mx = k$, дисперсия $Dx = 2k$, максимум функции плотности распределения $p(x)$ достигается при $x = k - 2$ (мода Mo при $k > 2$), асимметрия $Ax = 2\sqrt{2/k}$, эксцесс $Ex = 12 / Ax$. Это распределение составляет базу χ^2 критерия согласия Пирсона, в частности, как основа критерия близости распределений.

2.3.2. Распределение Стьюдента

Распределение Стьюдента – одно из популярнейших распределений для оценки доверительных интервалов – распределение отношения **стандартизованных нормальных случайных величин** к корню из величины χ^2, деленной на число степеней свободы k.

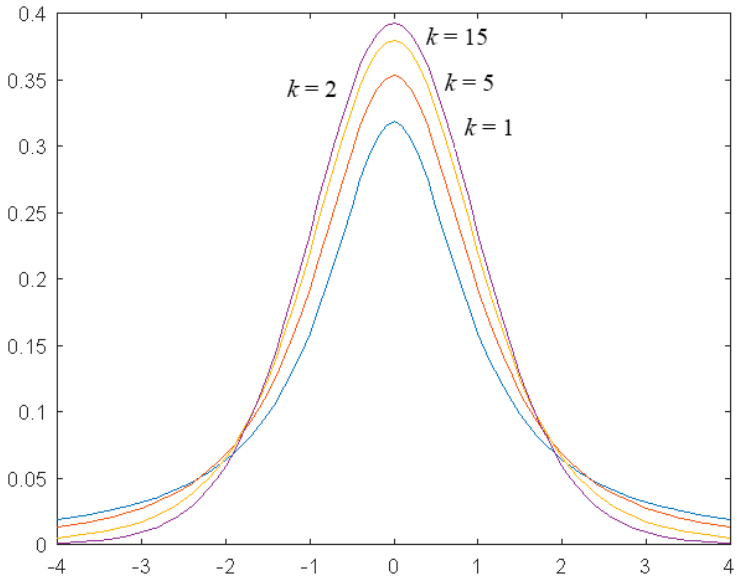


Рис. 2.22. Распределение Стьюдента

Если $y_0, y_1, \ldots, y_k$ – независимые стандартные нормальные случайные величины, то распределение случайной величины $t = \dfrac{y_0}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} y_i^2}}$ называется распределением Стьюдента (рис. 2.22) с k степенями свободы (t выступает как функция от k).

Распределение таких случайных величин непрерывно и имеет плотность

$$p(x) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}.$$

Функция распределения не представима в элементарных функциях, математическое ожидание равно нулю $Mx = 0$, дисперсия $Dx = \dfrac{k}{k-2}$, мода, медиана и асимметрия равны нулю, эксцесс $Ex = 3\dfrac{k-2}{k-4}$.

При $k = 1$ получается так называемое *распределение Коши*:

$$p(x) = \frac{1}{\pi(1+x^2)}, \quad F(x) = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg}(x), \quad |x| < \infty.$$

2.3.3. Распределение Фишера

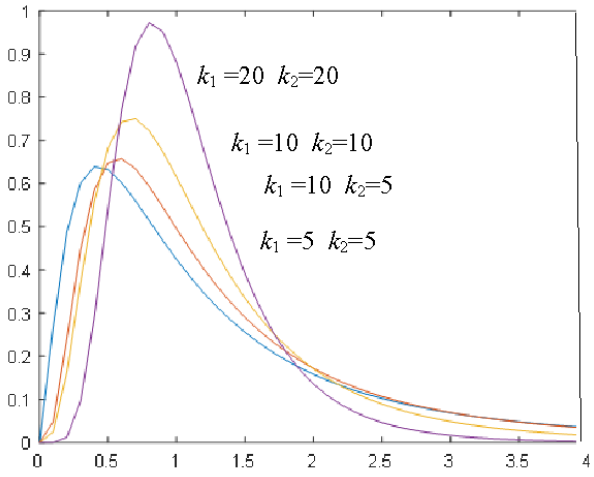


Рис. 2.23. Распределение Фишера

Плотность распределения (рис. 2.23) Фишера (распределения отношения дисперсий случайных хи-квадрат распределенных величин с различным числом степеней свободы k_1 и k_2)

$$p(x) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma(k_1/2)\,\Gamma(k_2/2)} \left(\frac{k_1}{k_2}\right)^{k_1/2} x^{\frac{k_1}{2}-1} \times$$

$$\times \left(1 + \frac{k_1}{k_2} x\right)^{-\frac{k_1+k_2}{2}}.$$

Математическое ожидание $Mx = k_1 / (k_2 - 2)$, дисперсия $Dx = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$, мода $Mo = \frac{k_1 - 2}{k_1}\frac{k_2}{k_2 + 2}$.

2.3.4. Распределение Колмогорова – Смирнова

Распределение Колмогорова – Смирнова предназначено для проверки гипотезы о принадлежности выборки некоторому закону распределения (непараметрические критерии).

Его функция распределения

$$F(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}, x > 0 \quad .$$

2.4. Эмпирические распределения и критерий согласия

Пусть в результате некоторого эксперимента получена выборка случайных значений объема N. Для упрощения технологии построения выборочного (эмпирического) распределения исходную последовательность значений случайной величины $\{x_i, i = 1, 2, \ldots, N\}$ упорядочиваем по возрастанию, получая так называемый *вариационный ряд*.

Интервал $[x_{\min} - \varepsilon, x_{\max} + \varepsilon]$ разбиваем на k подынтервалов длиной Δ. Подсчитываем число элементов исходной выборки, попавших в каждый из подынтервалов $\{n_j, j = 1, 2, \ldots, k\}$, и соответствующие эмпирические вероятности попадания в них $\{p_j = n_j / N, j = 1, 2, \ldots, k\}$, на основе которых и строится эмпирическое распределение с эмпирической плотностью $p^*(x)$ и эмпирической функцией (*кумулятой*) $F^*(x)$

$$F_j^* = \sum_{i=1}^{j} p_j^*, \; j = 1, 2, \ldots, k.$$

Знание $x_{\min}$, $x_{\max}$, Δ и указанных функций позволяет строить их гистограммы (рис. 2.24), привязанные к серединам соответствующих интервалов, и находить канонические оценки (моменты, медиану, моду).

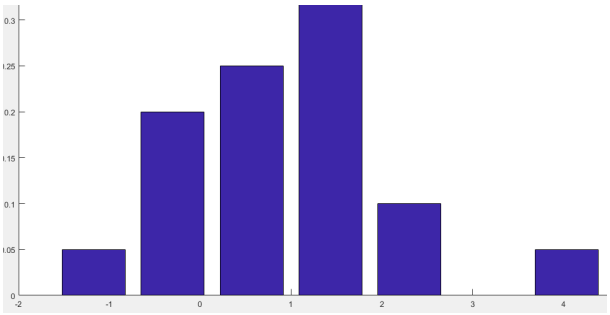


Рис. 2.24. Гистограмма эмпирического распределения

Простота такого поиска (оценки) эмпирического распределения кажущаяся, ибо при малом объеме выборки изменения в выборе k полностью меняют представления о характере искомых функций, в такой ситуации приходится даже задуматься о включении пограничных значений. При малых

значениях k трудно понять структуру плотности распределения, при больших k и относительно небольших N в отдельных подынтервалах обнаружится ничтожно малое число попаданий или их отсутствие (нулевая вероятность). Соответственно выбираем k интуитивно (так, чтобы в каждом подынтервале было не меньше 3 попаданий) или руководствуясь формулой Стерджесса, по которой можно получить первоначальную оценку k:

$$k = 1 + 3{,}3219 \ln(N) \text{ (здесь } 3{,}3219 = \log_2 10).$$

Если задуматься о последующем использовании найденной информации, то едва ли разумно хранить таблицы значений $p^*(x)$. Хотелось бы аппроксимировать найденное распределение каким-то из известных непрерывных распределений, созданных за многовековую историю практики статистического анализа.

Надежность выводов при анализе статистической выборки существенно зависит от типа распределения (для многих физических величин нормальное распределение неприемлемо). Тем не менее при больших выборках делается предположение об асимптотической нормальности выборочных значений (рис. 2.25) и для проверки гипотезы о нормальности эмпирической выборки привлекается известный хи-квадрат критерий согласия Пирсона.

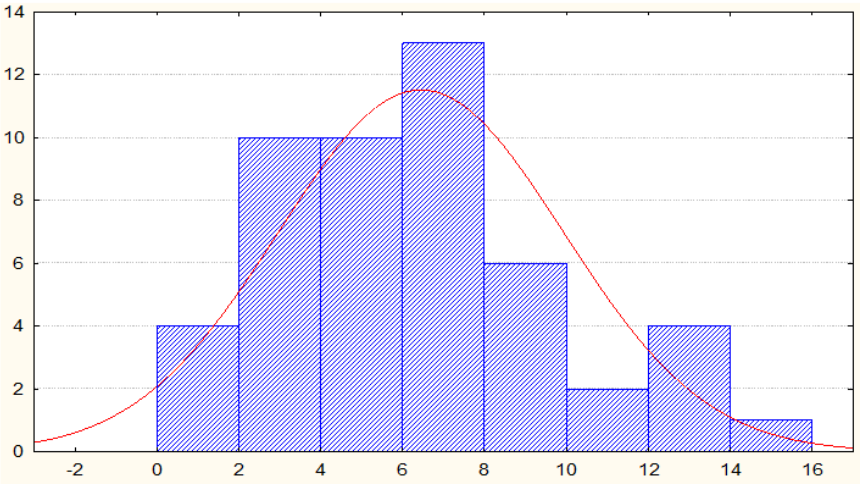


Рис. 2.25. Аппроксимация эмпирического распределения гауссовым

Для оценки близости эмпирического и некоторого теоретического распределений с известной (вычисляемой) плотностью $p(x)$ берется величина

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - Np_j)^2}{Np_j},$$

где N – объем выборки; k – число подынтервалов; n_j – число попаданий в j-й подынтервал; p_j – теоретическая вероятность принадлежности этому подынтервалу (G_{j-1}, G_j):

$$p_j = \int_{G_{j-1}}^{G_j} p(x)dx \quad .$$

Эта величина (χ^2) подчиняется хи-квадрат распределению Пирсона с $k-1$ степенями свободы и задает предельное значение, при котором гипотеза о близости распределений отвергается с той или иной вероятностью ошибки α.

Обратившись к нижеприведенной табл. 2.2 квантилей распределения Пирсона, обнаруживаем, что при $k=5$ эта оценка превышает 11 и гипотеза о согласованности распределений отвергается с вероятностью ошибки в 5 %.

Таблица 2.2. Таблица квантилей распределения Пирсона

α / k	0,950	0,750	0,250	0,100	0,050	0,010	0,005
1	0,00393	0,10153	1,32330	2,70554	3,84146	6,63490	7,87944
2	0,10259	0,57536	2,77259	4,60517	5,99146	9,21034	10,5966
3	0,35185	1,21253	4,10834	6,25139	7,81473	11,3448	12,8381
4	0,71072	1,92256	5,38527	7,77944	9,48773	13,2767	14,8602
5	1,14548	2,67460	6,62568	9,23636	**11,0705**	15,086	16,7496
6	1,63538	3,45460	7,84080	10,6446	12,5916	16,8119	18,5475
7	2,16735	4,25485	9,03715	12,0170	14,0671	18,4753	20,2777
8	2,73264	5,07064	10,2188	13,3615	15,5073	20,0902	21,9549
9	3,32511	5,89883	11,3887	14,6836	16,9189	21,6659	23,5893
10	3,94030	6,73720	12,5488	15,9871	18,3070	23,2092	25,1881
15	7,26094	11,03654	18,24509	22,30713	24,99579	30,57791	32,80132
20	10,85081	15,45177	23,82769	28,41198	31,41043	37,56623	39,99685
25	14,61141	19,93934	29,33885	34,38159	37,65248	44,31410	46,92789
30	18,49266	24,47761	34,79974	40,25602	43,77297	46,97924	50,89218

Иногда предлагается подынтервал с $N_j < 3$ объединять со смежным, тем самым меняя границы G_j укрупненных подынтервалов.

Для сопоставления двух распределений может быть использована проверка $\sup_x |p_1(x) - p_2(x)| < \varepsilon$ и, по крайней мере, вылавливание выбросов.

2.5. Генерация случайных величин и численное решение задач средствами MS Excel

Генерацию случайных чисел в среде MS Excel можно осуществить с помощью надстройки «Пакет анализа» (Data Analysis) и некоторых статистических функций рабочего листа. Для обращения к `Пакету анализа` необходимо выполнить команду `Данные – Анализ данных`. `Пакет анализа` позволяет решать в диалоговом режиме 19 различных задач, наиболее часто встречающихся в классической математической статистике (рис. 2.26) (хотя это и не мешает подходить к ним творчески).

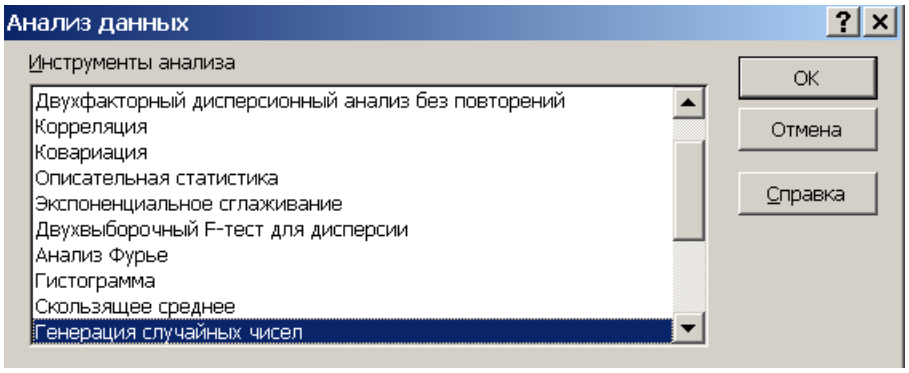


Рис. 2.26. Инструменты анализа данных

Генерация случайных чисел заполняет диапазон случайными величинами, заданными по одному из законов распределения: равномерному; нормальному; Бернулли (случайная величина X принимает значение 1 с вероятностью p и 0 с вероятностью $(1-p)$ (индикаторная случайная величина); биномиальному; Пуассона; модельному (позволяющему генерировать последовательности случайных чисел от a до b с шагом c, и возможностью повторения каждого числа и последовательности); дискретному (решающему задачу получения по имеющемуся распределению новых значений того же распределения).

Замечание. Инструмент генерации случайных чисел позволяет решить целый ряд задач численных методов (например, приближенное вычисление определенных интегралов методом статистических испытаний – методом Монте-Карло); имитационное моделирование изучаемых процессов и т. д.

Пример. Вычислить определенный интеграл $J = \int_0^1 x^2 dx$ методом Монте-Карло.

Решение. Вычисление определенного интеграла J равносильно нахождению площади D криволинейной трапеции под графиком функции $Y = f(X) = X^2$ (рис. 2.27).

Рассмотрим систему двумерных равномерно распределенных случайных величин (X, Y) на интервале от 0 до 1. При достаточно большом числе опытов N площадь D будет приближенно равна относительной частоте попадания точек $M_i(x_i, y_i)$ в область D (в силу закона больших чисел):

$$J \approx \frac{n}{N}.$$

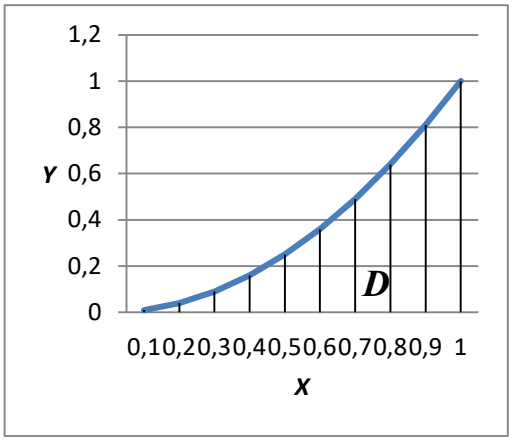


Рис. 2.27. Область D

Для генерации систем двух равномерно распределенных на интервале от 0 до 1 случайных величин используем инструмент `Генерация случайных чисел`. Заполним диалоговое окно для генерации 1000 пар указанных случайных чисел (рис. 2.28). В результате этого в диапазоне `A2:B10001` получим искомые пары случайных чисел. В ячейку `C2` введем формулу `=A1^2`; в ячейку `D2` `=Если(B2>C2;0;1)`. Последняя формула помещает в ячейку значение 0, если точка M_i не попадает в область D и значение 1 в противном случае.

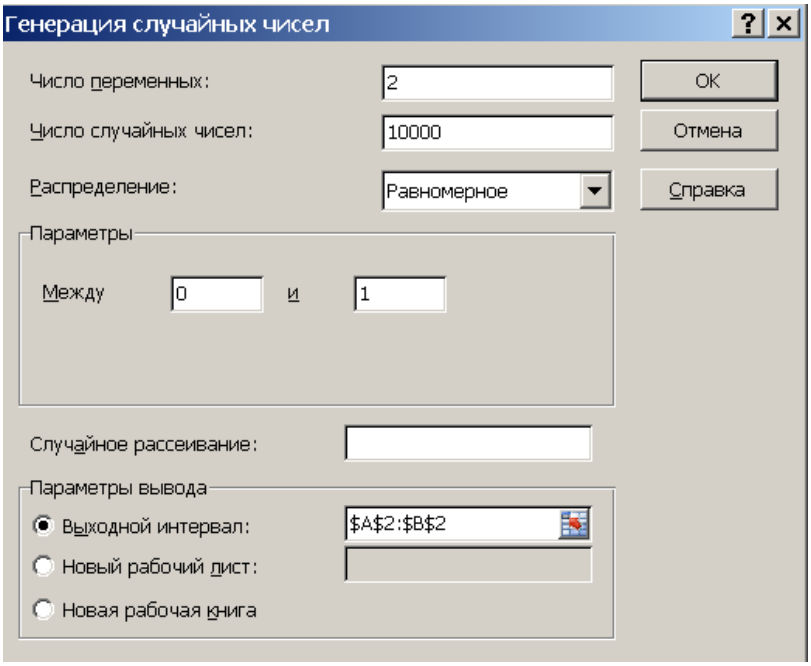


Рис. 2.28. Диалоговое окно генерации двумерных равномерно распределенных случайных величин

Выделим диапазон `C2:D2` и скопируем вниз до строки 10001. Найдем сумму значений в диапазоне `D2:D10001`, в результате получим, что $n = \sum m_i = 3337$ (рис. 2.29). Отсюда $J \approx \frac{n}{N} = \frac{3337}{10000} = 0{,}3337$. Применяя неравенство Чебышёва, имеем

$$P\left(\left|\frac{n}{N} - J\right| < \varepsilon\right) \geq 1 - \frac{J(1-J)}{\varepsilon^2 N} \geq 1 - \frac{1}{4\varepsilon^2 N}. \qquad (2.4)$$

Если мы зададим уровень значимости α, то неравенство, приведенное выше, будет всегда верно с гарантийной вероятностью $p = 1 - \alpha$ при $\alpha = \frac{1}{4\varepsilon^2 N}$. При заданных значениях ε и α можно определить необходимое число испытаний $N = \frac{1}{4\varepsilon^2 \alpha}$. В силу того, что неравенство Чебышёва дает нижнюю оценку вероятности, значение N будет завышено, например, в нашем случае при $\varepsilon = 0{,}001$ и $\alpha = 0{,}01$ получится $N = 25000000$. Точное значение $J = 0{,}33(3)$. В рассматриваемом примере точность $\varepsilon = 0{,}001$ достигается уже при 10000 испытаний. Существуют более точные методы оценки значения N, основывающиеся на предельных теоремах теории вероятностей.

D2 | fx | =ЕСЛИ(B2>C2;0;1)

	A	B	C	D	E	F
1	Xi	Yi	f(Xi)	mi		
2	0,382	0,100681	0,145924	1		
3	0,596484	0,899106	0,355793	0		
4	0,88461	0,958464	0,782534	0		
5	0,014496	0,407422	0,00021	0		
6	0,863247	0,138585	0,745195	1		
7	0,245033	0,045473	0,060041	1		
8	0,03238	0,164129	0,001048	0		
9	0,219611	0,01709	0,048229	1		
10	0,285043	0,343089	0,081249	0		
9999	0,878994	0,310404	0,772631	1		
10000	0,589068	0,885311	0,347001	0		
10001	0,54677	0,571673	0,298957	0		
10002	Итого	-	-	3367		

Рис. 2.29. Результат применения метода Монте-Карло

Контрольные вопросы

1) Чем отличаются друг от друга дискретные и непрерывные случайные величины?

2) Каким образом связаны функции плотности и распределения вероятностей случайной величины?

3) В чем заключается принцип перехода от вычисления параметров распределения случайной величины к вычислению оценок для выборочного распределения?

4) В чем отличие смещенных от несмещенных выборочных оценок параметров распределения случайной величины?

5) Какие случайные числа служат базой для применения методов Монте-Карло?

6) Каким образом строится эмпирическое распределение случайной величины?

7) Каким образом моделируются случайные величины с известным законом распределения?

8) Что представляет собой нормальное распределение случайной величины?

9) Каким образом распределение Пирсона связано с нормальным распределением?

10) Каким образом распределение Стьюдента связано с нормальным распределением и распределением Пирсона?

11) Как распределение Фишера связано с распределением Пирсона?

12) Каким образом интерпретируется геометрическое распределение?

13) Что представляет собой биномиальное распределение?

14) Какое распределение называется распределением Пуассона?

15) Что представляет собой стационарный пуассоновский поток?

16) В чем заключаются принципы моделирования случайных величин с заданным законом распределения?

17) С помощью чего моделируются случайные величины в MS Excel?

18) Какие задачи можно решать с помощью инструмента генерации случайных чисел?

Глава 3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Основные понятия

В книге [1] приведена цитата из «Золотого ключика» А. Н. Толстого:

…Сова приложила ухо к груди Буратино.

– Пациент скорее мертв, чем жив, – прошептала она и повернула голову назад на 180°.

Жаба долго мяла влажной лапой Буратино ... прошептала:

– Пациент скорее жив, чем мертв...

Народный лекарь Богомол прошелестел:

– Одно из двух, или пациент жив, или он умер...

Все три мудрых врачевателя делают свои выводы (строят гипотезы) на основе имеющегося опыта, статистики суждений о признаках жизни.

Немного об общепринятой терминологии.

Статистической гипотезой называют предположение о свойствах распределения или параметрах некоторой случайной величины, которые приходится проверять, опираясь на данные случайной выборки из теоретически бесконечной (в идеале) *генеральной совокупности*. Другими словами, итог проверки статистической гипотезы состоит в том, чтобы *по данным случайной выборки решить, гипотеза принимается или отклоняется с минимальным риском (вероятностью) ошибки*. В литературе принято проверяемую гипотезу называть «*нулевой гипотезой*» и обозначать H_0. К нулевым относят гипотезы, утверждающие, что *различие между сравниваемыми величинами отсутствует*, а наблюдаемые отклонения объясняются лишь случайностью. Остальные гипотезы, противопоставляемые H_0, называются *альтернативными* (конкурирующими) и обозначаются H_1.

Случайная величина K, определяющая условия для принятия или отклонения H_0, называется *статистическим критерием*. Совокупность W значений критерия, при которых гипотезу отвергают, называют *критической областью* ω (рис. 3.1).

Если гипотезу отклоняют, тогда как она верна, совершают *ошибку первого рода*, вероятность которой $\alpha = P(K \in \omega \mid H_0)$ называется *уровнем значимости*. Пограничное значение K, сопоставляемое вероятности α, называют критической точкой K_α (по неписаной традиции в эконометрике выбирают $\alpha = 0{,}05$).

Если гипотезу H_0 не отклоняют, а она неверна, совершают *ошибку второго рода* с вероятностью $\beta = P(K \in W - \omega \mid H_1)$. Величина

$1 - \beta = P(K \in \omega \mid H_1)$ определяет вероятность справедливого отклонения гипотезы H_0 и называется *мощностью критерия*.

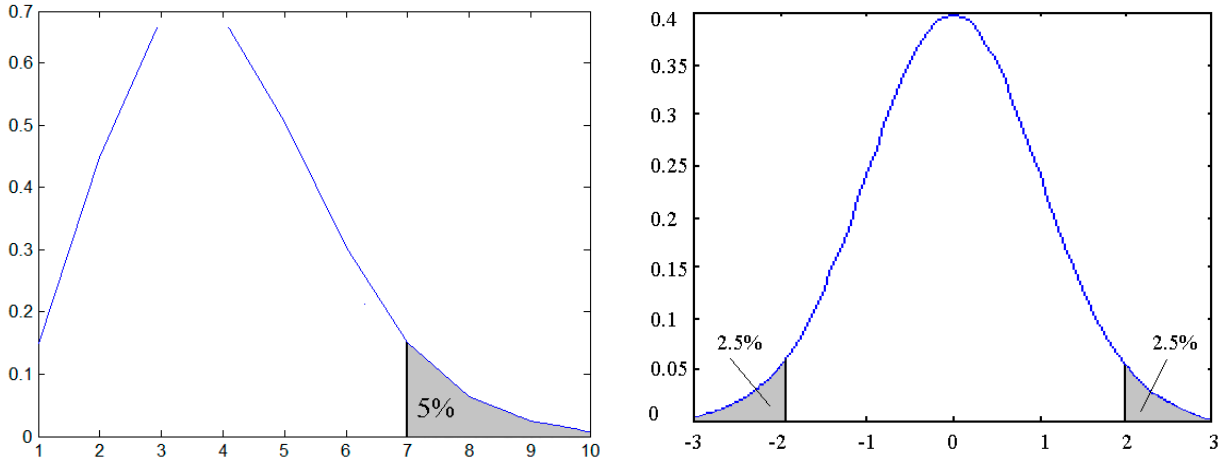


Рис. 3.1. Критическая область

При построении статистических критериев вероятности обоих типов ошибок α и β желательно свести к минимуму. Однако уменьшение ошибки одного типа ведет к увеличению ошибки другого типа. Единственный способ уменьшить одновременно ошибки обоих типов состоит в увеличении объема выборки, что на практике может оказаться слишком сложным делом.

Гипотеза называется *простой*, если она утверждает, что некий параметр θ может принять одно значение (например, $\theta = \theta_0$). *Сложной* считается гипотеза, утверждающая, что параметр может принять некоторые значения из заданного множества (например, $\theta > \theta_0$).

Сложная гипотеза может быть односторонней или двусторонней (гипотеза $H_1 : \theta \neq \theta_0$ по существу представляет собой альтернативу $\theta < \theta_0$ – левосторонняя альтернатива и $\theta > \theta_0$ – правосторонняя).

При двустороннем критерии границы критической области (рис. 3.1) выбираются так, чтобы вероятности попадания в левую и правую части были бы одинаковы и равны $\alpha / 2$ (при одностороннем область содержит всего одну зону, вероятность попадания в которую равна α). Об этом мы уже упоминали при знакомстве с понятием квантили.

Для проверки гипотезы о том или ином явлении необходимо наличие эталона. Например, при оценке рассеивания по дальности при стрельбе (заведомо нормальное распределение) при больших n достаточно нормировать данные и найти процент отклонений от нуля более чем 3. Даже не прибегая к таблицам, можно осознать, что он не должен превышать 5 % при «нормальной» стрельбе (правило трех сигм).

Если некто не увидел более 5 грамматических ошибок на журнальной странице, то для приема его на работу корректором журнала необ-

ходима предварительная установка предельно допустимого уровня. Сто лет назад знаменитое книгоиздательство И. Д. Сытина отвергло бы претендента при одной ошибке на 5 страниц. Увы, сегодня при такой требовательности и реальном уровне грамотности населения, отвергнув такого претендента, можно остаться вообще без корректора.

Приведем несколько примеров проверки статистических гипотез.

3.2. Гипотезы относительно биномиальной вероятности

Как решить вечный вопрос: кто умнее, мужчины или женщины? Нет объективного критерия, как отличить умного человека от «дурака» (в сумасшедших домах есть по-своему гениальные люди, а политический деятель, изрекающий явные глупости, на поверку провокатор, но не дурак), потому оценивается *IQ* (коэффициент интеллекта) мужчин и женщин и выясняется, случайна ли обнаруженная разница ($H_0 : IQ_м = IQ_ж$) или нет ($H_1 : IQ_м \neq IQ_ж$)? Обнаруженная разница случайна или есть основания ее отвергнуть?

Пусть, например, получена *большая* статистика из $n = 1000$ ответов, где женщины превзошли мужчин в $m = 511$ случаях.

Здесь имеем дело лишь с двумя возможными исходами (согласно гипотезе равновероятными). Поэтому естественно обратиться к биномиальному распределению (см. п. 2.2.1) при $p = 0{,}5$

$$P(x = m) = C_n^m p^m (1-p)^{n-m} = C_n^m 0{,}5^n;\ \mu = n\,p = 500;\ \sigma^2 = n\,p(1-p) = 250,$$

и оценить вероятность

$$P(x \le m) = \sum_{x=0}^{m} C_n^x p^x (1-p)^{n-x} = \sum_{x=0}^{m} C_n^x\, 0{,}5^{1000},$$

выбрать «порог» вероятности 0,95 (или иной) и проверить условие $P(x \le m) < 0{,}95$. При его выполнении нет оснований принять гипотезу превосходства женщин с вероятностью 95 %. В противном случае принимаем гипотезу превосходства на 5 % уровне.

В силу симметрии распределения при больших m проще проверить

$$P(x \ge m) = \sum_{x=0}^{n-m} C_n^x p^x (1-p)^{n-x} = \sum_{x=0}^{n-m} C_n^x\, 0{,}5^{1000} < 0{,}05.$$

Порядок возникающих здесь величин ($0{,}5^{1000}$; $9{,}3\cdot10^{-302}$) таков, что простой расчет $\sum_{x=1}^{m} C_{1000}^x$ через факториалы нереален и нужны хитроумные приемы типа $P(x = 511) = C_{1000}^{511}\ 0{,}5^{1000} = \prod_{i=1}^{489}\left(\frac{1}{4}\frac{1001-i}{i}\right) 0{,}5^{22} = 0{,}0198.$

Пороговые значения в таблицах (табл. 3.1) для подобного объема выборки отсутствуют, хотя при $n > 120$ они достаточно близки.

При больших выборках возможна аппроксимация биномиального распределения нормальным. Для этого подвергаем найденное m нормировке $t = \dfrac{m - \mu}{\sigma} = \dfrac{m - n\,p}{\sqrt{n\,p(1-p)}} = 0{,}696$. Обратившись к таблицам нормального распределения для уровня значимости $1 - \alpha = 0{,}95$, обнаруживаем $0{,}696 < t_{крит} = 1{,}96$. Вывод: *нет оснований отвергать нулевую гипотезу* (обратите внимание на мягкость и осторожность вывода в этой формулировке).

Если бы превосходство женщин обнаружилось в 535 или более случаях, то $t = 2{,}21 > t_{крит} = 1{,}96$ и мужчинам пришлось бы ссылаться на неравные условия или на принцип «сила есть – ума не надо».

Таблица 3.1. Доверительные пределы вероятности биномиального распределения

n*\ *p	**0,5**	**0,9**	**0,95**	**0,99**
1	1,000	6,31	12,7	63,7
2	0,816	2,92	4,30	9,92
3	0,765	2,35	3,18	5,84
4	0,741	2,13	2,77	4,60
5	0,727	2,02	2,57	4,03
6	0,718	1,943	2,45	3,71
7	0,711	1,895	2,36	3,50

n*\ *p	**0,5**	**0,9**	**0,95**	**0,99**
8	0,706	1,860	2,31	3,36
9	0,703	1,833	2,26	3,25
10	0,700	1,812	2,23	3,17
15	0,691	1,746	2,13	2,95
25	0,684	1,708	2,06	2,79
60	0,679	1,671	2,00	2,66
120	0,677	1,658	1,98	2,62

3.3. Гипотезы относительно полиномиальных вероятностей и критерий хи-квадрат

Проверка *гипотезы относительно полиномиальных вероятностей* много сложнее. Если в биноминальном случае лишь два исхода, то здесь число исходов равно k с вероятностями p_i, $i = 1, 2, \ldots, k$. Нулевая гипотеза предполагает совпадение ожидаемой вероятности p_i с эмпирической n_i / n (n_i – число исходов i-го типа).

При больших n ($n > 25$) проверка базируется на аппроксимации полиномиального распределения χ^2-распределением. Отыскивается оценка $\chi^2 = \sum_{i=1}^{k} \dfrac{(n_i - n\,p_i)^2}{n\,p_i}$ и выясняется, не превосходит ли она, например, α%-ю точку χ^2-распределения с $k - 1$ степенями свободы.

Например, при 50-кратном бросании кости выпали 12 шестерок, 9 пятерок, 9 четверок, 6 троек, 9 двоек и 5 единиц. Проверяя гипотезу

о том, что кость «правильная», получаем оценку $\chi^2 = 3{,}76$. Сравнение с табличными значениями при пяти степенях свободы показывает незначимость оценки и не позволяет объявить кость фальшивой.

Для других случайных величин проверка гипотез является более изощренной. В литературе приведены десятки подходов для задач приема промышленной продукции, контроля качества и т. п.

3.4. Критерий Стьюдента

Критерий Стьюдента предназначен для сравнения средних M_x и M_y двух *нормально распределенных* случайных величин, полученных по выборкам объема n и m, на базе t-распределения. Здесь выдвигается гипотеза $H_0 : M_x = M_y$, справедливая, если разница между выборочными средними незначима и объясняется случайным выбором элементов.

Существуют несколько вариантов выбора критерия.

Гипотезы: $H_0 : M_x = M_y$ и $H_1 : M_x \neq M_y$. В силу симметрии распределения ищут квантиль $t_{кр}$ с учетом не α, а $\alpha / 2$. Если эмпирическое $|t_{эмп}| < t_{кр}$, то нет оснований отвергнуть нулевую гипотезу, в противном случае нет оснований ее принять.

Гипотезы: $H_0 : M_x = M_y$, $H_1 : M_x > M_y$. В этом случае строят одностороннюю (правостороннюю) критическую область.

Гипотеза $H_0 : M_x = R$, где R – некоторая константа. Находят значение $t = \dfrac{M_x - R}{\sigma_x}\sqrt{n}$ (число степеней свободы $f = n - 1$).

Гипотеза $H_0 : M_x = M_y$ в предположении равенства дисперсий $\sigma_x^2 = \sigma_y^2$. Находят

$$t = \frac{M_x - M_y}{g}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}},\ g = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n + m - 1}},$$

где s_x^2, s_y^2 – несмещенные оценки дисперсий, число степеней свободы равно $f = n + m - 1$.

При той же гипотезе в предположении неравенства дисперсий $\sigma_x^2 \neq \sigma_y^2$ (f – число степеней свободы) отыскивается

$$t = \frac{M_x - M_y}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}},\ f = \left[\frac{\left(\frac{s_x^2}{n}+\frac{s_y^2}{m}\right)^2}{\left(\frac{s_x^2}{n}\right)^2 \Big/ {}_{n+1} + \left(\frac{s_y^2}{m}\right)^2 \Big/ {}_{m+1}} - 2\right].$$

В предположении неизвестных дисперсий при равенстве объемов выборок требуется поиск $t = \frac{D}{S_D}\sqrt{n}$,

где $D = M_x - M_y$, $S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(|y_i - x_i| - D)^2}$, $f = n - 1$.

3.5. Критерий Фишера

Критерий Фишера, наряду с другими многочисленными приложениями, применяют для проверки величины дисперсии случайной переменной и проверки равенства дисперсий двух аналогичных выборок случайных чисел.

Пусть построена выборка объема n для некоторой случайной величины с нормальным распределением $N(\mu, \sigma^2)$. Получив несмещенную оценку выборочной дисперсии S^2, отыскиваем величину

$$\chi^2 = (n-1)\frac{S^2}{\sigma^2},\ f = n - 1$$

и сравниваем ее, например, с 97,5%-й квантилью при проверке гипотезы о равенстве дисперсий или с 95%-й при гипотезе $S^2 \geq \sigma^2$ или $S^2 \leq \sigma^2$ (следует подчеркнуть необходимость нормальности эмпирического распределения).

Например, в конструкторском бюро установлено, что кучность стрельбы нового автомата по дальности подчинена нормальному закону распределения $N(\mu = 0, \sigma = 7)$. Измерения производились в сантиметрах. При попытке серийного производства на заводе было проведено 11-кратное испытание с полученными отклонениями r_i. При этом посчитанная оценка дисперсии (относительно центра мишени) оказалась $S^2 = \frac{1}{11-1}\sum_{i=1}^{11} r_i^2$ = 169 см^2. Насколько вероятно столь значительное отклонение от эталона? Отыскав значение $\chi^2 = 10 \cdot 169 / 49 = 34{,}5$, из таблицы χ^2-распределения при 10 степенях свободы видим, что даже при вероятности ошибки 0,5 % табличное значение равно 25,2 < 34,5. Вывод: с уверенностью 99,5 % можно утверждать, что на заводе нарушена технология производства (ссылки на малое число стрельб безосновательны).

Пусть имеются две выборки объема n_1 и n_2 с выборочными дисперсиями S_1^2 и S_2^2 из двух нормально распределенных генеральных совокупностей с дисперсиями σ_1^2 и σ_2^2. Выдвинув гипотезу равенства дисперсий ($H_0 : \sigma_1^2 = \sigma_2^2$), вычисляем значение критерия Фишера по формуле $F = R_1 / R_2$ ($R = S / \sigma$). Если значение критерия $F > 1$, берем число степеней свободы $p = n_1 - 1$, $q = n_2 - 1$. В противном случае берем критерий

равным $1 / F$ и $p = n_2 - 1$, $q = n_1 - 1$. После чего остается обратиться к таблицам F_{pq}-распределения для заданного уровня значимости.

Например, пусть объемы выборок равны 11 и 21, а оценки дисперсий оказались равными 5,9 и 11,2. Поскольку отношение $5{,}9 / 11{,}2 < 1$, берем обратное значение $F = 11{,}2 / 5{,}9 = 1{,}90$ и $p = 20$, $q = 10$. На 5%-м уровне $1{,}90 < F_{20,10} = 2{,}35$, и нет оснований утверждать, что дисперсия первой совокупности меньше дисперсии второй.

3.6. Критерий Колмогорова – Смирнова

Критерий Колмогорова – Смирнова, наряду с другими, предназначен для проверки гипотезы H_0 о том, что случайная величина x распределена по закону $p(x)$ на основании выборки объема n.

В качестве критерия выбирается максимальная разница между функциями плотности гипотетического и эмпирического распределений

$$D = \max_x \left| p(x) - p_{\text{эмп}}(x) \right|.$$

Доказано, что для любой функции $p(x)$ при неограниченном росте числа испытаний n вероятность неравенства $D\sqrt{n} \ge \lambda$ стремится к предельной функции распределения Колмогорова – Смирнова

$$K(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}, \lambda > 0.$$

Поэтому для проверки гипотезы H_0 вычисляют $\lambda = D\sqrt{n}$ и значение $K(\lambda)$. Если $1 - K(\lambda) < \alpha$, где α – уровень значимости, то гипотезу о равенстве функций распределений нет оснований отвергать (в противном случае нет оснований ее принять).

Этим же критерием можно пользоваться и для проверки гипотезы о совпадении распределений двух случайных величин по выборкам с объемами n и m:

$$S_D = D = \max_x \left| p_1(x) - p_2(x) \right|, \quad \lambda = D\sqrt{\frac{mn}{m+n}}.$$

3.7. Непараметрические критерии

Для проверки гипотезы о близости распределений по выборкам x_i, y_i $(i = 1, 2, ..., n)$ может быть использован *критерий знаков*. Здесь отыскиваются разности $r_i = x_i - y_i$, которые можно рассматривать как случайные величины с двумя исходами (+ или –) с вероятностями 0,5 (нулевые разности можно исключить). Отношение числа положительных (или отрицательных) исходов к объему выборки может характеризовать степень доверия к гипотезе $H_0 : M_x = M_y$.

Мало чем, по существу, отличается *критерий Манна – Уитни*. Обозначим буквой m наименьшее из количеств соответствующих (положительных или отрицательных) исходов. Вероятность того, что случайная величина примет m значений одного знака в предположении гипотезы близости распределений, равна $p = C_n^m / 2^n = \frac{1}{2^n} \frac{n!}{m!(n-m)!}$.

И в случае $p < \alpha$ нет оснований отвергать гипотезу.

Для проверки гипотезы о равенстве средних при отсутствии гарантированной нормальности распределений применяют *критерий Уилкоксона*. Так, если выясняется результат воздействия некоторого мероприятия на значение какого-то показателя (медицинского препарата на вес человека, приказа ректора на посещаемость занятий студентами и т. п.), отыскиваются разности между значениями до и после и выдвигается гипотеза: распределение разностей симметрично относительно нуля.

Для реализации критерия абсолютные величины разностей ранжируются в порядке возрастания (ранги от 1 до n). Каждому рангу присваивается знак соответствующей разности и отыскивается сумма положительных рангов U (можно отрицательных). Если эта сумма превышает критическое значение (табл. 3.2), гипотеза отвергается.

Критерий Уилкоксона можно использовать при небольшом объеме выборки (до 25), поскольку при большом n распределение значений этого критерия приближается к нормальному.

Таблица 3.2. Значения критерия Уилкоксона при $\alpha = 0{,}05$

n	7	8	9	10	11	12	13	14	15	16	17	18	19	20
U	25	31	37	45	53	61	70	80	90	101	112	124	137	150

При большом объеме выборки аналогично U отыскивается величина $T = \frac{U - \frac{n(n+1)}{4}}{\sqrt{n(n+1)\frac{2n+1}{24}}}$, сравниваемая с квантилью нормального распределения (если $|T| > 1{,}96$, нулевую гипотезу с 95%-й гарантией следует отклонить).

Идеологически близок к критерию знаков *критерий медианы*. Пусть имеется выборка объема n и установлено наличие в ней m «хороших» (плюсов) и $n - m$ «плохих» (минусов) наблюдений. При больших n применяют критериальную оценку $T = \frac{2m-n}{\sqrt{n}}$ в сопоставлении с квантилью нормального распределения.

Существует множество других непараметрических критериев, связанных с анализом эффектов условий эксперимента, – критерии *Фридмана*, *Пейджа* и другие.

Контрольные вопросы

1) В чем заключается основной смысл проверки статистических гипотез?

2) Что принято считать нулевой гипотезой H_0?

3) Что собой представляет статистический критерий?

4) Что такое критическая область.

5) В чем заключается ошибка первого рода при проверке статистических гипотез?

6) Что представляет собой ошибка второго рода при проверке статистических гипотез?

7) Чем различаются простые и сложные гипотезы?

8) В чем различие двусторонней и односторонней гипотез?

9) Каким образом используются критическое значение при проверке статистических гипотез и уровень значимости?

10) В чем заключается гипотеза относительно биномиальной вероятности?

11) Что представляет собой гипотеза относительно полиномиальных вероятностей?

12) Для чего используется критерий Стьюдента?

13) Какие гипотезы проверяются с помощью критерия Фишера?

14) В чем заключается смысл критерия Колмогорова – Смирнова?

15) Для проверки каких гипотез используется критерий Пирсона χ^2?

16) В чем заключается принципиальное отличие непараметрических критериев?

17) Для чего используется критерий знаков?

18) Какие гипотезы проверяются с помощью критерия Манна – Уитни?

19) Для проверки каких гипотез используется критерий медианы?

20) В каких случаях для проверки гипотезы о равенстве средних используется критерий Уилкоксона?

Глава 4. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Факт взаимосвязей между явлениями в природе и обществе не вызывал сомнений даже у древних обитателей дельты Нила и междуречья Тигра и Евфрата XIV века до нашей эры. Почему-то возникали пожары, эпидемии, нашествия саранчи и солнечные затмения, «крокодил не ловился и не рос кокос». Чаще всего предпочитали искать причины в воле неких богов, общение с которыми обычно поручалось избранным. Творческое сотрудничество физиков, математиков и представителей других естественных наук, пройдя через поиск искусственного золота, эликсира бессмертия и гомункулуса, постепенно установило базовые принципы и соотношения в сфере электротехники, механики, радиологии и даже генетики. Конечно, до сих дней бытует мнение, что «не повезет, если черный кот дорогу перейдет», «сглазили», «наслала порчу» и подобные суеверия. Находятся и те, кто буквально понимают написанное В. Маяковским и повторенное в «Маленьком принце» Антуаном де Сент-Экзюпери: «… если звезды зажигают – значит – это кому-нибудь нужно». В обыденной жизни нет гарантий отсутствия так называемого *белого шума*, сопровождающего даже прописные истины. Приходится во избежание неожиданностей создавать некий запас прочности в технике и экономике и пытаться строить эффективную «защиту от дурака». Мы говорим о наличии зависимости случайных величин, если их значения как-то систематически согласованы друг с другом в имеющихся у нас наблюдениях.

Любая статистика подтвердит, что рост человека явно связан с его весом (обычно высокие люди тяжелее низких), не вдаваясь в причины этого явления. С другой стороны, наличие связи между успеваемостью студентов по математике и числом лейкоцитов в их крови не столь очевидно, хотя и нельзя категорически ее отрицать.

Назначение статистики в том и состоит, чтобы помочь объективно оценить *степень зависимости* между случайными величинами.

Основной способ выявления этой оценки связан с понятием коэффициента корреляции по Пирсону, применяемого для исследования линейной взаимосвязи двух и более переменных, измеренных в метрических шкалах на одной и той же выборке, оценки пропорциональности изменчивости переменных.

Начнем с простейшего случая двух факторов. Пусть задана выборка объема N для двух случайных величин X и Y со значениями средних μ_x и μ_y.

Величина

$$\operatorname{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{4.1}$$

называется *ковариацией* X и Y.

Положительная ковариация свидетельствует об одинаковых тенденциях в поведении величин, отрицательная – об уменьшении значений одной при росте другой.

Нулевая ковариация вовсе не гарантирует независимости величин, а свидетельствует лишь об отсутствии линейной связи.

В частности, при $Y \equiv X$

$$\operatorname{cov}(x, x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 = \sigma_x^2. \tag{4.2}$$

Иногда эти оценки представляют симметрической *матрицей ковариаций*

$$\begin{bmatrix} \operatorname{cov}(x,x) & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & \operatorname{cov}(y,y) \end{bmatrix} \tag{4.3}$$

(аналогичная матрица строится в случае не только двух, но и бóльшего числа переменных).

Величина

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y},\ -1 \le r_{xy} \le 1 \tag{4.4}$$

называется коэффициентом *корреляции* и характеризует *степень линейной зависимости* между X и Y; близость $|r_{xy}|$ к 1 свидетельствует о наличии явно неслучайной такой связи.

4.1. Парная корреляция и линейная регрессия

Не вдаваясь в терминологические отличия понятий корреляции и регрессии, будем применять термин *регрессия* в сочетании с терминами *уравнение* или *линия*. Простая линейная *корреляция* (рис. 4.1) определяет степень, с которой значения двух переменных пропорциональны друг другу и на графике зависимость между ними можно представить прямой линией (с положительным или отрицательным углом наклона), которую называют *прямой регрессии* (термин *регрессия* введен Ф. Гальтоном в 1886 году).

Так, для двумерной выборки $\{(x_i, y_i),\ i = 1, 2, \ldots, N)\}$ коэффициент корреляции определяется в форме (4.4), где

$$\mathrm{cov}(x,y)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y),$$

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)^2,\sigma_y^2=\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_y)^2.$$

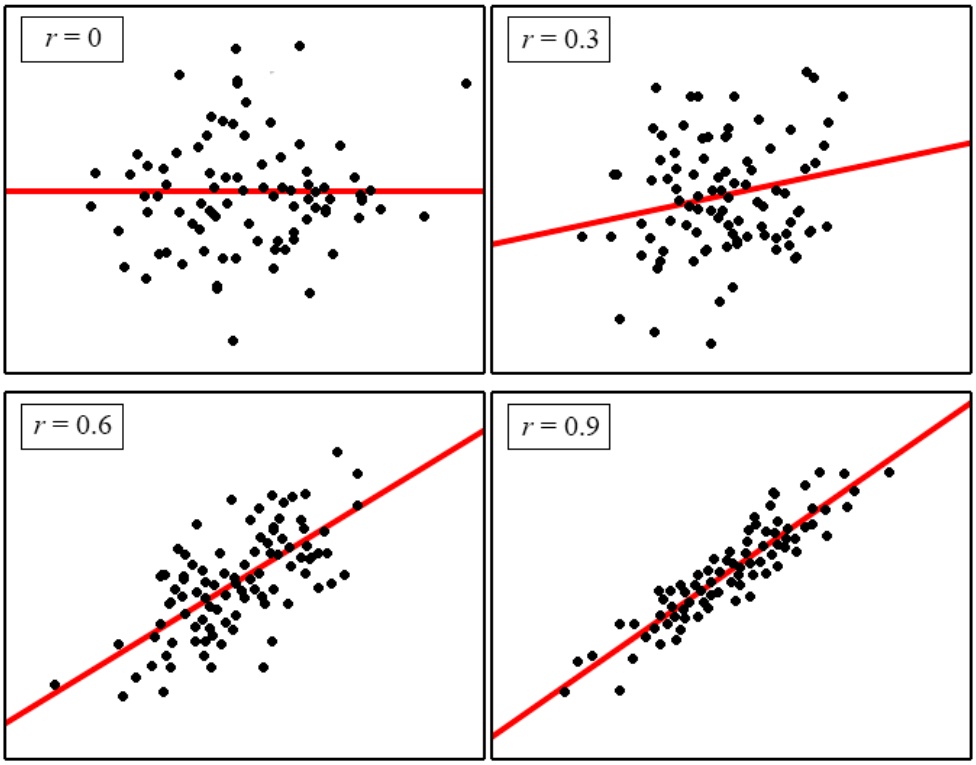


Рис. 4.1. Многообразие линейной корреляции

При традиционном обозначении для средней величины

$$\overline{xy}=\frac{1}{N}\sum_{i=1}^{N}x_iy_i\,, \qquad (4.5)$$

коэффициент корреляции и соответствующие дисперсии можно представить в более изящной (не всегда рациональной с точки зрения вычислений) форме:

$$r_{xy}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\sigma_x\sigma_y},\ \sigma_x^2=\overline{x^2}-\bar{x}^2,\ \sigma_y^2=\overline{y^2}-\bar{y}^2. \qquad (4.6)$$

Квадрат коэффициента корреляции называют *коэффициентом детерминации*, интерпретируя его как долю вариации, общую для двух переменных («степень» связанности двух переменных).

Линия регрессии подбирается по общеизвестному *методу наименьших квадратов* (К. Ф. Гаусс, 1801 г.), требующему, чтобы сумма *квадратов* отклонений по оси *Y* (расстояний) от наблюдаемых точек до линии регрессии была минимальной.

Так, если выбрать линию регрессии в виде $Y = a + b\,x$, то, потребовав минимума

$$Q(a,b) = \frac{1}{N}\sum_{i=1}^{N}(a + bx_i - y_i)^2 \ , \tag{4.7}$$

получаем систему двух линейных уравнений относительно двух неизвестных величин a и b:

$$\frac{\partial Q(a,b)}{\partial a} = 2\,\frac{1}{N}\sum_{i=1}^{N}(a + b\,x_i - y_i) = 0$$

$$\frac{\partial Q(a,b)}{\partial b} = 2\,\frac{1}{N}\sum_{i=1}^{N}(a + b\,x_i - y_i)\,x_i = 0$$

(уравнения получаются за счет обращения в нуль частных производных по a и b), откуда и находим

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2} = r_{xy}\,\frac{\sigma_y}{\sigma_x}\,,\ a = \overline{y} - b\,\overline{x}\,, \tag{4.8}$$

то есть уравнение прямой, проходящей через точку $(\overline{x}, \overline{y})$:

$$y = \overline{y} + r_{xy}\,\frac{\sigma_y}{\sigma_x}\left(x - \overline{x}\right). \tag{4.9}$$

Иногда коэффициент b называют коэффициентом регрессии, однако лучше таким понятием не пользоваться.

Поиск коэффициента корреляции r_{xy} не вызывает затруднений, но какова его надежность? Как и при проверке статистических гипотез, при проведении регрессионного анализа надо не только получить саму оценку коэффициента, но и оценить ее значимость. Естественно, что значимость коэффициента корреляции зависит от объема выборки N.

В предположении, что распределение отклонений наблюдений от прямой регрессии для y $\varepsilon_i = y_i - a - b\,x_i$ является нормальным, существует критерий проверки гипотезы об отсутствии связи с помощью статистики

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}, \tag{4.10}$$

распределенной по закону Стьюдента с $N - 2$ степенями свободы. Если эмпирическая оценка не превышает двустороннюю критическую (для уровня $\alpha / 2$), то нет оснований отвергать гипотезу об отсутствии взаимосвязи. Качество аппроксимации (значимость регрессии) характеризуется отношением *среднего квадрата регрессии* к среднему квадрату отклонений

$$F_{N-2,1} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{y})^2}{\frac{1}{N}\sum_{i=1}^{n}(Y_i - y_i)^2}, \qquad (4.11)$$

– статистикой, подчиняющейся распределению Фишера (если отношение велико, то нулевая гипотеза отвергается). В случае парной линейной регрессии *F*-критерий эквивалентен приведенной выше оценке по Стьюденту.

Для построения доверительного интервала коэффициентов уравнения регрессии требуется найти несмещенную дисперсию оценки (иногда ее называют *остаточной дисперсией*):

$$D_{\text{ост}} = \frac{1}{N-2}\sum_{i=1}^{N}\left(Y_i - y_i\right)^2 . \qquad (4.12)$$

Тогда доверительный интервал для коэффициента *b*

$$\Delta b = \pm t\sqrt{\frac{D_{\text{ост}}\sum_{i=1}^{N} x_i^2}{N^2\sigma_x^2}}, \qquad (4.13)$$

где t – критерий Стьюдента для уровня значимости $1 - \alpha / 2$ с $N - 2$ степенями свободы, а доверительный интервал для *a*

$$\Delta a = \pm t\sqrt{\frac{D_{\text{ост}}}{N^2\sigma_x^2}} . \qquad (4.14)$$

Основываясь лишь на коэффициентах корреляции, не всегда можно *гарантировать* причинно-следственную зависимость между переменными.

В пьесе Э. Ростана «Шантеклер» петух подметил, что всякий раз, когда он запоет, восходит солнце, и пришел к выводу, что именно он и вызывает солнце на небосклон. Очевидна корреляция между ущербом от пожара и числом пожарных, тушивших пожар. Из этого не следует, что для уменьшения ущерба следует уменьшать число вызванных пожарных. Иногда обнаруживаются *ложные корреляции* (рис. 4.2), то есть корреляции, обусловленные влияниями каких-то других, неизвестных нам факторов (переменных).

Неоднородность в выборке также является существенным фактором, смещающим (в ту или иную сторону) выборочную корреляцию. Так, если данные о растительности собраны в пустыне Калахари и близ Подкаменной Тунгуски, то искать корреляцию по совокупности призна-

ков следует весьма осторожно (распределения по некоторым составляющим могут оказаться отнюдь не унимодальными и далекими от нормального). Если такое явление возможно, то лучше вести анализ корреляции отдельно для каждого подмножества (для выделения подмножеств можно попытаться использовать, например, *кластерный анализ*).

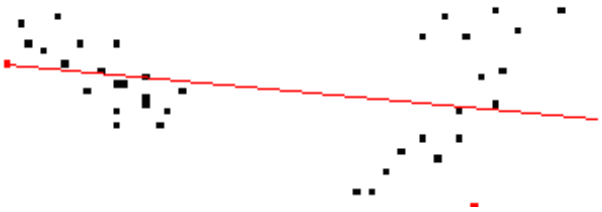


Рис. 4.2. Ложная корреляция

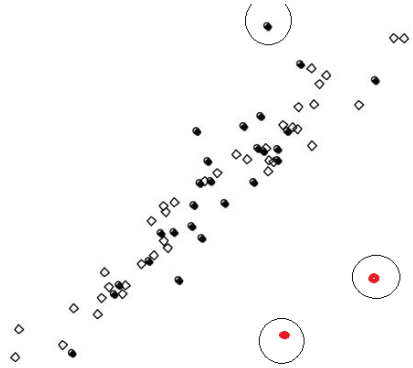


Рис. 4.3. Эллипс рассеивания и выбросы

Серьезную опасность для оценки корреляции и регрессии представляют *выбросы* (рис. 4.3) – нетипичные, резко выделяющиеся наблюдения. Даже единичный выброс способен существенно изменить наклон прямой и, следовательно, значение корреляции.

Общепринятого метода автоматического удаления выбросов не существует. В двумерном случае может помочь *диаграмма рассеяния*. В общем случае применяют численные методы удаления выбросов. Например, исключаются значения, которые выходят за границы $\pm 2\,\sigma$ от выборочного среднего. Однако во многих случаях выбросы представляют бо′льший интерес, чем сама выборка.

4.2. Множественная линейная регрессия

Заметим, что выше мы строили практически все суждения в предположении нормальности распределения одномерных случайных величин.

Для двумерных независимых случайных величин (рис. 4.4) с одинаковой дисперсией нетрудно найти плотность распределения

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}.$$

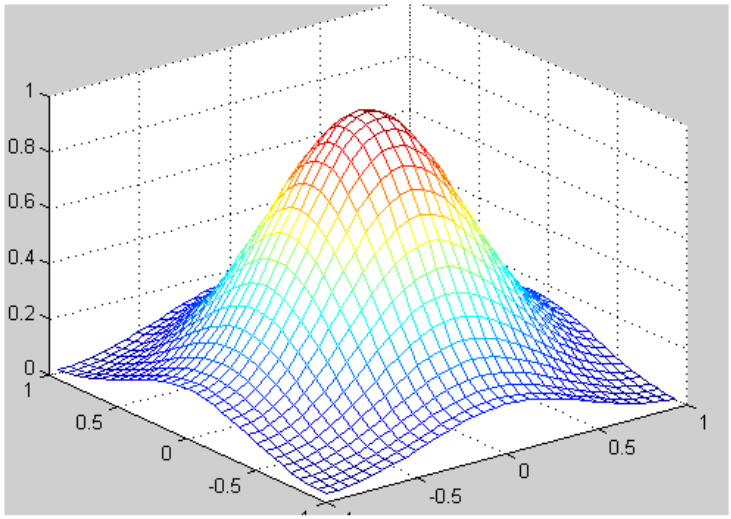


Рис. 4.4. Двумерное нормальное распределение независимых случайных величин

Если величины не являются независимыми ($|\rho_{xy}| > 0$), то

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} e^{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho_{xy}\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y}\right]},$$

которую следует иметь в виду (при существенной нелинейной связи).

Общее назначение множественной регрессии (Пирсон, 1908) состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и одной зависимой переменной. Например, агент по продаже недвижимости может внести в отдельные позиции своего каталога не только размер квартиры, число комнат, размер кухни, этаж, но и удаление от центра города, и наличие остановок муниципального транспорта, средний доход населения в этом районе, субъективную оценку продажной стоимости. Было бы интересно посмотреть, связаны ли эти характеристики жилья (если да, то каким образом) с ценой, по которой оно было продано.

А как связана заработная плата отечественного специалиста по компьютерным технологиям с его возрастом, трудовым стажем, удаленностью от Москвы, количеством вузов города, знанием иностранного языка, баллами ЕГЭ по математике?

Множественная регрессия позволяет оценить, какие факторы являются лучшими предикторами успешной учебы в средней школе или процветания фирмы.

Обратимся к ситуации выяснения взаимосвязи между переменной y и m факторами по выборке объема N наблюдений

$$\{(y_i, X_i)), i = \overline{1,n}\}, \qquad X_i = (x_{1i}, ..., x_{mi}) \tag{4.15}$$

(иногда, по техническим соображениям, мы будем переменную y включать в вектор предикторов X с индексом 0).

Ставится задача – подобрать линейную функцию, так называемое *уравнение множественной регрессии*

$$Y(X) = a_0 + \sum_{j=1}^{m} a_j x_j \ , \tag{4.16}$$

аппроксимирующее выборочные данные в смысле минимизации:

$$\frac{1}{N}\sum_{i=1}^{N}(Y(X_i) - y_i)^2 \tag{4.17}$$

(выше мы рассматривали идею метода наименьших квадратов и пример ее использования для построения линии регрессии).

Поиск неизвестных коэффициентов сводится к решению системы из $m + 1$ линейных уравнений:

$$\begin{cases} a_0 + \sum_{j=1}^{m} a_j \overline{x_j} = \overline{y} \\ a_0 \overline{x_k} + \sum_{j=1}^{m} a_j \overline{x_j x_k} = \overline{y}, \ k = \overline{1, m} \end{cases} \tag{4.18}$$

Перепишем (4.16) в *стандартизованном масштабе*

$$\frac{y - \overline{y}}{\sigma_y} = \sum_{j=1}^{m} \beta_j \frac{x_j - \overline{x_j}}{\sigma_{x_j}}, \tag{4.19}$$

откуда следует

$$a_0 = \overline{y} - \sum_{j=1}^{m} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \overline{x_j}, \quad a_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}}, \ j = \overline{1, m}. \tag{4.20}$$

Коэффициенты уравнения (4.19) позволяют оценить относительную их значимость в модели (вклад разброса данного фактора в общий разброс результирующего признака).

Вспомнив, что имеет место

$$r_{kj} = \frac{\overline{x_k x_j} - \overline{x_k}\,\overline{x_j}}{\sigma_k \sigma_j}, \tag{4.21}$$

и считая $y \equiv x_0$, систему (4.19) можно переписать в виде

$$\sum_{j=1}^{m} r_{jk} \beta_j = r_{0k}, \ k = \overline{1, m}. \tag{4.22}$$

Симметрическая положительно определенная матрица коэффициентов парной корреляции

$$D = \begin{bmatrix} 1 & r_{01} & r_{02} & \dots & r_{0m} \\ r_{10} & 1 & r_{12} & \dots & r_{1m} \\ r_{20} & r_{21} & 1 & \dots & r_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{m0} & r_{m1} & r_{m2} & \dots & 1 \end{bmatrix}, \qquad (4.23)$$

дает материал для последующих действий по оптимизации модели множественной регрессии.

Так, обнаружив наличие тесной связи между некоторыми двумя признаками, один из них можно удалить.

Величина $R = \sqrt{1 - \frac{|D|}{D_{00}}}$, где D_{00} – определитель матрицы, полученной из D вычеркиванием нулевой строки и нулевого столбца, называется *коэффициентом множественной регрессии*.

По существу, R представляет собой долю дисперсии Y, объясненную линейной зависимостью с $X_1, X_2, ..., X_m$

$$R = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}}, \qquad (4.24)$$

где несмещенная оценка остаточной дисперсии с учетом числа степеней свободы

$$\sigma^2_{\text{ост}} = \frac{1}{n-m-1} \sum_{i=1}^{n} (Y(X_i) - y_i)^2 . \qquad (4.25)$$

Так при $R = 0{,}8$ остается 60 % необъясненного разброса значений Y.

По аналогии с парной регрессией предлагается критерий проверки гипотезы об отсутствии связи с помощью статистики

$$t = \frac{R\sqrt{n-m-1}}{\sqrt{1-R^2}}, \qquad (4.26)$$

отвечающей распределению Стьюдента с $n - m - 1$ степенями свободы. Если эмпирическая оценка не превышает двустороннюю критическую (для уровня значимости $\alpha / 2$), то нет оснований отвергать гипотезу об отсутствии взаимосвязи.

Доверительный интервал для коэффициентов a_k определяется

$$\Delta a_k = \pm t \sqrt{\frac{\sigma^2_{\text{ост}}}{n\,\sigma^2_{x_k}}}\,, \qquad (4.27)$$

где t – критерий Стьюдента для уровня $1 - \alpha / 2$ с $f = n - m - 1$.

Значимость коэффициентов множественной регрессии можно оценивать и с помощью критерия Фишера.

Наряду с множественным коэффициентом корреляции существует понятие *частного коэффициента корреляции* как коэффициента парной корреляции между двумя признаками y, x_k при фиксированных значениях множества остальных признаков $Z = \{X \setminus x_k\}$.

В случае $m = 3$ эти значения можно получить из соотношений

$$r_{01\ 2} = \frac{r_{01} - r_{02}\,r_{12}}{\sqrt{(1 - r^2_{02})\,(1 - r^2_{12})}}\,,\ r_{02\ 1} = \frac{r_{02} - r_{01}\,r_{12}}{\sqrt{(1 - r^2_{01})\,(1 - r^2_{12})}}\,. \qquad (4.28)$$

В общем случае частные коэффициенты рассчитываются непосредственно при фиксации значений остальных признаков на уровне средних значений.

Что делать, если часть данных о значениях некоторых факторов в таблице наблюдений пропущена? Можно удалить целиком соответствующую строку таблицы (наблюдение), но это чревато потерей информации – понижается число степеней свободы и соответственно надежность выводов. Чаще идут по пути замены пропущенных данных средним значением. Этот путь сохраняет состоятельность оценок, но искусственно уменьшает разброс данных и тем самым уменьшает корреляцию.

4.3. Нелинейная регрессия

Корреляция Пирсона хорошо подходит для описания линейной зависимости. Однако реальная регрессия может быть *нелинейной*, и соответствующая линия регрессии выбирается либо по диаграмме рассеяния, либо из опыта предшественников, либо по соображениям «здравого смысла». Нелепо искать линейную связь между явно или опосредованно взаимно аналитически вычисляемыми факторами $y = F(x)$ и $x = F^{-1}(y)$ (например, затратами на единицу продукции и объемом продукции на единицу затрат).

Забавно, когда для наукообразия по трем наблюдениям строят кривую второго порядка (параболу), получают нулевую остаточную дисперсию и коэффициент корреляции 0,9999995. Уже Евклид постулировал, что через 2 точки на плоскости однозначно можно провести единственную прямую. Для N точек, отражающих некую случайность, выбор

аппроксимирующего многочлена порядка $N-1$ устраняет наличие случайности (число степеней свободы должно быть больше нуля!).

Нелепо подменять статистический анализ традиционной для прикладной математики аппроксимацией табличной функции, например алгебраическими многочленами высших порядков. Дать им осмысленное профессиональное объяснение нереально.

«Джентльменские наборы» популярных аппроксимирующих нелинейных функций (рис. 4.5) в пакетах прикладных программ включают гиперболическую функцию $y = a + b / x$, квадратичную $y = a + bx + cx^2$, линейно-логарифмическую $y = a + b \ln(x)$, показательную $y = \exp(a + bx)$, алгебраические и тригонометрические многочлены и другие, линейные относительно своих коэффициентов, поиск которых методом наименьших квадратов сводится к решению систем линейных уравнений.

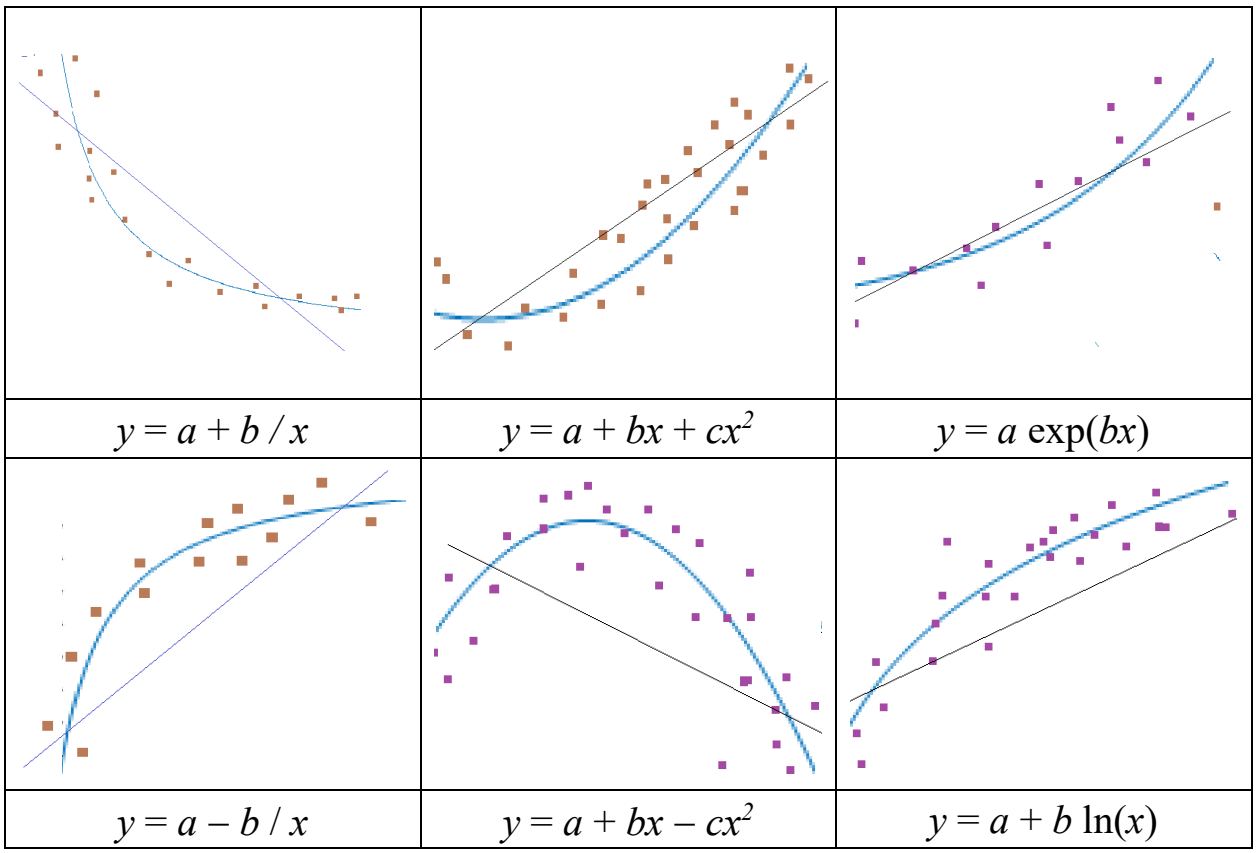


Рис. 4.5. Популярные нелинейные связи

Для линеаризации этих нелинейных функций (приведения к линейному виду) в представлении для линейной регрессии (4.8) достаточно значения x заменить обратными значениями, их логарифмами и т. п. При этом следует помнить, линеаризация нелинейных связей на базе усреднения данных привносит значимую погрешность (см. рис. 4.5). Построение аппроксимирующей функции для всего диапазона данных иногда разумнее заменить квадратичным сплайном [25].

Есть функции нелинейные относительно своих коэффициентов, не поддающиеся достаточно простой аппроксимации методом наименьших квадратов. Так мультипликативная модель

$$y = a_0 \prod_{j=1}^{m} x_j^{a_j} \tag{4.29}$$

(обобщение известной модели Кобба – Дугласа) логарифмированием приводится к системе $m+1$ линейных алгебраических уравнений. А при *существенно нелинейной регрессии* типа *логистического уравнения*

$$y = \frac{a_0}{1 + a_1 e^{a_2 x}} \tag{4.30}$$

приходится решать систему нелинейных уравнений каким-либо из итерационных методов (Ньютона, наискорейшего спуска, покоординатного спуска и пр.).

Можно использовать искусственные приемы численного анализа. Например, привести модель (4.30) к виду

$$a_0 \tfrac{1}{y} = 1 + a_1 e^{a_2 x} \tag{4.31}$$

и, фиксируя a_2 в разумном диапазоне, решить линейную систему

$$\begin{cases} a_0 \left(\overline{\frac{1}{y^2}} \right) - a_1 \left(\overline{\frac{1}{y} e^{a_2 x}} \right) = \left(\overline{\frac{1}{y}} \right) \\ a_0 \left(\frac{1}{y} e^{a_2 x} \right) - a_1 \overline{e^{2 a_2 x}} = \overline{e^{a_2 x}} \end{cases}, \tag{4.32}$$

и оценить соответствующее значение остаточной дисперсии для той или иной формы представления уравнения регрессии (4.30) или (4.31). Получив оценки для выбранного a_2, выявляем направление последующих смещений в сторону минимума остаточной дисперсии и продолжаем начатый итерационный процесс достижения требуемой точности.

Иногда при нелинейной регрессии вместо термина *коэффициент корреляции* используют термин *регрессионное отношение* и вместо символа r – символ ρ.

4.4. Метод главных компонент и факторный анализ

Метод главных компонент как метод своеобразного понижения размерности за счет поиска попарно ортогональных направлений максимальной вариации исходных данных появился в 1901 году в работе Пирсона, но реальное его использование связано с появлением компьютеров и совершенствованием методов решения проблемы собственных значений матрицы.

Пусть имеется выборка из n наблюдений над m признаками X, характеризующаяся вектором средних значений $\mu = [\mu_1, \mu_2, ..., \mu_m]$ и матрицей ковариаций (или парных коэффициентов корреляции)

$$\mathrm{cov}(x_k, x_j) \equiv \sigma_{kj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \mu_{x_k})(x_{ij} - \mu_{x_j}), \quad k, j = \overline{1, m} \qquad (4.33)$$

(напоминаем, что при $k = j$ ковариация совпадает с дисперсией k-го признака (фактора) и что матрица ковариаций, деленная на стандартные отклонения соответствующих признаков, дает матрицу парных коэффициентов корреляции).

Наша цель – это поиск линейных комбинаций (обобщенных факторов)

$$Y_s = \sum_{k=1}^{m}\alpha_{sk}X_k, \quad s = \overline{1, q} \le m, \qquad (4.34)$$

таких, что:

1) $\mathrm{cov}(Y_k, Y_j) = 0$ при всех $k \neq j$;

2) $D_{Y1} \ge D_{Y2} \ge ... \ge D_{Yq}$;

3) суммарная дисперсия всех Y_s совпадает с суммарной дисперсией всех исходных признаков.

Первое из условий утверждает отсутствие корреляции между всеми Y_s (геометрически это можно интерпретировать как взаимную ортогональность (перпендикулярность) векторов. Второе предлагает лишь размещение факторов по уровню значимости.

Для понимания сущности и цели метода обратимся к случаю двух факторов. На рис. 4.6 в привычной для нас ортогональной системе координат X_1 и X_2 представлены результаты наблюдений – точки с соответствующими координатами (наличие корреляции очевидно). Попробуем найти такой перенос и поворот осей координат с сохранением их ортогональности, чтобы в этих координатах корреляция была подавлена.

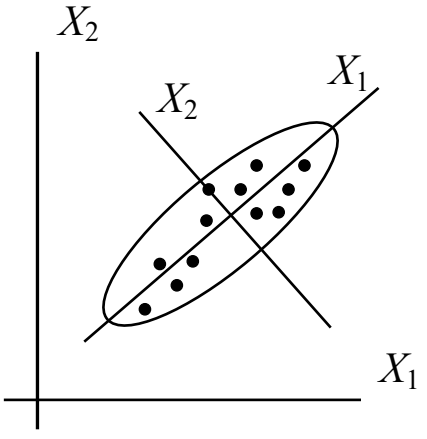


Рис. 4.6. Случай двух факторов

Без труда можно переместить центр эллипса рассеивания в начало координат X_1 и X_2 и подвергнуть обычной нормировке $(z - \mu) / \sigma$. Из отношения μ_2 / μ_1 находим угол θ поворота координатных осей и переходим к новой прямоугольной системе координат Y_1, Y_2:

$$y_1 = x_1 \cos(\theta) - x_2 \sin(\theta),$$
$$y_2 = x_1 \sin(\theta) + x_2 \cos(\theta).$$

Исходные точки оказываются сконцентрированными в эллипсе, центр которого определяется средними значениями исходных признаков, а полуоси определяются дисперсиями так называемых главных компонент. Как правило, дисперсия факторов в новой системе координат оказывается меньше дисперсии исходных факторов. Если обнаружится для дисперсий новых факторов (компонент) $D(Y_2) << D(Y_1)$, то Y_2 можно отбросить (его вклад в общую дисперсию незначителен).

Обратным преобразованием

$$x_1 = y_1 \cos(\theta) + y_2 \sin(\theta),$$
$$x_2 = -y_1 \cos(\theta) + y_2 \sin(\theta)$$

можем вернуться к исходным факторам, по крайней мере, удалив случайные выбросы.

В общем случае математический аппарат метода главных компонент сводится к поиску собственных чисел λ_s ($s = 1, 2, \ldots, m$) и нормированных собственных векторов $a_s = (a_{s1}, a_{s2}, ..., a_{sm})$ ковариационной матрицы.

Если понимание методов решения систем линейных алгебраических уравнений не требует особой математической подготовки за пределами неполной средней школы и программирование простейших из них доступно начинающему программисту (во всех библиотеках программ систем программирования есть таковые), то методы решения проблемы собственных значений представляют вершину методов линейной алгебры.

Не вдаваясь в детали, можно считать, что проблема состоит в поиске ненулевых решений так называемого характеристического уравнения $A\,X = \lambda\,X$ (здесь A – ковариационная симметрическая матрица m-го порядка, λ – вектор собственных значений того же порядка, X – матрица собственных векторов). Благодаря симметричности матрицы A численное решение несколько упрощается, но остается нетривиальным. Для достижения цели исследователь должен ознакомиться с соответствующими методами или обратиться к библиотекам программ.

Так в библиотеке MatLab имеется команда `[X,v]=eig(A)` (`v` – диагональная матрица собственных чисел, `X` – матрица нормированных собственных векторов).

Выполнив ее, получаем

```
  1 2 3      0.8484  0.7163 -0.1198    0.2179  0 0
A=1 4 9    X=-0.5150 0.6563 -0.3295  v=0 1.8393  0
  1 8 27     0.1222 -0.2371 -0.9365    0 0      29.9428
```

Значения собственных чисел определяют дисперсию соответствующих главных компонент и отношение некоторого собственного числа к общей их сумме или к сумме дисперсий исходных признаков, то есть определяют ту долю дисперсии, которая может быть объяснена предлагаемой комбинацией исходных признаков.

Например, получив собственные значения (дисперсию, объясненную последовательными факторами):
6,11837 1,80068 0,47289 0,40800 0,31722 0,29330 0,19581 0,17043
0,13797 0,08533
(сумма равна 10), можно найти соответствующий вклад в общую дисперсию (%)
61,1837 18,0068 4,7289 4,0800 3,1722 2,9330 1,9581 1,7043 1,3797 0,8533.

Сколько факторов следует оставить? Можно ориентироваться на сохранение определенного суммарного процента или на широко используемый *критерий Кайзера*, предлагающий брать только факторы с собственными значениями, бо′льшими единицы.

Иногда главные компоненты имеют четко выраженный физический смысл (если в предложенном наборе данных о больных выделяется компонента, характеризующая состав крови, и ей отвечает собственное число, составляющее 75 % суммы всех собственных чисел, то возрадуйтесь – вы нашли эффективный способ диагностики).

Анализируя влияние десятков факторов на уровень итогового показателя благополучия жителей Кузбасса, выделяем их группировки и значимость по признакам производственных возможностей, экологии, географии и т. д. Разумеется, ряд факторов присутствует в нескольких группах с соответствующим весом.

Очевидна возможность обратного преобразования (от главных компонент к исходным)

$$X_s = \sum_{k=1}^{m} \beta_{sk} Y_k, \; s = \overline{1, m}, \tag{4.35}$$

и к тому же можно доказать равенство $\beta_{sk} = \alpha_{ks}$, откуда получаем так называемую *модель главных компонент*

$$X_s = \sum_{k=1}^{m} \alpha_{ks} Y_k, \; s = \overline{1, m}. \tag{4.36}$$

В сущности, метод главных компонент дает обоснованные решения для так называемого *факторного анализа*, главными целями которого также являются *сокращение* числа переменных (редукция данных) и *определение структуры* взаимосвязей между переменными, то есть *классификация переменных*.

Если в анализе главных компонент предполагается, что должна быть использована *вся* изменчивость переменных, то в факторном анализе используется только изменчивость переменной, общая и для других переменных. В большинстве случаев эти два метода приводят к весьма близким результатам.

Отыскав собственные числа и собственные векторы матрицы ковариаций, берем в качестве *общих факторов* первые $q \le m$ главных компонент Y_i с весами $F_i = Y_i / V(\lambda_i)$, где

$$V(Y_i) = \sum_{k=1}^{m}\sum_{s=1}^{m} \alpha_{ik}\alpha_{is}\sigma_{ks} . \qquad (4.37)$$

Величины $l_{ij} = \alpha_{ji} V(Y_i)^{1/2}$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, q$ называются *факторными нагрузками* и характеризуют коэффициент корреляции между i-м исходным признаком и j-м фактором.

Величины $h_i^2 = \sum_{j=1}^{q} l_{ji}^2$ определяют так называемые оценки *общности*.

Соотношение

$$X_i = \sum_{j=1}^{q} l_{ij}F_j + e_i, i = \overline{1, m}, \qquad (4.38)$$

где e_i – оценки вклада специфических факторов, называется *факторной моделью*.

Общие факторы здесь взаимно некоррелированные и имеют единичную дисперсию. Получив представление исходных факторов через главные компоненты (взятые частично или полностью), получаем возможность выяснить структуру главных компонент – попытаться дать интерпретацию этих факторов.

4.5. Пробит- и логит-анализ

Нередко зависимая переменная (переменная отклика) бинарна по своей природе, то есть может принимать только два значения. Например, пациент может выздороветь или нет, кандидат на должность может пройти или провалить тест при приеме на работу и т. п. Во всех этих случаях нас может интересовать зависимость между «непрерывными» переменными и одной бинарной, зависящей от них. Едва ли в этом случае можно использовать стандартную множественную регрессию. Если сопоставить зависимой переменной значения 0 и 1, то нет гарантии, что мы не получим модель с предсказываемыми значениями, большими единицы и меньшими нуля. Задача регрессии может быть сформулирована иначе: вместо предсказания бинарной переменной мы пред-

сказываем *непрерывную* переменную со значениями на отрезке [0, 1]. Наибольшее распространение в этой области получили регрессионные модели *логит* и *пробит*.

В модели *логит-регрессии* предсказываемые значения для зависимой переменной принадлежат отрезку [0, 1] при любых значениях независимых переменных. Это достигается применением регрессионного уравнения

$$y = \frac{\exp(b_0 + \sum_{j=1}^{n} b_j x_j)}{1+\exp(b_0 + \sum_{j=1}^{n} b_j x_j)} \quad , \tag{4.39}$$

которое не зависит от коэффициентов регрессии и значений x и дает $y \in [0, 1]$ (может интерпретироваться как вероятность p).

В принципе так называемым логистическим преобразованием $q = \ln[p / (1 - p)]$ можно получить обычную модель линейной регрессии для произвольного по величине значения

$$q = b_0 + \sum_{j=1}^{m} b_j x_j \, . \tag{4.40}$$

Может использоваться и *обобщенная логит-регрессия*

$$y = \frac{b_0}{1 + b_1 e^{b_2 x}}, \tag{4.41}$$

позволяющая отклику y произвольно меняться внутри фиксированного интервала.

Пробит-регрессия рассматривает бинарную зависимую переменную как отклик на изменения некоторой основной, нормально распределенной переменной с диапазоном принимаемых значений от минус до плюс бесконечности (вся числовая прямая). Например, человек по отношению к какому-то мероприятию может быть «против», сомневаться или решительно «за». В любом случае будет бинарный отклик на мероприятие (участие или отказ).

4.6. Ранговая корреляция

Как было указано ранее, некоторые случайные величины представляются в *порядковой шкале*. Каждому значению (объекту) такой величины присваивается *ранг* (порядковый номер) $1, 2, ..., n$, где n – количество объектов. Меры взаимосвязи между парами величин (признаков) называются в статистике *коэффициентами ранговой корре-*

ляции. Наиболее популярны ранговые корреляции по Спирмену и Кендаллу.

Коэффициент ранговой корреляции, названный по имени психолога Ч. Спирмена (1904 г.), определяет степень взаимосвязи для двух случайных величин (признаков), значения которых ранжированы от 1 до n. Здесь рассчитываются сумма квадратов ранговых разностей

$$D = \sum_{i=1}^{n} (R_{xi} - R_{yi})^2 \qquad (4.42)$$

и сам коэффициент ранговой корреляции

$$r_S = 1 - \frac{6D}{n(n^2 - 1)}, \qquad (4.43)$$

изменяющийся в диапазоне от –1 до 1.

В самом деле, если ранги значений для обоих признаков одинаковы, то $D = 0$ и $r_s = 1$. Если они противоположны, то при четных n

$D = [1 - n]^2 + [2 - (n - 1)]^2 + ... + [n - 1]^2 = n\,(n^2 - 1) / 3$ и $r_s = -1$.

В ситуации независимых признаков ожидаемое значение $r_s = 0$ и дисперсия $Dr_s = 1 / (n - 1)$. По отклонению r_s от нуля можно судить о степени зависимости или независимости признаков.

При малых n можно прибегнуть к таблице предельных значений распределения коэффициента r_s при различных уровнях значимости

$\alpha \setminus n$	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**25**	**30**
0,05	0,90	0,83	0,71	0,64	0,60	0,56	0,44	0,38	0,34	0,31
0,025	1,00	0,89	0,79	0,74	0,68	0,65	0,52	0,44	0,40	0,36

Если учесть, что при $n > 10$ величина $r_s\sqrt{n-1}$ асимптотически нормальна с параметрами $N(0, 1)$, то при больших выборках значимость r_s можно определять величиной $t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$ и сравнивать эту величину с предельным значением $t_{1-\alpha/2}$ для критерия Стьюдента с $f = n - 2$ степенями свободы.

Если есть уверенность в нормальном распределении признаков, то при больших выборках можно использовать в качестве оценки коэффициента ранговой регрессии величину $2\sin\left(\frac{\pi r_s}{6}\right)$.

Пусть 10 студентов протестированы по математике и истории с оценками по стобалльной шкале:

Математика	95 90 87 84 75 70 61 60 58 55
История	92 94 83 79 58 61 47 72 62 68

Запишем ранги студентов по тестам и квадраты их разностей:

Математика	1	2	3	4	5	6	7	8	9	10
История	2	1	3	4	9	8	10	5	7	6
D_i	1	1	0	0	16	4	9	9	4	16

Отыскав $D = \sum d_i^2 = 1+1+16+4+9+9+4+16 = 60$, вычислим выборочный коэффициент ранговой корреляции Спирмена между оценками по двум тестам $r_s = 1 - \frac{6D}{n\left(n^2-1\right)} = 1 - \frac{6 \cdot 60}{10 \cdot 99} = \frac{7}{11}$. Гипотеза о положительной корреляции приемлема на 95%-м уровне (редкий случай гармонии в образовании). Тем не менее $r_s^2 = 0{,}4$ объясняет наличие связи лишь на 40 %.

Еще один пример столетней давности о зависимости между средним размером имения (десятин) и сословным рангом владельца (конец XIX века) на основе приведенной статистики землевладения:

Сословие (*X*)		**Размер имения (*Y*)**	
название	**ранг**	**количество десятин**	**ранг**
Дворяне	1	400	1
Купцы	2	70	3
Мещане	3	150	2
Крестьяне	4	40	4

Здесь $D = (1-1)^2 + (2-3)^2 + (3-2)^2 + (4-4)^2 = 2$,
$r_s = 1 - 6 \cdot 2 / (4 \cdot 15) = 0{,}8$. Получилась довольно существенная связь между сословиями и землевладением (неудивительно, что крестьянский вопрос в европейской части царской России не был решен ни Пугачевым, ни Столыпиным).

При ранжировании признаков нередко возникает ситуация, когда два (или большее число) значений получают одинаковые ранги (такие значения называют *связанными*). В этом случае ранг связанных объектов берется равным среднему значению тех рангов, которые имели бы эти значения, если они были бы различны. Если число связанных рангов невелико, то при вычислении ранговой корреляции можно пользоваться введенными здесь формулами, в противном случае эти формулы несколько усложняются.

Так при большом числе связанных рангов отыскивают значения

$$t_x = \sum_i \frac{t_{xi}^3 - t_{x_i}}{12};\ t_y = \sum_i \frac{t_{yi}^3 - t_{y_i}}{12}, \qquad (4.44)$$

где t_{xi}, t_{yi} – количества связанных рангов, если хотя бы одно из этих значений отлично от нуля, то берется

$$A = \frac{n^3 - n}{12} - t_x; B = \frac{n^3 - n}{12} - t_y;\ r_s = \frac{A + B + D}{2\sqrt{AB}}. \qquad (4.45)$$

Критерий Кендалла (*тау*-критерий) определяется в форме

$$\tau = \frac{S^+ - S^-}{\frac{1}{2}n(n-1)} = \frac{S}{\frac{1}{2}n(n-1)}, \qquad (4.46)$$

где S^+ и S^- определяются следующим образом. В приведенном примере таблица упорядочена по возрастанию рангов X (это существенно для вычисления S) и дополнена двумя столбцами, заполняемыми согласно соотношения рангов X и Y.

Сословие (*X*)		**Размер имения (*Y*)**		S^+	S^-
название	**ранг**	**количество десятин**	**ранг**		
Дворяне	1	400	1	3	0
Купцы	2	70	3	1	1
Мещане	3	150	2	1	0
Крестьяне	4	40	4	0	0
			Сумма	**5**	**1**

Для значений в колонке S^+ находим количество последующих с бóльшим рангом: ранг 1 первого значения меньше рангов всех последующих трех значений – в первую строку столбца заносим число 3; ранг второго значения, равный 3, меньше только одного из последующих – во вторую строку этого столбца заносим число 1 и т. д.

При заполнении столбца S^- выясняем число последующих строк с меньшими рангами.

Отыскав суммы по столбцам, получаем оценку по Кендаллу $\tau = \frac{5-1}{\frac{1}{2}\cdot 4(4-1)} = 0{,}67$, более осторожную в сравнении с оценкой по Спирмену.

Утверждают, что

$$-1 \le 3\,\tau - 2\,r_s \le 1. \qquad (4.47)$$

Для проверки значимости коэффициента Кендалла (при $n > 10$) по Стьюденту вычисляют значение

$$t_k = \frac{S}{\sqrt{n(n-1)\,(2n+5)/18}}\ . \qquad (4.48)$$

Коэффициент Кендалла часто используется для оценки согласованности мнений независимых экспертов (судей), в частности, в отношении баллов, выставленных одному и тому же субъекту.

Если исследуется взаимосвязь 3 признаков, то с помощью коэффициентов τ можно найти *коэффициент частной ранговой корреляции*, позволяющий оценить степень «чистой» взаимосвязи двух ранговых признаков, устранив влияние третьего:

$$\tau_{12\,3} = \frac{\tau_{12} - \tau_{13}\,\tau_{23}}{\sqrt{(1-\tau_{13}^2)\,(1-\tau_{23}^2)}}\,. \qquad (4.49)$$

Множественный коэффициент ранговой корреляции W (называемый также *коэффициентом конкордации*) предназначен для измерения связи произвольного числа ранговых признаков. Для примера возьмем нижеприведенную таблицу $n = 7$ наблюдений над $m = 3$ ранговыми признаками, упорядоченную по рангам первого признака (Δ – отклонение суммы рангов от среднего значения). Если ранжировки по разным признакам совпадают (или близки), то суммарные ранги будут сильно различаться. Если же все m ранжировок слабо согласованы, суммарные ранги объектов будут почти одинаковыми и близкими к их средней сумме, равной в нашем случае $S = m\,(n + 1) / 2 = 12$.

№	**Ранги**			**Сумма рангов**	**Δ**	**Δ²**
1	1	5	7	13	1	1
2	2	3	6	11	–1	1
3	3	4	5	12	0	0
4	4	6	4	14	2	4
5	5	1	3	9	–3	9
6	6	7	2	15	3	9
7	7	2	1	10	–2	4
Средняя сумма				84 / 7 = 12	**Сумма**	28

Отыскав S^* – сумму квадратов отклонений найденных сумм от средней (здесь 28), нормируют ее делением на максимально возможное значение, равное $m^2\,(n^3 - n) / 12$.

Таким образом, формула для вычисления коэффициента конкордации имеет следующий вид:

$$W = \frac{12\,S^*}{m^2\,(n^3 - n)}\,. \qquad (4.50)$$

Значения W заключены в интервале [0, 1]. Равенство $W = 0$ означает полную несогласованность m ранжировок. Если же $W = 1$, то все

m ранжировок совпадают. Значимость полученной величины W может быть проверена по критерию χ^2:

$$\chi^2 = \frac{12S^*}{mn(n+1)} \qquad (4.51)$$

с числом степеней свободы $k = n - 1$.

Для нашего примера $W = 0{,}11$, $\chi^2_{ф} = 2$, $k = 6$. Для уровня значимости $\alpha = 0{,}05$ критическое значение $\chi^2_{кр} = 12{,}6$. Поскольку фактическое значение $\chi^2_{ф}$ меньше критического, то гипотеза об отсутствии связи между рассматриваемыми ранговыми признаками не отклоняется, то есть коэффициент W в данном случае не является значимым.

Коэффициенты ранговой корреляции могут использоваться не только для качественного анализа взаимосвязи двух ранговых признаков, но и при определении степени связи между показателями различной природы. В этом случае значения количественного признака упорядочиваются, и им приписываются соответствующие ранги. Разумеется, возникает определенное «огрубление» исходной информации.

4.7. Корреляционно-регрессионный анализ в среде MS Excel

Рассмотрим пример построения линейного уравнения регрессии между полной себестоимостью добычи 1 т угля Y (млн руб.) и средней суточной добычей угля на шахте X_1 (т), удельным весом комбайновой проходки выработки X_2 (%) (табл. 4.1).

Таблица 4.1. Исходные данные

№ п/п	Y	X_1	X_2	№ п/п	Y	X_1	X_2
1	3	29	17	**11**	2,5	28	15
2	4	40	25	**12**	3,5	30	18
3	3	36	15	**13**	2	25	16
4	3,2	32	17	**14**	4	48	23
5	2	23	15	**15**	2,2	30	18
6	3,5	45	18	**16**	3,2	40	18
7	3,5	38	17	**17**	3,9	40	25
8	4	40	25	**18**	3,6	38	23
9	3,8	50	19	**19**	2,6	29	18
10	4	47	23	**20**	2,5	25	17

Решение. Введем исходные данные в диапазон `A1:C21` рабочего листа MS Excel (рис. 4.7).

Исследуем корреляцию факторов, выполнив команду `Данные – Анализ данных – Корреляция`. Заполним параметры диалогового окна (рис. 4.8).

В результате получим корреляционную матрицу (табл. 4.2).

	A	B	C
1	Y	X1	X2
2	3	29	17
3	4	40	25
4	3	36	15
5	3,2	32	17
6	2	23	15
7	3,5	45	18
8	3,5	38	17
9	4	40	25
10	3,8	50	19
20	2,6	29	18
21	2,5	25	17

Рис. 4.7. Исходные данные

Таблица 4.2. Корреляционная матрица

	Y	X_1	X_2
Y	1		
X_1	0,853056	1	
X_2	0,778766	0,615448	1

Проанализируем матрицу парных коэффициентов межфакторной корреляции (табл. 4.3).

Факторы X_1 и X_2 не коллинеарны, так как $|r_{X1\,X2}| \le 0{,}7$. Следовательно, оба фактора оставляем для построения уравнения регрессии.

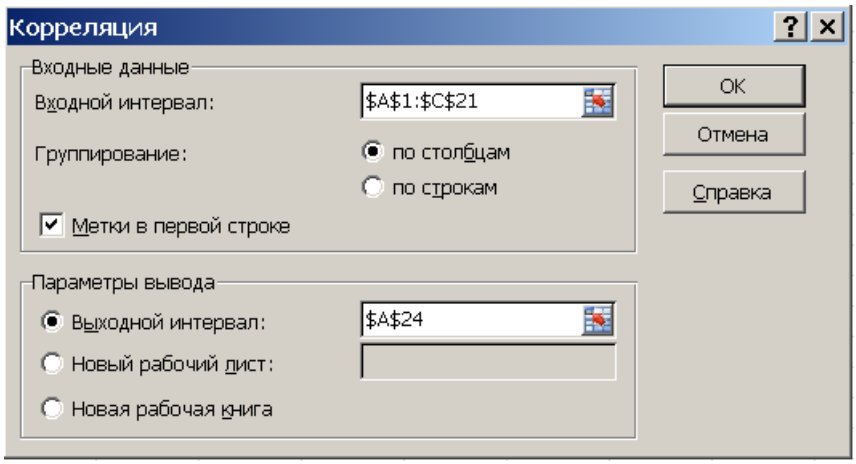


Рис. 4.8. Диалоговое окно «Корреляция»

Заполним диалоговое окно инструмента `Пакет анализа Регрессия` (рис. 4.9), выполнив команду `Данные – Анализ данных – Регрессия`. В результате выбора кнопки `OK` получим таблицу итогов регрессионного анализа (табл. 4.4).

Дисперсионный анализ показывает, что уравнение является значимым при уровне значимости $\alpha = 2{,}68686 \cdot 10^{7}$. Множественный коэффициент корреляции $R = 0{,}912$, то есть полученное уравнение достаточно хорошо описывает изучаемую взаимосвязь между факторами. Коэффициент детерминации $R^2 = 0{,}83$ – это означает, что 83 % вариации результативного признака (Y) объясняется вариацией факторных переменных (X_1, X_2).

Таблица 4.3. Матрица межфакторной корреляции

	X_1	X_2
X_1	1	
X_2	0,615448	1

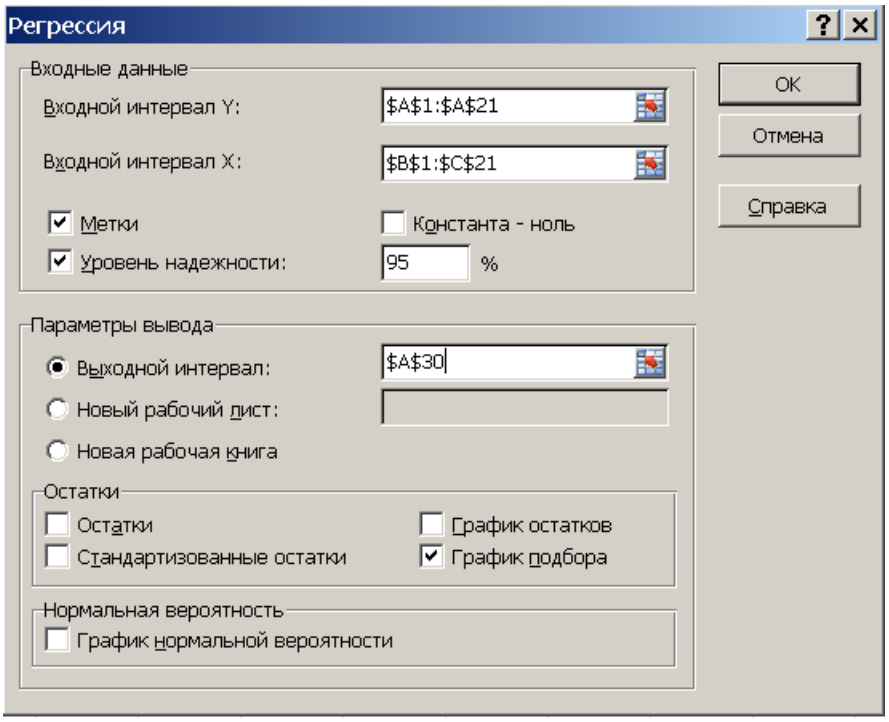


Рис. 4.9. Диалоговое окно «Регрессия»

Таблица 4.4. Вывод итогов

Регрессионная статистика

Множественный R	0,91178744
R-квадрат	0,83135634
Нормированный R-квадрат	0,81151591
Стандартная ошибка	0,30111587
Наблюдения	20

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	2	7,598597	3,799298	41,90213	2,68686E-07
Остаток	17	1,541403	0,090671		
Итого	19	9,14			

	Коэффи-циенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95,0%
Y-пересечение	-0,16039	0,39214	-0,40902	0,68764	-0,98773	0,66695
X1	0,05099	0,01071	4,76118	0,00018	0,02839	0,07359
X2	0,08076	0,02499	3,23243	0,00489	0,02805	0,13348

Согласно табл. 4.4 искомое уравнение регрессии имеет вид

$$Y = 0{,}051X_1 + 0{,}081X_2 - 0{,}16.$$

Причем доверительный интервал при уровне значимости 5 %:

1) для коэффициента при X_1 – (0,028; 0,074);
2) для коэффициента при X_2 – (0,028; 0,133);

3) для свободного члена – (–0,988; 0,667).

MS Excel позволяет также анализировать парные регрессионные зависимости (остатки и многое другое). Так в данном примере зависимость Y от X_1 имеет вид, представленный инструментом `Регрессия` на рис. 4.10.

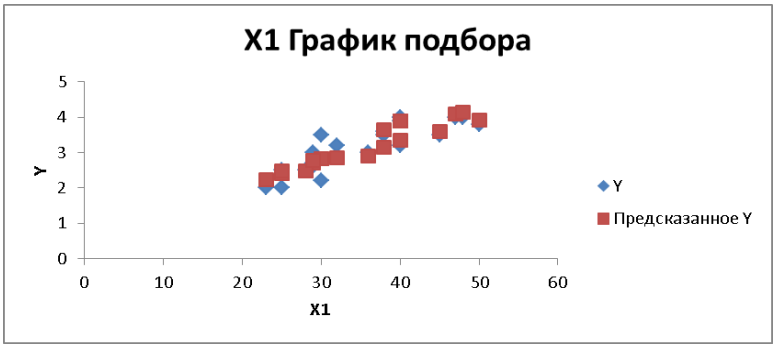


Рис. 4.10. График подбора

Проанализируем полученное уравнение регрессии. Изучалась зависимость между результативным признаком – «полная себестоимость добычи 1 т угля» (Y) и факторными признаками «средняя суточная добыча угля на шахте» (X_1), «удельный вес комбайновой проходки выработки» (X_2). Из параметров найденного уравнения видно, что повышение средней суточной добычи угля на шахте на 1 т приводит к увеличению полной себестоимости добычи 1 т угля в среднем на 0,051 млн руб. Увеличение удельного веса комбайновой проходки выработки на 1 % приводит к увеличению полной себестоимости добычи 1 т угля в среднем на 0,081 млн руб.

Контрольные вопросы

1) В чем смысл корреляционной связи?
2) Что характеризует коэффициент парной корреляции?
3) Что собой представляет линейная парная регрессия?
4) Каким образом связаны линейная регрессия и коэффициент парной корреляции?
5) В чем суть метода наименьших квадратов?
6) Как коэффициент детерминации связан с коэффициентом корреляции?
7) Какой критерий используется для проверки гипотезы об отсутствии связи?
8) Каким образом можно оценить качество аппроксимации (значимость регрессии)?
9) Как строится доверительный интервал для коэффициентов уравнения регрессии?

10) Какое влияние на коэффициент корреляции и регрессионное уравнение оказывают неоднородность выборки и выскакивающие варианты?

11) В чем заключается основное назначение множественной линейной регрессии?

12) Каким образом вычисляется и что собой представляет коэффициент множественной регрессии?

13) Что собой представляют и как рассчитываются частные коэффициенты корреляции?

14) Чем отличаются линейная и нелинейная регрессии?

15) Как выглядит логистическое уравнение?

16) С какой целью используется метод главных компонент?

17) В каких ситуациях используется коэффициент ранговой корреляции?

18) Для чего используется и каким образом вычисляется коэффициент ранговой корреляции Кендалла?

19) Для чего используется и каким образом вычисляется коэффициент ранговой корреляции Спирмена?

20) Каким образом вычисляется и что позволяет оценить коэффициент частной ранговой корреляции?

21) Для чего предназначен множественный коэффициент ранговой корреляции?

22) В чем заключается особенность модели логит-регрессии?

23) Каким образом в модели пробит-регрессии рассматривается бинарная зависимая переменная?

24) Какими средствами корреляционно-регрессионного анализа располагает MS Excel?

Глава 5. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ – один из методов статистического анализа, который, как и критерий Стьюдента, предназначен для поиска зависимостей в экспериментальных данных путем исследования значимости различий в средних значениях за счет сравнения выборочных дисперсий. В отличие от *t*-критерия он допускает сравнение средних значений трех и более групп. В зарубежной литературе дисперсионный анализ часто обозначается как ANOVA – анализ вариативности (Analysis of Variance).

Идеология дисперсионного анализа предложена Фишером в 1920 году. Как утверждают литературные источники, «дисперсионный анализ был разработан для обработки данных, полученных в ходе специальных экспериментов, и считался единственным методом корректного исследования причинных связей». Метод применялся для оценки экспериментов в растениеводстве, а позднее в психологии, педагогике, медицине, экономике и др.

В задачах, решаемых этим методом, присутствует численно представленный *отклик*, на который воздействует несколько переменных номинальной природы (в шкале наименований).

Примерами могут служить эксперименты по выбору рационов откорма скота, эффективных сортов картофеля, размещению травматологических пунктов на автомагистралях и аптечных киосков в городе, рекламы в СМИ и т. д. Эти эксперименты могут закончиться обнаружением ответа на вопрос о существовании статистически значимого отличия между итогами поставленных экспериментов.

Существует много моделей одно- и многофакторного дисперсионного анализа и сопутствующих критериев [3].

5.1. Однофакторный дисперсионный анализ

Остановимся на рассмотрении простейшего случая – *однофакторного дисперсионного анализа*.

Пусть в итоге эксперимента получены *m* групп наблюдений с разными значениями средних $\mu = (\mu_1, \mu_2, ..., \mu_m)$, но одинаковыми дисперсиями

$$\sigma_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{ik} - \mu)^2 \, , \, k = 1, 2, \ldots, m \qquad (5.1)$$

(n_k – количество наблюдений в *k*-й группе, $\sum_{k=1}^{m} n_k = n$).

Ниже приведен пример с разными средними по каждой группе, но с одинаковой внутригрупповой дисперсией.

Если все наблюдения группировать воедино, обнаруживаем, что общая дисперсия много выше дисперсии отдельных групп.

Группа	Наблюдения	Среднее	Дисперсия	Общее среднее	Общая дисперсия
1	2 3 1	2	2/2	4	28/5
2	6 7 5	6	2/2		

Иными словами, внутригрупповая изменчивость меньше общей изменчивости (относительно общего среднего). Проверка значимости такого различия (существенна ли обнаруженная разница между средними или это случайность?) составляет предмет дисперсионного анализа.

Здесь ищут оценку отношения доли межгрупповой дисперсии к внутригрупповой с помощью *F*-критерия (действительно ли отношение дисперсий значимо больше 1?). Существуют две модели подхода к дисперсионному анализу (рис. 5.1) – с так называемыми *фиксированными* и *случайными эффектами.*

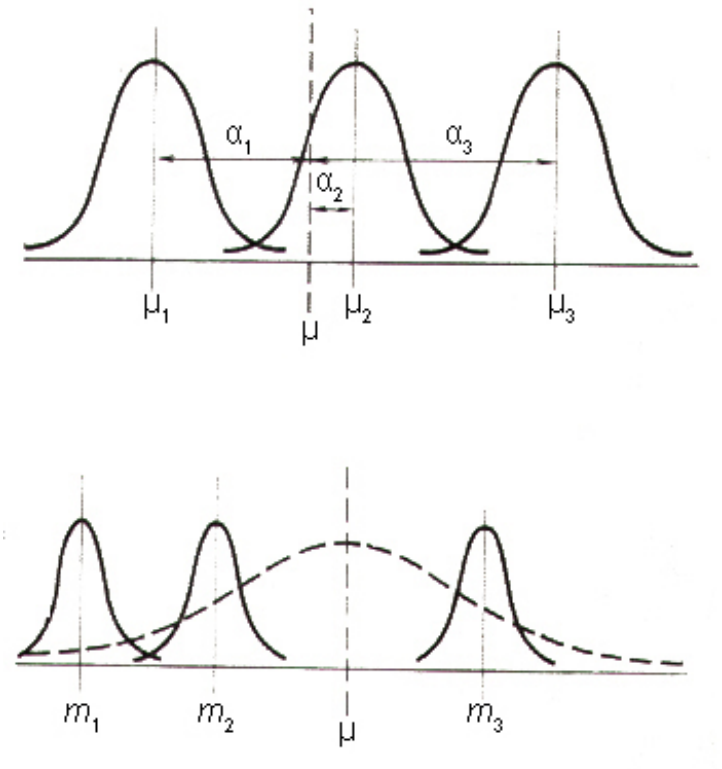


Рис. 5.1. Модели подхода к дисперсионному анализу (вверху – с фиксированными эффектами; внизу – со случайными)

В *модели с фиксированными эффектами* каждое наблюдаемое значение базируется на общем среднем μ и фиксированных эффектах α_i, определяемых условиями эксперимента (фиксированные виды диеты, партийная принадлежность губернаторов, погода и т. п.).

На верхнем из приведенных рисунков (рис. 5.1) [3] показаны 3 группы наблюдений с разными средними и одинаковой внутригрупповой дисперсией. Отклонения средних по группам от общего среднего объясняется условиями эксперимента (например, в трех регионах проведены выборы, и влияние на результат зависит от степени вмешательства административного ресурса и т. п.).

На нижнем рисунке (см. рис. 5.1) отклонения обусловлены случайностью. Процедура однофакторного дисперсионного анализа представлена в табл. 5.1.

Таблица 5.1. Процедура однофакторного дисперсионного анализа

Источник дисперсии	**Сумма квадратов**	**Степени свободы**	**Средний квадрат**	***F*-отношение**
Межгрупповая	$SS_B = \sum_{k=1}^{m} n_k (\overline{x_k} - \overline{x})^2$	$\nu_B = m - 1$	$MS_B = \frac{SS_B}{\nu_B}$	
Внутригрупповая	$SS_R = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \overline{x_k})^2$	$\nu_R = n - m$	$MS_R = \frac{SS_R}{\nu_R} = S^2$	$F = \frac{MS_B}{MS_R}$
Полная	$SS_T = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \overline{x})^2$	$\nu_T = n - 1$		

Обратившись к первой модели, проверяем гипотезу равенства нулю всех значений α_i (или, что то же, $\mu_1 = \mu_2 = ... = \mu_m = \mu$). Например, при испытании $m = 4$ видов «диет» выбраны группы по 9 испытуемых ($n = 36$) и получены средние эффекты по каждой группе: 4,10; 4,63; 4,75; 5,04. Насколько значимы эти различия?

Оценка общего среднего здесь равна сумме этих эффектов, деленной на 4, то есть 4,63. Отклонения α_i от среднего составили 4,10 – 4,63; 4,63 – 4,63; 4,75 – 4,63; 5,04 – 4,63. Сумма их квадратов, умноженная на 9, составила 4,23. Из исходной таблицы наблюдений (здесь не приведена) найдены внутригрупповая и полная дисперсия – 14,06 и 18,29. В итоге имеем:

Источник дисперсии	**Сумма квадратов отклонений**	**Степени свободы**	**Средний квадрат**	***F*-отношение**
Межгрупповая	$SS_B = 4{,}23$	$\nu_B = 3$	1,41	
Внутригрупповая	$SS_R = 14{,}06$	$\nu_R = 32$	0,44	3,21
Полная	$SS_T = 18{,}29$	$\nu_T = 35$		

Если сравнить найденное значение F с процентилями распределения Фишера для уровней значимости 5 % и 10 % $F(3; 32; 0{,}05) = 8{,}61$ или $F(3; 32; 0{,}10) = 5{,}17$, то обнаруживается сомнение относительно нулевой гипотезы и приходится сравнивать эффект диет попарно на основе *t*-критерия.

Во второй *модели* нас интересует оценка дисперсии внешних эффектов – проверка гипотезы $\sigma_a^2 = 0$ (фактор не вносит никакого вклада в дисперсию).

Так в приведенном выше примере за счет равенства объемов

$$\mu = 4{,}63;\ \sigma^2 = 0{,}44;\ \sigma_a^2 = (1{,}41 - 0{,}44) / 9 = 0{,}11;$$

$$F = (\sigma^2 + 9\sigma_a^2) / \sigma^2 = 3{,}21.$$

В случае *двухфакторного дисперсионного анализа* рассматриваются сочетания и группировки отношений между двумя группами факторов.

5.2. Двухфакторный дисперсионный анализ

В случае *двухфакторного дисперсионного анализа* рассматриваются два типа отношений между группами факторов *A* и *B* – *пересечение* и *группировка*.

Пересечение возникает, если в плане эксперимента представлены все сочетания групп (например, все сочетания m_1 диет и m_2 комплексов физических упражнений). Другими словами, имеется сетка сочетаний (ячеек) ij ($i = 1, 2, \ldots, m_1$; $j = 1, 2, \ldots, m_2$), для каждого из которых выполнено n_{ij} наблюдений (обычно предполагают постоянство числа наблюдений). Такая модель называется *двухфакторным планом* (*двухфакторной классификацией*).

Говорят, что *фактор B группируется фактором A*, если каждая группа фактора *B* встречается в паре с единственной группой группирующего фактора *A*. Общее число комбинаций здесь меньше $m_1 \times m_2$. Такая модель $B(A)$ называется *двухфакторной иерархической моделью*.

Обратимся к модели пересечений при постоянстве $n_{ij} = K$.

Если в однофакторной модели наблюдаемые значения базировались на общем среднем и фиксированных (или случайных) эффектах для группы, то здесь $X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i\,j} + \varepsilon_{ijk}$, то есть наряду с эффектами факторов присутствует эффект взаимодействия факторов.

Обозначим $\overline{X}_i$ – средние по строкам, через $\overline{X}_j$ – по столбцам и $\overline{X}$ – по всем элементам и построим таблицу. На базе такой таблицы (табл. 5.2) и осуществляется проверка гипотез:

1) на отсутствие эффектов взаимодействия

$$F = \frac{MS_{AB}}{MS_R},\ \nu_1 = (m_1 - 1)\,(m_2 - 1),\ \nu_2 = m_1\,m_2\,(K - 1); \qquad (5.2)$$

2) на отсутствие эффектов фактора *A*

$$F = \frac{MS_A}{MS_R},\ \nu_1 = m_1 - 1,\ \nu_2 = m_1\,m_2\,(K - 1); \qquad (5.3)$$

3) на отсутствие эффектов фактора *B*

$$F = \frac{MS_B}{MS_R},\ \nu_1 = m_2 - 1,\ \nu_2 = m_1\,m_2\,(K - 1). \qquad (5.4)$$

Таблица 5.2. Процедура двухфакторного дисперсионного анализа

Источник дисперсии	Сумма квадратов	Степени свободы	Средний квадрат
Фактор *A*	$SS_A = m_B K \sum_{i=1}^{m_1} (\overline{x_{i..}} - \overline{x}_{...})^2$	$\nu_A = m_1 - 1$	$MS_A = \frac{SS_A}{\nu_A}$
Фактор *B*	$SS_B = m_A K \sum_{j=1}^{m_2} (\overline{x_{.j.}} - \overline{x}_{...})^2$	$\nu_B = m_2 - 1$	$MS_B = \frac{SS_B}{\nu_B}$
Взаимодействие *AB*	$SS_{AB} = K \sum_{j=1}^{m_1} \sum_{j=1}^{m_2} (\overline{x_{ij.}} - \overline{x}_{i..} - \overline{x}_{.j.} + \overline{x}_{...})^2$	$\nu_{AB} = (m_1 - 1)$ $(m_2 - 1)$	$MS_{AB} = \frac{SS_{AB}}{\nu_{AB}}$
Остаток (ошибка)	$SS_R = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{K} (x_{ijk} - \overline{x_{ij.}})^2$	$\nu_R = m_1 m_2 (K - 1)$	$MS_R = \frac{SS_R}{\nu_R}$
Полная	$SS_T = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{K} (x_{ijk} - \overline{x_{...}})^2$	$\nu_T = m_1 m_2 K - 1$	

Например, при испытании влияния на объем выдыхаемого азота 4 видов диет [7, с. 249] выбраны группы из 3 испытуемых мужчин и 3 женщин и получены оценки, приведенные в табл. 5.3.

Таблица 5.3. Результаты испытаний

	Диеты				**Средние**
	1	**2**	**3**	**4**	
Мужчины	4,079 4,859 3,540	4,368 5,668 3,752	4,169 5,709 4,416	4,928 5,609 4,940	4,6697
Женщины	2,870 4,648 3,847	3,578 5,393 4,374	4,403 4,496 4,688	4,905 5,208 4,806	4,4347
Средние	3,9738	4,5222	4,6468	5,0658	4,5522

На основании этих данных строим расчетную таблицу (табл. 5.4), из которой находим эмпирические оценки $F_{3,16} = 0{,}03$, $F_{3,16} = 2{,}48$, $F_{1,16} = 0{,}68$ и обнаруживаем при уровне значимости $\alpha = 0{,}05$ отсутствие значимых факторов (соответствующие оценки равны 8,70 и 246).

Не останавливаясь на других вариантах двухфакторного анализа, заметим, что «мир по своей природе сложен и многомерен». Ситуации, когда некоторое явление полностью описывается одной или двумя переменными, чрезвычайно редки.

Заметим, что многофакторный дисперсионный анализ, несмотря на многие свои недостатки, предпочтительнее последовательного срав-

нения двух выборок при разных уровнях факторов с помощью *t*-критерия.

Таблица 5.4. Результаты расчетов

Источник дисперсии	Сумма квадратов (*SS*)	Степени свободы	Средний квадрат (*MS*)
Диета	3,6491	3	1,2164
Пол	0,3314	1	0,3314
Диета × Пол	0,0428	3	0,0143
Остаток	7,8353	16	0,4897
Полная	11,8586	23	

5.3. Технология решения задач дисперсионного анализа с применением MS Excel

`Пакет анализа` позволяет решать в диалоговом режиме три задачи дисперсионного анализа, наиболее часто встречающихся в классической математической статистике (рис. 5.2).

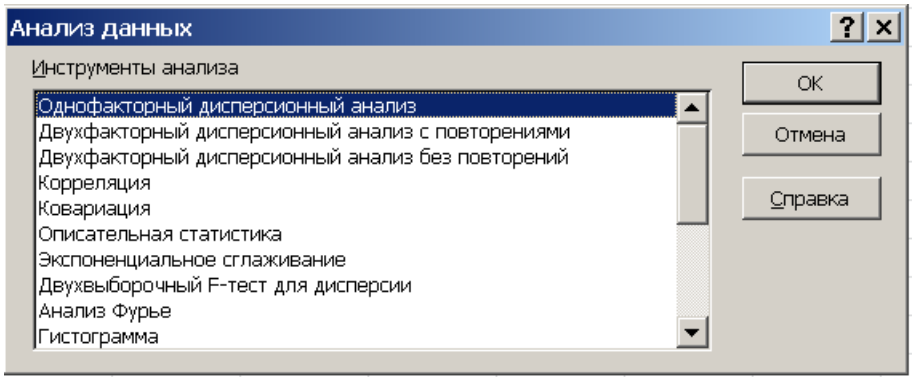


Рис. 5.2. Инструменты MS Excel для решения задач дисперсионного анализа

Однофакторный дисперсионный анализ позволяет статистически обосновать существенность влияния факторного признака *A* на результативный *F*.

Замечание. В `Пакете анализа` нет выделения моделей *дисперсионного анализа* по виду факторов.

Однофакторный дисперсионный анализ используется для проверки гипотезы о сходстве средних значений двух или более выборок, принадлежащих одной и той же генеральной совокупности. Этот метод распространяется также на тесты для двух средних (к которым относится, например, *t*-критерий). То есть если для разных уровней фактора *A* средние отличаются незначительно, следует принять гипотезу о существенном влиянии факторного признака *A* на результативный *F*.

Пример 1. Необходимо проверить статистическую существенность влияния катализатора *A* на химическую реакцию. Результаты измерений при 5 уровнях фактора *A* приведены в таблице.

A1	***A2***	***A3***	***A4***	***A5***
3,2	2,6	2,9	3,7	3
3,1	3,1	2,6	3,4	3,4
3,1	2,7	3	3,2	3,2
2,8	2,9	3,1	3,3	3,5
3,3	2,7	3	3,5	2,9
3	2,8	2,8	3,3	3,1

Решение. Введем исходные данные в диапазон `A1:A7` листа MS Excel. Заполним параметры диалогового окна (рис. 5.3).

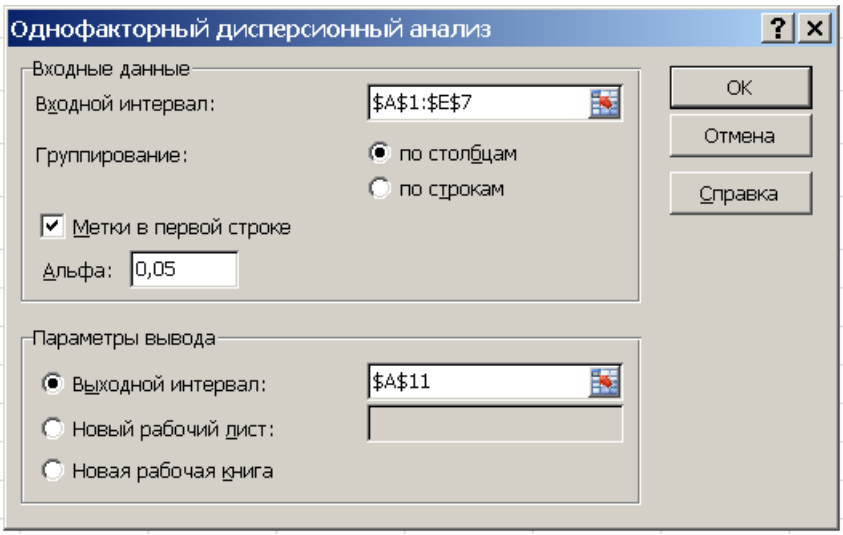


Рис. 5.3. Диалоговое окно однофакторного дисперсионного анализа

`Группирование: по столбцам` (рис. 5.3) – именно такое расположение имеют уровни фактора *A*. Отметим `Метки в первой строке` (там расположены уровни фактора *A*). В `Выходном интервале` достаточно отметить левую верхнюю ячейку выходного интервала `$A$11`. После нажатия кнопки `OK` получим таблицу однофакторного дисперсионного анализа (табл. 5.5).

Двухфакторный дисперсионный анализ с повторениями представляет собой более сложный вариант однофакторного анализа, включающего более чем одну выборку для каждой группы данных. Двухфакторный дисперсионный анализ позволяет статистически обосновать существенность влияния факторных признаков *A* и *B* и взаимодействия факторов (*A* и *B*) на результативный фактор *F*.

Таблица 5.5. Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
A1	6	18,5	3,08333333	0,0296667
A2	6	16,8	2,8	0,032
A3	6	17,4	2,9	0,032
A4	6	20,4	3,4	0,032
A5	6	19,1	3,18333333	0,0536667

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	1,342	4	0,3355	9,3540892	9,164E-05	2,75871047
Внутри групп	0,896667	25	0,03586667			
Итого	2,238667	29				

Пример 2. У 60 сотрудников предприятия фиксировалась среднечасовая выработка в натуральных единицах продукции. Данные обследования представлены в таблице MS Excel (табл. 5.6).

Таблица 5.6. Исходные данные

	A	B	C	D
1	Стаж	Возраст		
2		от 25 до 35 лет	от 35 до 45 лет	от 45 до 55 лет
3	от 1 до 4 лет	19	19	18
4		20	20	19
5		20	20	20
6		20	23	21
7		22	25	23
8	от 4 до 7 лет	30	20	19
9		31	29	25
10		32	30	25
11		32	31	26
12		34	31	26
13	от 7 до 10 лет	35	36	24
14		35	40	24
15		39	41	24
16		40	42	25
17		41	45	25
18	свыше 10 лет	40	28	20
19		40	31	24
20		41	35	25
21		41	36	31
22		42	40	32

Необходимо оценить существенность влияния возраста и стажа на производительность труда.

Решение. На основе таблицы MS Excel (табл. 5.6) заполним диалоговое окно (рис. 5.4).

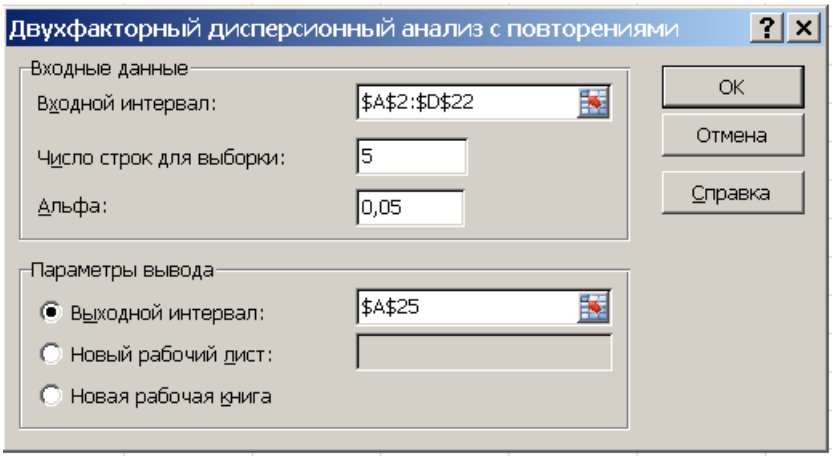


Рис. 5.4. Диалоговое окно двухфакторного дисперсионного анализа с повторениями

Следует отметить, что в исходных данных должно быть одинаковое число строк (повторений наблюдений), в противном случае рекомендуется ввести оценки пропущенных наблюдений, выбрав их таким образом, чтобы минимизировать остаточную дисперсию (иначе нужно ввести среднее значение других наблюдений в ячейках), причем нельзя включать эти оценки при подсчете соответствующих степеней свободы. В случае пропуска данных в дисперсионном анализе без повторений необходимы более сложные методы импутирования.

После нажатия кнопки `OK` получим итоговую табл. 5.7. Дисперсионный анализ с повторениями позволяет оценить существенность влияния факторов *A* (стажа), *B* (возраста) и их взаимодействия (факторов *A* и *B*) на среднечасовую выработку продукции в натуральных единицах. Так как:

$F_{A\text{расч}} = 66{,}8189 > F_{A\text{кр}} = 2{,}7980;$

$F_{B\text{расч}} = 48{,}9791 > F_{B\text{кр}} = 3{,}1907;$

$F_{AB\text{расч}} = 9{,}7456 > F_{AB\text{кр}} = 2{,}2945,$

то следует признать статистически значимым влияние стажа (фактор *A*), возраста (фактор *B*), и их взаимодействия (факторы *A* и *B*) на производительность труда сотрудников. Итоговая таблица позволяет более детально рассмотреть свойства отдельных групп (например, возраст от 35 до 45 лет и стаж от 4 до 7 лет).

В MS Excel имеется хорошая возможность наглядного представления изучаемых процессов или явлений с помощью мастера диаграмм. Для этого предварительно выделим данные (см. табл. 5.6), затем выберем вкладку `Вставка`, потом выберем тип диаграммы `Диаграммы – График`. В результате после преобразований получим рис. 5.6.

Таблица 5.7. Двухфакторный дисперсионный анализ с повторениями

ИТОГИ	от 25 до 35 лет	от 35 до 45 лет	от 45 до 55 лет	Итого
от 1 до 4 лет				
Счет	5	5	5	15
Сумма	101	107	101	309
Среднее	20,2	21,4	20,2	20,6
Дисперсия	1,2	6,3	3,7	3,542857
от 4 до 7 лет				
Счет	5	5	5	15
Сумма	159	141	121	421
Среднее	31,8	28,2	24,2	28,06667
Дисперсия	2,2	21,7	8,7	19,6381
от 7 до 10 лет				
Счет	5	5	5	15
Сумма	190	204	122	516
Среднее	38	40,8	24,4	34,4
Дисперсия	8	10,7	0,3	60,4
свыше 10 лет				
Счет	5	5	5	15
Сумма	204	170	132	506
Среднее	40,8	34	26,4	33,73333
Дисперсия	0,7	21,5	25,3	50,6381
Итого				
Счет	20	20	20	
Сумма	654	622	476	
Среднее	32,7	31,1	23,8	
Дисперсия	68,53684211	66,62105263	13,32631579	

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	1842,533333	3	614,1777778	66,81898	3,7E-17	2,798061
Столбцы	900,4	2	450,2	48,97915	2,56E-12	3,190727
Взаимодействие	537,4666667	6	89,57777778	9,745542	5,2E-07	2,294601
Внутри	441,2	48	9,191666667			
Итого	3721,6	59				

Пример 3. Рассмотрим решение задачи из примера 2, взяв в каждой ячейке среднее значение наблюдений (табл. 5.8). В результате заполнения диалогового окна (аналогично окну примера 2) получим итоговую таблицу двухфакторного дисперсионного анализа без повторений (табл. 5.9).

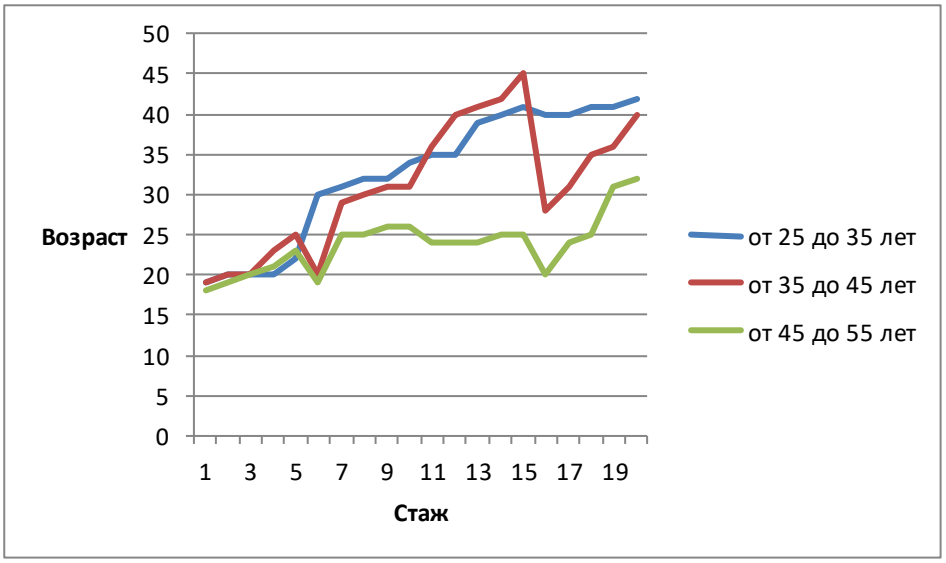


Рис. 5.6. Изменение среднечасовой выработки по категориям

Таблица 5.8. Исходные данные (средние значения наблюдений)

Стаж	Возраст		
	от 25 до 35 лет	от 35 до 45 лет	от 45 до 55 лет
от 1 до 4 лет	20,2	21,4	20,2
от 4 до 7 лет	31,8	28,2	24,2
от 7 до 10 лет	38	40,8	24,4
свыше 10 лет	40,8	34	26,4

Таблица 5.9. Двухфакторный дисперсионный анализ без повторений

ИТОГИ	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
от 1 до 4 лет	3	61,8	20,6	0,48
от 4 до 7 лет	3	84,2	28,06666667	14,45333
от 7 до 10 лет	3	103,2	34,4	76,96
свыше 10 лет	3	101,2	33,73333333	51,89333
от 25 до 35 лет	4	130,8	32,7	83,58667
от 35 до 45 лет	4	124,4	31,1	68,33333
от 45 до 55 лет	4	95,2	23,8	6,746667

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Строки	368,5	3	122,8355556	6,856363	0,02295241	4,757062663
Столбцы	180,1	2	90,04	5,0258	0,05222744	5,14325285
Погрешность	107,5	6	17,91555556			
Итого	656,1	11				

Данные дисперсионного анализа (см. табл. 5.9) свидетельствуют о том, что фактор A (стаж) существенно влияет на производительность труда (так как $F_{A\text{расч}} = 6{,}8563 > F_{A\text{кр}} = 4{,}7570$), а фактор B (возраст) статистически существенного влияния не оказывает (так как $F_{B\text{расч}} = 5{,}0258 < F_{B\text{кр}} = 5{,}1432$). Различия в выводах примеров 2 и 3 можно объяснить тем, что в примере 3 не учитывались конкретные наблюдения в ячейках, а рассматривались только их средние значения. Поэтому результат дисперсионного анализа с повторениями является более значимым, что соответствует и самому смыслу задачи. Так как очевидно – на производительность труда влияют и стаж, и возраст, и их взаимодействие (стаж и возраст), следует вывод, что при организации наблюдений необходимо для каждого уровня факторов рассматривать возможно большее количество элементов в ячейках.

Контрольные вопросы

1) Что проверяется с помощью дисперсионного анализа?
2) В чем заключается основа дисперсионного анализа?
3) Как вычисляется внутригрупповая дисперсия?
4) Относительно какого среднего вычисляется межгрупповая дисперсия?
5) Почему внутригрупповая дисперсия не может быть больше межгрупповой?
6) С помощью какого критерия проверяется значимость различий дисперсий?
7) Опишите модель дисперсионного анализа с фиксированными эффектами.
8) В чем отличие модели дисперсионного анализа со случайными эффектами от модели с фиксированными эффектами?
9) Какое взаимодействие факторов рассматривается в двухфакторном дисперсионном анализе?
10) Какая модель называется двухфакторным планом?
11) Какие суммы квадратов вычисляются при двухфакторном дисперсионном анализе?
12) Назовите три проверяемые гипотезы при двухфакторном дисперсионном анализе.
13) Как проверяется отсутствие эффекта взаимодействия факторов?
14) Как проверяется отсутствие эффекта фактора A?
15) Как проверяется отсутствие эффекта фактора B?

Глава 6. КЛАСТЕРНЫЙ АНАЛИЗ

Задачи *кластерного анализа* (*распознавания образов*) возникают в различных сферах деятельности и связаны с проблемой *таксономии* – поиска групп объектов, объединяемых на основании общих признаков. Образ, кластер, таксон – сочетание терминов, между которыми, с неформальной точки зрения, трудно усмотреть четкие различия. Существование человека и всего живого на Земле так или иначе неотделимо от постоянного решения проблемы распознавания образов. По каким-то признакам кошка, не знакомая с курсом сравнительной биологии, отличает корову и собаку, ребенок – доброго и злого человека. Каждый человек вырабатывает свои представления о красоте, о музыке и звукоизвержении, уме и глупости. Врач, пользуясь созданными его предшественниками многотомными классификаторами признаков заболеваний, накопленным опытом и здравым смыслом, авторитетно ставит диагноз.

Согласно принятому определению, *распознавание образов* – это отнесение исходных данных к определенному классу с помощью выделения существенных признаков, характеризующих эти данные, из общей массы несущественных данных. Все возрастающая значимость этого направления в современной науке и технике связана с развитием вычислительной техники, необходимостью общения человека и компьютера, появлением роботизированных систем. Ведущие советские журналы по кибернетике 50-х годов излагали проблемы и успехи машинного перевода технических текстов, но за прошедшие годы компьютеры так и не научились переводить сонеты Шекспира. Однако появились превосходные поисковые системы Интернета, где на запрос «распознавание образов» Вы получите десятки тысяч справочных ссылок (за вами остается лишь «навозну кучу разбирая, найти жемчужное зерно» – а вы уверены в своих способностях и познаниях в разведке полезных ископаемых?).

По результатам геологоразведки прогнозируются месторождения нефти в западносибирском регионе и алмазов в Якутии, библиотекари создают систематизированные каталоги литературы, службы безопасности распознают «своих – чужих» по глазам и тембру голоса, произносящего «Сезам, откройся», зенитчики распознают быстродвижущиеся объекты, компьютерщики учат ЭВМ распознаванию рукописного текста и т. д. Методы, применяемые в вычислительной технике для выполнения поставленных задач, многообразны. Распознавание может производиться по форме, цвету, положению, шаблону и т. п. Но так или иначе, оно связано с кластеризацией.

Кластер (англ. cluster – скопление) определяется как объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определенными свойствами.

Таксономию определяют как «учение о принципах и практике классификации и систематизации сложноорганизованных иерархически соотносящихся сущностей». Первая реальная постановка задачи таксономии связана с именем Карла Линнея (1707–1778), который в ботанике стремился построить систематическое описание растений, чтобы по совокупности признаков изучаемого растения найти его родственников. Термин *таксономия* впервые был предложен в 1813 году О. Декандолем, занимавшимся классификацией растений, и до поры, до времени применялся только в биологии. Как бы ни называлась классификационная группировка в системе классификации, выделяющая определенную группу объектов по некоторому признаку, мы не будем заниматься формальными отличиями понятий образа, кластера и таксона.

В основе методов кластеризации лежит понятие *метрики* – способа определения *расстояния* между элементами множества (разумеется, расстояние здесь понимается существенно шире, чем привычное для нас расстояние по прямой между двумя точками плоскости). Чем меньше это расстояние, тем более похожими считаются изучаемые объекты, тем больше оснований отнести их к одной группе. Обычно элементы задаются в виде набора чисел, а метрика в виде некоторой функции. Методику отнесения элемента к какому-либо кластеру называют *решающим правилом*.

Математический аппарат распознавания образов достаточно молод (60-е годы прошлого века) и практически беспомощен без компьютера с его быстродействием (не лишними являются и возможности экранной графики). Приступая к кластерному анализу, мы реализуем то или иное решающее правило выбором конкретного эвристического алгоритма, не выдвигая в начальной стадии исследования каких-то априорных гипотез относительно классов (таксонов).

Популярный и достаточно универсальный подход к кластеризации связан с алгоритмом так называемой *древовидной кластеризации*, где пытаются объединить объекты (например, животных) в достаточно большие кластеры по мере сходства. Типичным результатом такой кластеризации является приведенное ниже иерархическое дерево (рис. 6.1), построенное по принципу уменьшения уникальности. Так, в ботанике на нижнем его уровне находятся объекты анализа, например, растения *рода* «ель», «пихта», «сосна», «лиственница», которые по совокупности общих, характерных только для них признаков, объединяют в *семейство* «сосновые», которое вместе с кипарисовыми и тиссовыми по аналогич-

ным принципам объединяются в *класс* «хвойные», который вместе с другими образует *подотдел* «голосеменных» и т. д. (отдел, подотдел, класс, подкласс, отряд, семейство, род).

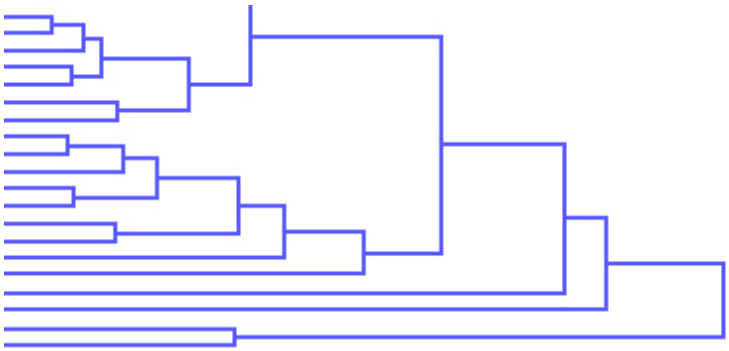


Рис. 6.1. Древовидная классификация

Так, человек по современной классификации в биологии принадлежит к приматам, млекопитающим, позвоночным и животным. Если построить соответствующее дерево, то по биологическим признакам человек оказывается «ближе» к голозадым макакам, чем к благородным овчаркам или сенбернарам.

Если принять за так называемое *расстояние количество несовпадающих признаков* – показателей, измеренных в какой-то числовой шкале (разницу в цвете глаз по стандартному спектру, в росте, калорийности, цене, этажности и др.), то объекты объединяются по степени их близости.

В качестве меры степени близости обычно используются следующие представления:

1) евклидово расстояние $\sqrt{\sum_i (x_i - y_i)^2}$;

2) манхэттенское расстояние $\sum_i | x_i - y_i |$;

3) расстояние Чебышёва $\max_i | x_i - y_i |$;

4) процент несогласия – число несовпадений / общее число.

Наряду с упомянутыми в математической статистике используется обобщение понятия евклидова расстояния – расстояние Махаланобиса как мера расстояния между заданной точкой и центром масс, деленное на ширину эллипсоида в направлении заданной точки. Эта мера предложена индийским экономистом и статистиком Махаланобисом Прасанта Чандра (1893–1972) в 1936 году.

На первом шаге процесса объединения каждый объект считается отдельным кластером и расстояния между объектами определяются выбранной мерой. В дальнейшем возникает вопрос о правиле объединения двух кластеров, о способе определения расстояния между кластерами,

необходимом при близости объектов из разных кластеров (*поиске ближайших* или *наиболее удаленных соседей*).

Метод *одиночной связи* строится на выборе пары объектов в двух кластерах, расстояние между которыми меньше расстояния между любой иной парой объектов этих кластеров (при поиске ближайшего соседа).

Метод *полной связи* выступает более жесткой альтернативой ему и строится на поиске пары наиболее удаленных объектов из разных кластеров, которые находятся друг от друга дальше всех остальных пар объектов.

Другие методы строятся на оценке *среднего расстояния* между какими-то парами объектов.

Так метод *невзвешенного попарного среднего* использует поиск среднего расстояния между *всеми* парами объектов.

Метод *взвешенного попарного среднего* аналогичен методу невзвешенного попарного среднего, но учитывает размеры кластеров (число содержащихся в них объектов), и при вычислениях эти размеры используются в качестве весовых коэффициентов.

Используется и *центроидный метод* с многочисленными вариациями, базирующимися на поиске расстояния между центрами тяжести кластеров. В простейшем варианте *невзвешенного центроидного метода* расстояние между двумя кластерами определяется как расстояние между их центрами тяжести. При *взвешенном центроидном* (*медианном*) *методе* при вычислениях учитываются размеры кластеров.

Популярным методом кластеризации является *метод K-средних*. В отличие от вышеприведенных методов априори выдвигается гипотеза относительно числа кластеров. Здесь дается указание о построении *K* различных кластеров, расположенных на возможно больших расстояниях друг от друга. Затем пытаются выяснить принадлежность объектов к ним так, чтобы минимизировать суммарное квадратичное отклонение элементов кластеров от их центров (минимизировать изменчивость *внутри* кластеров и максимизировать изменчивость *между* кластерами). Можно задать начальные центры кластеров, если такая информация доступна. Предполагается, что выбрано подходящее число кластеров, а в анализ включены все существенные переменные. При неправильном выборе числа кластеров полученные результаты могут быть далеки от ожидаемого. Не вдаваясь в детали технологии поиска, отметим ее родство с дисперсионным анализом.

Важную роль в реализации распознавания образов играет и масштабирование переменных. Если переменные имеют различный масштаб измерений (например, рубли, годы и количество), то и результаты

кластеризации могут быть некорректными. Необходимо подумать и о стандартизации переменных.

Если выработано «решающее правило» на базе множества признаков (факторов), то можно попробовать удалить из него отдельные факторы и сравнить получаемую точность с исходной, определяя тем самым *информативность* (значимость) фактора (вспомните метод главных компонент – см. п. 4.4).

Глава 7. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

7.1. Основные понятия и определения

Наблюдения над каким-то явлением, характер которого меняется во времени, дают последовательность значений

$$u_1 = u(t_1),\ u_2 = u(t_2),\ \ldots,\ u_n = u(t_n), \qquad (7.1)$$

называемую *временным* или *динамическим рядом* (рис. 7.1). В виде динамических рядов представляют динамику валового национального продукта в России за последние 20 лет, спрос на продукцию, динамику цен, валютных курсов и т. д. В силу ограниченности наших знаний или из-за ошибок в наблюдениях значения $u(t)$ могут быть случайными. Заметим, что t может быть не только временем, но и пространственной характеристикой (например, удаленностью от областного центра или высотой над уровнем Мирового океана).

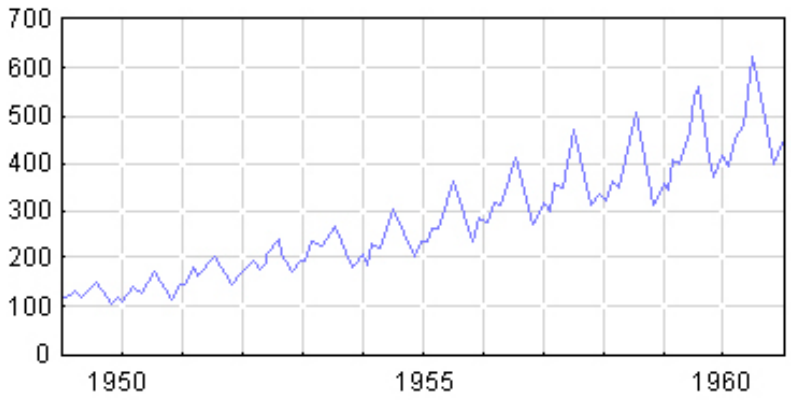


Рис. 7.1. Временной ряд

В принципе, переменная t и функция $u(t)$ могут быть как дискретными, так и непрерывными. Даже не ссылаясь на физическую сущность дискретности, мы вынуждены прибегать к замене (оцифровке) «непрерывного» процесса в силу природной ограниченности восприятия пространства и времени. Едва ли мы можем (и едва ли кому-то это надо) фиксировать биржевые курсы с точностью до минуты, прирост числа жителей Земли за день или рост благосостояния чаще, чем через месяц. Так, урожай овса оценивается с *неизменным* годичным интервалом, перепись населения или поголовья тигров проводится с установленными интервалами (годы, месяцы), данные об осадках фиксируются *в виде суммарных значений* за избранный отрезок времени. Даже выбранные интервалы не всегда соблюдаются – забастовки диспетчеров, землетрясения и т. п. Когда мы видим на экране телевизора статистику достижений Н-ской области в производстве молока или кардиограмму человека в виде ломаных или гладких линий, это иллюзия в угоду нашему эстетическому восприятию – плод выполнения процедуры `LINE` (провести прямую линию между двумя точками экрана) и аппроксимации таблично заданной функции непрерывной с требуемой точностью.

Анализ временных рядов преследует две цели:

1) выяснить природу динамического ряда и дать ее формальное описание;

2) прогнозировать будущие значения по прошлым наблюдениям.

Предполагается, что данные содержат *систематическую составляющую* (обычно включающую несколько факторов) и *случайную ошибку* (*белый шум*).

По мнению ряда авторов, в реальных ситуациях систематическая составляющая складывается из трех компонент:

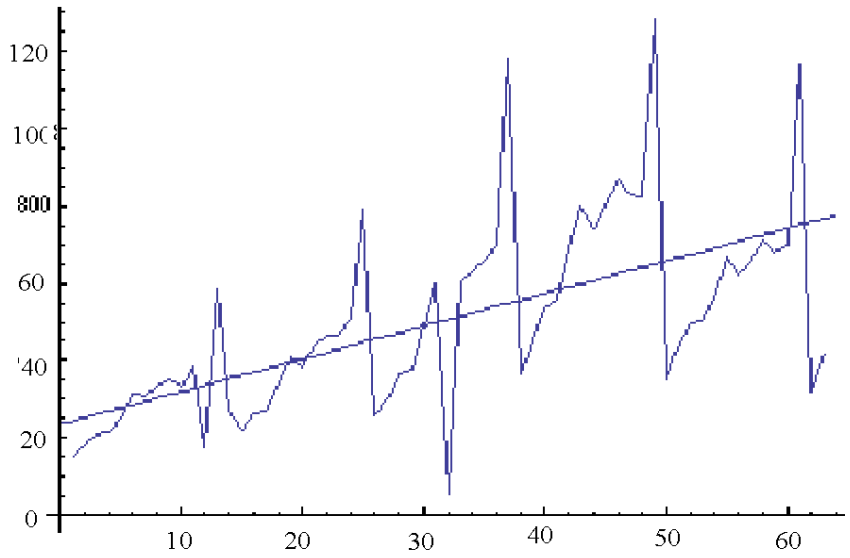


Рис. 7.2. Тренд временного ряда

1) *тренда* (систематического движения – рис. 7.2);

2) *регулярных колебаний* относительно тренда;

3) *эффекта сезонности* (циклов)

Что касается двух последних, многие авторы предпочитают их не разделять и сезонность считать частным случаем регулярных колебаний и даже их отождествлять.

Под трендом временного ряда (рис. 7.2) понимают некое *устойчивое изменение в течение долгого времени*. Однако само понятие «долгое» относительно: при изучении климата столетие – ничто в сравнении с геологическими периодами и колоссальный срок для жизни бабочки. Так заявление о потеплении климата определяется статистикой лишь последних столетий, но многие климатологи объявляют это лишь очередным циклом в жизни Земли (ведь Гренландия – *норв.* Granland, Зеленая страна – предстала таковой викингам 1000 лет назад).

Что касается эффекта сезонности, иногда он выясняется относительно легко (спрос на теплую обувь, уровень потребления газа, яйценоскость деревенских кур связаны с годовым оборотом Земли вокруг Солнца, всплески числа медицинских справок о болезни в вузовской поликлинике в периоды очередной экзаменационной сессии, аварийность грузового автотранспорта – с понедельниками, а посещение занятий – с субботами). Одно и то же явление может быть подвергнуто нескольким признаками сезонности.

Выявив тренд и регулярные колебания, мы не можем дать гарантию, что оставшиеся колебания носят лишь случайный характер. Есть, конечно, возможность статистику из 10 наблюдений аппроксимировать алгебраическим многочленом 9-й степени и тем самым объявить все случайности (погрешности инструментария и человеческого фактора) закономерностью.

На основе вольных и невольных экспериментов и субъективных суждений выработаны некоторые критерии случайности. Один из них связан с вылавливанием экстремальных точек в ряде – *пиков* ($u_i < u_{i+1} > u_{i+2}$) и «ям» ($u_i > u_{i+1} < u_{i+2}$) (концы ряда при этом не учитываются). Математическое ожидание числа таких точек и дисперсия равны соответственно

$$\frac{2}{3}(n-2) \text{ и } (16\,n - 29) / 90. \qquad (7.2)$$

Например, при $n = 10$ ожидаемое количество равно 5,3 и стандартное отклонение $\sigma \approx 1{,}2$. Например, получив 8 экстремальных точек за пределами двух сигм, мы с большой уверенностью можем утверждать наличие невыясненных регулярных колебаний.

Интервал между двумя экстремальными точками называют *фазой*. Математическое ожидание фаз длины d

$$N_d = \frac{2(n-d-2)\left(d^2+3d+1\right)}{(d+3)!}, \; (d = 0, 1, ..., n-3). \qquad (7.3)$$

Так при $n = 10$ значения N_d равны 2,6667; 2,9167; 1,1000; 0,2639; 0,0460; 0,0061; 0,0006; 0,00004.

В случае больших n анализ на случайность можно свести к сравнению получаемого распределения с ожидаемым по критерию χ^2.

Другой критерий связан с асимптотически нормальным распределением знаков разностей – числа интервалов убывания (возрастания), математическое ожидание для которого равно $(n-1)/2$ и дисперсия равна $(n+1)/12$.

Имеются варианты аналогичного подхода, основывающиеся на сравнении всех пар элементов ряда, а не только соседних, с привлечением критерия Спирмена.

7.2. Анализ тренда и сглаживание временных рядов

Не существует *автоматического* способа обнаружения тренда во временном ряде. Если тренд является монотонным (устойчиво возрастает или устойчиво убывает), то анализировать такой ряд обычно нетрудно с помощью аппроксимации полиномом какой-то заданной степени, экспонентой, логистической кривой $u(t) = A / (1 + B\,e^{-Ct})$ и т. п. Тем

не менее выбор хорошей аппроксимации часто не дает специалисту понимание природы изучаемого явления.

Если временные ряды содержат значительную ошибку, первым шагом выделения тренда является *сглаживание* – способ локального усреднения данных, где случайные компоненты взаимно погашают друг друга. Здесь обычно строится полином, аппроксимирующий лишь некоторую часть ряда (*окно*). По нему отыскивается «сглаженное» значение, например, для середины окна; далее берется окно той же длины, но со следующего элемента и т. д. – идея *скользящего среднего*, где каждый член ряда заменяется простым или взвешенным средним m соседних членов. Этот метод часто сглаживает случайные колебания и позволяет применять ряд, очищенный даже от сезонности, при долгосрочном прогнозировании.

Пусть дан отрезок ряда длины $2m + 1$:

$$u_{-m}, u_{-(m-1)}, \ldots, u_0, \ldots, u_m.$$

В примитивном варианте линейного сглаживания (рис. 7.3) новое U_0 равно

$$U_0 = \frac{1}{2m+1} \sum_{t=-m}^{m} u_t \,. \qquad (7.4)$$

t	1	2	3	4	5	6	7	8	9	10
$u(t)$	4	6	5	8	10	8	14	12	14	10
$u_1(t)$	4	5	6,33	7,67	8,67	10,67	11,33	13,33	12	10
$u_2(t)$	4	5	6,33	7,56	9	10,22	11,78	12,22	12	10

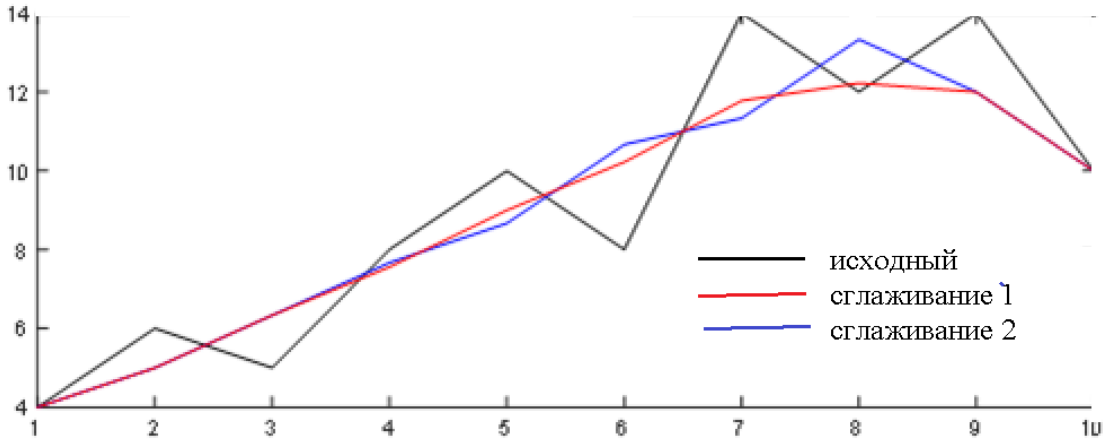


Рис. 7.3. Сглаживание методом скользящего среднего ($m = 1$)

Если вести сглаживание полиномом степени p, то его коэффициенты легко найти, решая систему (вспомните метод наименьших квадратов)

$$\frac{\partial}{\partial a_k}\sum_{i=-m}^{m}(a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p - u_i)^2 = 0, \qquad k = \overline{0, p}. \quad (7.5)$$

Например, при $m = 3$, $p = 3$ с учетом того, что при нечетных k сумма t^k обращается в нуль, получаем систему

$$\begin{cases} 7a_0 + 28a_2 = \sum_{t=-3}^{3} u_t \\ 28a_1 + 196a_3 = \sum_{t=-3}^{3} u_t \\ 28a_0 + 196a_2 = \sum_{t=-3}^{3} u_t \\ 196a_1 + 1588a_3 = \sum_{t=-3}^{3} u_t \end{cases}$$

откуда 21 $a_0 = 7\,\Sigma u_t - \Sigma t^2 u_t$ и

$$U_0 \equiv a_0 = \tfrac{1}{21}[-2u_{-3} + 3u_{-2} + 6u_{-1} + 7u_0 + 6u_1 + 3u_2 - 2u_3]. \quad (7.6)$$

Если взят временной ряд для $t = 1, 2, 3, \ldots$

0, 1, 8, 27, 64, 125, 216, 343, 512, 729, 1000,

то значение тренда для $m = 3$ (подставьте и убедитесь!) совпадает с 27 (ряд отражает тренд $(m - 1)^3$ без флуктуаций).

Если взять четвертый элемент равным 26 (с ошибкой), то сглаженная его оценка равна 26,6667.

u	0	1	8	26	60	126	216	343	512	729	1000
U	0	1	8	25,7	62,7	124	215,8	343,5	512	729	1000
	0	1	8	26,1	62,8	124,1	215,8	343,1	512	729	1000

При квадратическом 5-точечном сглаживании

$$a_0 = \tfrac{1}{35}[-3u_{-2} + 12u_{-1} + 17u_0 + 12u_1 - 3u_2], \quad (7.7)$$

а при 7-точечном совпадает с приведенным кубическим. Существует множество аналогичных сглаживающих оценок при $p = 2, 3, \ldots, 5$ и $m = 5, 7, \ldots, 21$.

Вместо среднего можно использовать и медиану значений, попавших в окно. Ряд авторов полагает, что *медианное сглаживание* обычно приводит к более гладким или, по крайней мере, более «надежным» кривым, по сравнению со скользящим средним с тем же самым окном.

Существуют приемы сглаживания, базирующиеся на *конечных разностях*, где разности некоторого порядка отбрасываются и используются интерполяционные формулы.

Ряды с относительно небольшим количеством наблюдений и систематическим расположением точек могут быть сглажены с помощью *кубических сплайнов* (напомним, что *сплайн-аппроксимация* сглаживает не только функцию, но и ее производные).

Недостаток рассмотренного метода скользящего среднего в том, что он не дает тренда для начала и конца временного ряда (особенно это неприятно для конца ряда из-за невозможности экстраполяции в будущее). Однако никто не запрещает найти решение системы уравнений не только относительно a_0, но и других коэффициентов и построить сглаживание для других элементов ряда. Например:

$$U_3 = \tfrac{1}{42}[-2u_{-3} + 4u_{-2} + u_{-1} - 4u_0 - 4u_1 + 8u_2 + 39u_3]. \qquad (7.8)$$

Иногда применяют неоднократное сглаживание временного ряда, но это чревато «превращением развесистого дерева в телеграфный столб».

Какой должна быть длина окна? Чем она больше, тем сильнее сглаживание. При большом окне, содержащем несколько колебаний, компонента значительно сглаживается. Если же период колебаний больше длины окна, то циклическая компонента ряда будет восприниматься как тренд.

Если периодические колебания отсутствуют, порядок скользящей средней подбирают начиная с наименьшего, укрупняя до тех пор, пока в средней не будет выступать тенденция развития процесса.

Популярным методом прогнозирования многих временных рядов является э*кспоненциальное сглаживание*.

Простая идея – с течением времени прошлое постепенно оказывается менее значимым для решений текущего момента. Способы плавки стали в XIX веке или передачи данных в недалеком прошлом устаревают из года в год. Даже этика Аристотеля и десяти заповедей забывается в процессе непрерывной вульгаризации общества. Первые приложения экспоненциального сглаживания связаны с обнаружением подводных лодок, использованием систем наведения во время Второй мировой войны, прогнозированием спроса на запасные части и др.

При *простом экспоненциальном сглаживании* используется сглаживание скользящим средним, в котором последним наблюдениям приписываются бо′льшие веса, чем предпоследним, и учитываются *все* предшествующие наблюдения ряда. Формула такого сглаживания имеет вид

$$S_t = \alpha\, u_t + (1 - \alpha)\, S_{t-1}. \qquad (7.9)$$

Результат сглаживания зависит от параметра α. Как же выбрать лучшее значение параметра α? Обычно берут $0 < \alpha < 1$, хотя некоторые

авторы принимают $0 < \alpha < 2$. Одни предлагают (на основе практики) брать $\alpha < 0{,}30$, другие утверждают иное. Разумно брать значения в диапазоне от 0 до 1 с некоторым шагом и выбирать такое α, для которого сумма квадратов разницы между наблюдаемыми значениями и прогнозом минимальна. Индикаторами качества подбора могут служить средняя ошибка, средняя абсолютная или относительная ошибка $100\ (u_t - S_t) / u_t$ и др.

Заметим, что в зависимости от выбора параметра α (в частности, если α близко к 0), начальное значение S_0 может оказать существенное воздействие на прогноз для многих последующих наблюдений. С ростом числа наблюдений влияние S_0 уменьшается.

7.3. Анализ сезонных колебаний

В ряде случаев периоды сезонных колебаний известны, в других они укрыты на фоне других колебаний и шума. Очевидно, что продажа ювелирных изделий и цветов в России обычно возрастает в периоды перед Новым годом, Рождеством и 8 Марта, затраты на ремонт овощехранилищ в июле–августе, продажа елочных игрушек в конце декабря, куриных яиц перед Пасхой, а аварий и прогулов на производстве по понедельникам. Все подобные явления достаточно устойчивы по времени (незначительный сдвиг по фазе) и объему (по амплитуде), но в кризисные годы для некоторых явлений в соответствующих временных рядах это сезонное явление практически теряется на уровне шума, и компьютерный анализ подвергнет сомнению наличие сезонного колебания. Разумеется, объемы продажи яиц значимы лишь для их производителей и в общем бюджете отрасли их роль сводится к некому усредненному показателю.

Есть факторы, значимые для бюджета многих стран, в динамике которых одни специалисты усматривают систематические регулярные колебания и даже объявляют конкретную периодичность, а другие столь же авторитетно опровергают. Конечно, цена на нефть периодически меняется от 30 до 100 долларов за баррель, но периодичность меняется в зависимости от темпов развития экономики Китая, добычи на сланцевых месторождениях США и непредугаданных причуд американского президента.

В основе анализа сезонности лежит понятие *автокорреляционной функции* – показателя взаимосвязи между функцией и ее сдвинутой копией от величины временного сдвига. Если исходная функция (временной ряд) строго периодическая, то на графике автокорреляционной

функции тоже будет строго периодическая функция, и тогда можно судить о периодичности исходной функции.

Формально периодическая зависимость может быть определена как сериальная зависимость порядка k между каждым u_t и u_{t+k}. Ее можно измерить с помощью *коэффициентов автокорреляции* (то есть корреляции между самими членами ряда). Величину k обычно называют *лагом* (эквивалентные термины *сдвиг*, *запаздывание*):

$$\rho_k = \frac{\frac{1}{n-k+1}\sum_{t=0}^{n-k} A_t B_{t+k}}{\sqrt{\left[\frac{1}{n-k+1}\sum_{t=0}^{n-k} A_t^2\right]\left[\frac{1}{n-k+1}\sum_{t=0}^{n-k} B_{t+k}^2\right]}}; \qquad (7.10)$$

$$A_t = u_t - \frac{1}{n-k+1}\sum_{i=0}^{n-k} u_i, \quad B_{t+k} = u_{t+k} - \frac{1}{n-k+1}\sum_{i=0}^{n-k} u_{i+k}.$$

На практике используют формулу с центрированием относительно среднего и нормировкой посредством общей дисперсии:

$$\rho_k = \frac{M\{(u_t-\mu)(u_{t+k}-\mu)\}}{\sigma^2} = \frac{1}{n-k+1}\sum_{t=0}^{n-k}\frac{(u_t-\mu)(u_{t+k}-\mu)}{\sigma^2}, \; k = 0, 1, \ldots, n-1, \quad (7.11)$$

но для рядов небольшой длины при значительных k она может привести к коэффициентам, превышающим 1.

Последовательность коэффициентов *автокорреляции* называют *коррелограммой* (*автокоррелограммой*). Она показывает, как часто и с каким запаздыванием изменение показателя u_t сказывается на последующих значениях. Сдвиг, которому соответствует наибольший коэффициент автокорреляции, дает временное запаздывание (*лаг*). Найдя линейный тренд и удалив его (рис. 7.4), получаем автокоррелограмму (рис. 7.5), где виден лаг порядка 29–30.

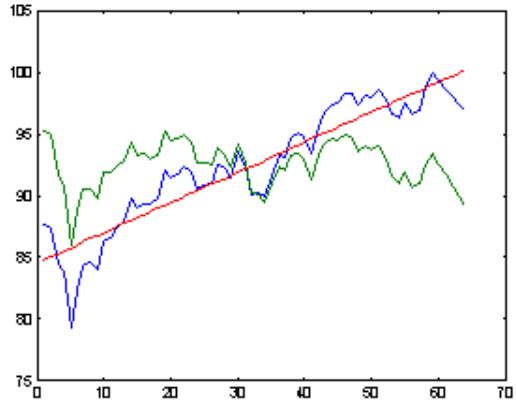


Рис. 7.4. Исходный ряд, тренд, ряд без тренда

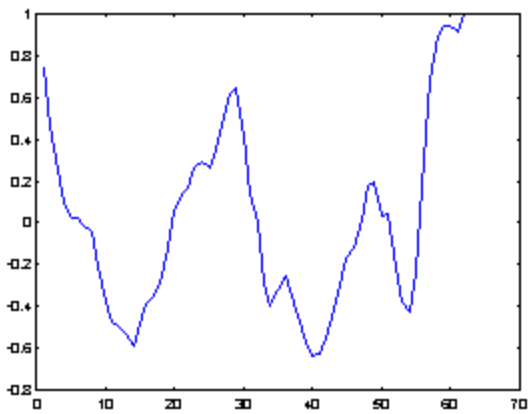


Рис. 7.5. Автокоррелограмма

Математическая модель авторегрессии без учета тренда (или после его исключения) имеет вид

$$u_t = \sum_{i=1}^{k} a_i u_{t-i} + \varepsilon, \qquad (7.12)$$

где a_i – параметры, оцениваемые по временному ряду с помощью метода наименьших квадратов (так называемые *коэффициенты линейного предсказания*); ε – независимая случайная величина.

После обнаружения некоторой периодической зависимости с лагом k с целью ее удаления из последующего анализа из каждого i-го элемента ряда вычитается $(i - k)$-й элемент. Такой прием позволяет определить скрытые периодические составляющие ряда. Так как автокорреляции на последовательных лагах зависимы, то удаление некоторых автокорреляций изменит другие сезонные составляющие и сделает их более заметными.

Для анализа временных рядов применяют и аппарат *теории случайных процессов*, и, в частности, *спектральный анализ*.

Как известно, случайные процессы, характеристики которых в любом интервале времени неизменны и не зависят от точки начала отсчета, называются *стационарными*. В экономической практике приходится в большинстве случаев иметь дело с нестационарными случайными процессами, но на определенных интервалах времени они могут быть близки к стационарным.

Аппарат спектрального (*гармонического*) анализа базируется на представлении функции отрезком ряда Фурье

$$u(t) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(\frac{2\pi k}{n} t\right) + b_k \sin\left(\frac{2\pi k}{n} t\right) \right], \qquad (7.13)$$

где n – число точек, k – номер и $c_k = \sqrt{a_k^2 + b_k^2}$ – амплитуда гармоники,

$$a_0 = \frac{1}{n}\sum_{i=1}^{n} u_i, \; a_k = \frac{2}{n}\sum_{i=1}^{n} u_i \cos\left(\frac{2\pi k}{n} i\right), \; b_k = \frac{2}{n}\sum_{i=1}^{n} u_i \sin\left(\frac{2\pi k}{n} i\right), \; k = 1, 2, \ldots .$$

Дисперсия, учитываемая одной гармоникой, определяется как $c_k^2 / 2$, и поскольку гармоники не коррелируют друг с другом, то отношение их суммы к начальной дисперсии определяет долю учитываемой дисперсии.

Существует алгоритм *быстрого преобразования Фурье* (БПФ), где время выполнения анализа ряда длины N пропорционально величине $N \log_2(N)$, а число элементов ряда должно быть равным степени 2.

В последние 70 лет появилось великое множество алгоритмов *анализа временных рядов*, основанных на сочетании регрессионного анализа, спектрального анализа и эвристических подходов, вокруг которых не утихают споры. Существует многообразие специфических подходов к анализу информации, предлагаемой статистикой рынка ценных бумаг, страхового дела. Хотя никаких революционных идей в обработке данных здесь не возникло, существует много фирм, дающих вполне правдоподобные прогнозы.

В заключение разговора о временных рядах отметим, что успех реального поиска тренда и тем более регулярных колебаний зависит от конкретного приложения и субъективных мнений исследователей. В литературе любознательный читатель может обнаружить сотни серьезных и не очень публикаций на эту тему.

7.4. Технологии анализа временных рядов средствами MS Excel

Пример 1. Для временного ряда курса акций фирмы IBM (табл. 7.1) требуется рассчитать трех- и семичленные скользящие средние, графически сравнить результаты.

Таблица 7.1. Временной ряд курса акций фирмы IBM (долл.)

t	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
y_t	510	497	504	510	509	503	500	500	500	495	494	499
t	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
y_t	502	509	525	512	510	506	515	522	523	527	523	528

	A	B
1	t	y t, долл.
2	1	510
3	2	497
4	3	504
5	4	510
23	22	527
24	23	523
25	24	528

Рис. 7.6. Исходные данные

Решение. Введем исходные данные в диапазон `A1:B25` листа MS Excel (рис. 7.6). Вычислим трех- и семичленные скользящие средние, выполнив команду `Данные – Анализ данных – Скользящее среднее`. Заполним параметры диалоговых окон (рис. 7.7, 7.8). В обоих случаях выбираем `OK`.

Результаты расчетов представлены в табл. 7.2. Графический анализ наглядно показывает (рис. 7.9), что чем больше длина интервала сглаживания, тем более гладкий ряд получается на выходе модели.

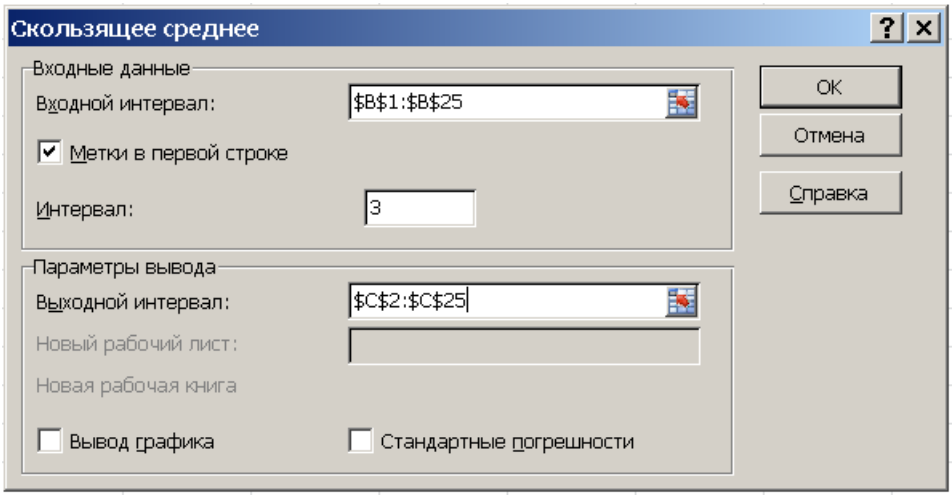


Рис. 7.7. Диалоговое окно «Скользящее среднее» ($l = 3$)

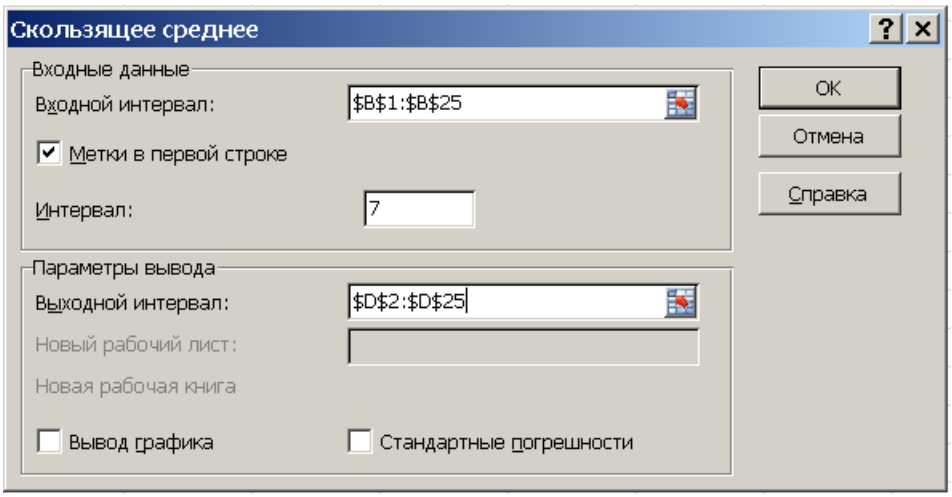


Рис. 7.8. Диалоговое окно «Скользящее среднее» ($l = 7$)

Таблица 7.2. Сглаживание временного ряда курса акций фирмы IBM (долл.) с помощью скользящих средних

t	y_t, долл.	Скользящие средние		t	y_t, долл.	Скользящие средние	
		$l = 3$	$l = 7$			$l = 3$	$l = 7$
1	510	–	–	**13**	502	503,3	505,1
2	497	503,7	–	**14**	509	512,0	507,3
3	504	503,7	–	**15**	525	515,3	509,0
4	510	507,7	504,7	**16**	512	515,7	511,3
5	509	507,3	503,3	**17**	510	509,3	514,1
6	503	504,0	503,7	**18**	506	510,3	516,1
7	500	501,0	502,4	**19**	515	514,3	516,4
8	500	500,0	500,1	**20**	522	520,0	518,0
9	500	498,3	498,7	**21**	523	524,0	520,6
10	495	496,3	498,6	**22**	527	524,3	–
11	494	496,0	499,9	**23**	523	526,0	–
12	499	498,3	503,4	**24**	528	–	–

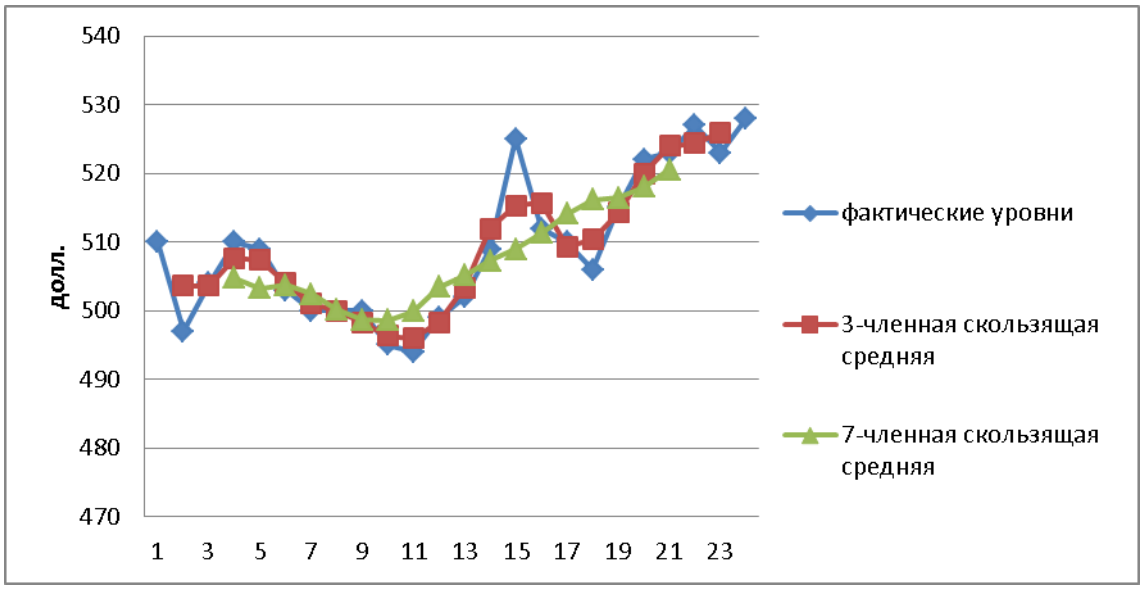


Рис. 7.9. Сглаживание временного ряда курса акций с помощью скользящих средних

Скользящие средние являются важным инструментом, имеющим разнообразные приложения в статистических исследованиях. Они широко используются в техническом анализе при изучении финансовых и товарных рынков. Скользящие средние применяются при оценивании сезонной составляющей во временных рядах, в процедурах сезонной корректировки, часто на практике используются совместно с моделями кривых роста.

Пример 2. Для временного ряда объема продаж продукции фирмы (табл. 7.3) требуется рассчитать экспоненциальную среднюю. Расчеты произвести для трех различных значений параметров адаптации: а) $\alpha = 0{,}1$; б) $\alpha = 0{,}5$; в) $\alpha = 0{,}9$. Сравнить графически исходный временной ряд и экспоненциально сглаженные временные ряды при различных значениях параметра адаптации. Указать, какой временной ряд носит более гладкий характер.

Таблица 7.3. Временной ряд объема продаж продукции фирмы

***t*, номер квартала**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
y_t, объем продаж (тыс. шт.)	235	234	227	222	218	199	197	203	208
***t*, номер квартала**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
y_t, объем продаж (тыс. шт.)	212	217	232	230	220	213	213	219	

Решение. Введем исходные данные в диапазон A1:B18 рабочего листа MS Excel (рис. 7.10). Вычислим значения экспоненциальной средней при α = 0,1, выполнив команду Данные – Анализ данных – Экспоненциальное сглаживание. Затем заполним параметры диалогового окна (рис. 7.11). В диалоговом окне в графе Фактор затухания следует поставить значение 1 – α. После чего выбираем OK. Аналогичны настройки и вычисления для α = 0,5 и α = 0,9.

	A	B
1	t, номер квартала	yt, объем продаж (тыс. шт.)
2	1	235
3	2	234
4	3	227
5	4	222
16	15	213
17	16	213
18	17	219

Рис. 7.10. Исходные данные

Результаты расчетов экспоненциально сглаженных рядов при различных значениях параметров адаптации представлены в табл. 7.4. Заполнение табл. 7.4. осуществлялось в предположении, что прогноз на первый квартал совпал с фактическим значением.

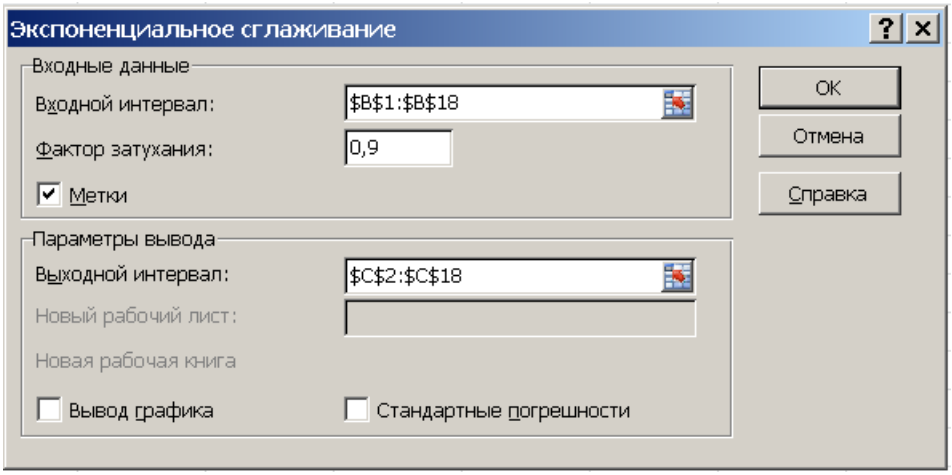


Рис. 7.11. Диалоговое окно «Экспоненциальное сглаживание» (α = 0,1)

На рис. 7.12 наглядно проявляется влияние значения параметра адаптации на характер сглаженного ряда. При α = 0,1 экспоненциальная средняя носит более гладкий характер, так как в этом случае в наибольшей степени поглощаются случайные колебания временного ряда.

Пример 3. Используя режим Анализ Фурье, вычислить коэффициенты прямого дискретного преобразования Фурье (ДПФ) от временной последовательности, значения которой приведены в столбце B документа MS Excel, изображенного на рис. 7.13.

Таблица 7.4. Экспоненциальные средние для временного ряда объема продаж продукции фирмы

***t*, номер квартала**	***y_t*, объем продаж (тыс. шт.)**	**Экспоненциальная средняя**		
		α = 0,1	**α = 0,5**	**α = 0,9**
1	235	235,0	235,0	235,0
2	234	235,0	235,0	235,0
3	227	234,9	234,5	234,1
4	222	234,1	230,8	227,7
5	218	232,9	226,4	222,6
6	199	231,4	222,2	218,5
7	197	228,2	210,6	200,9
8	203	225,1	203,8	197,4
9	208	222,8	203,4	202,4
10	212	221,4	205,7	207,4
11	217	220,4	208,8	211,5
12	232	220,1	212,9	216,5
13	230	221,3	222,5	230,4
14	220	222,1	226,2	230,0
15	213	221,9	223,1	221,0
16	213	221,0	218,1	213,8
17	219	220,2	215,5	213,1

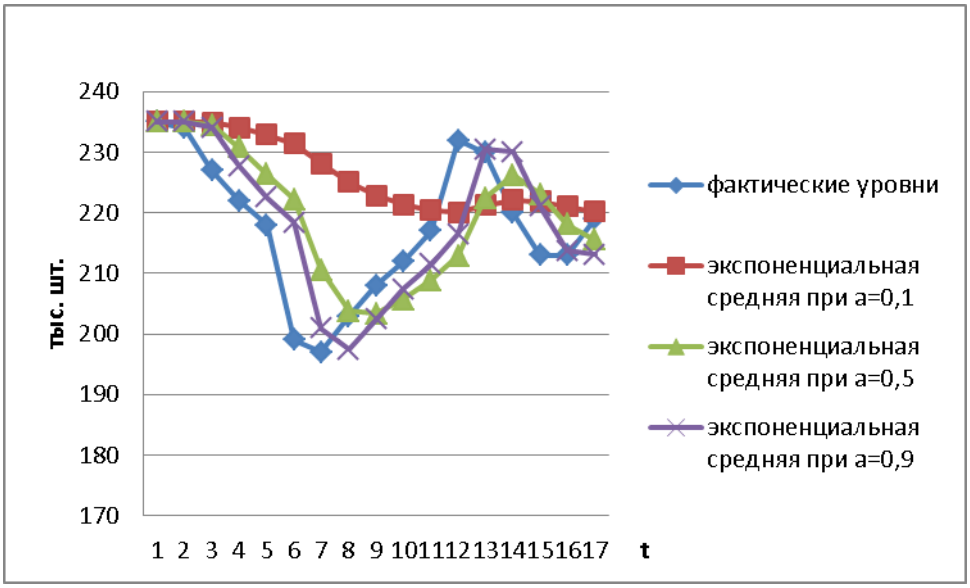


Рис. 7.12. Экспоненциальное сглаживание при различных значениях параметра адаптации

Решение. Обратимся к режиму Анализ Фурье, выполнив команду Данные – Анализ данных – Анализ Фурье, и зададим в появившемся диалоговом окне необходимые параметры (рис. 7.13):

1) Входной интервал – диапазон ячеек, содержащих вещественные данные, к которым применяется ДПФ;

2) Метки в первой строке – включается, если первая строка содержит заголовки;

3) Выходной интервал – вводится адрес ячеек выходного диапазона;

4) Инверсия – включается, если необходимо вычислить обратное ДПФ.

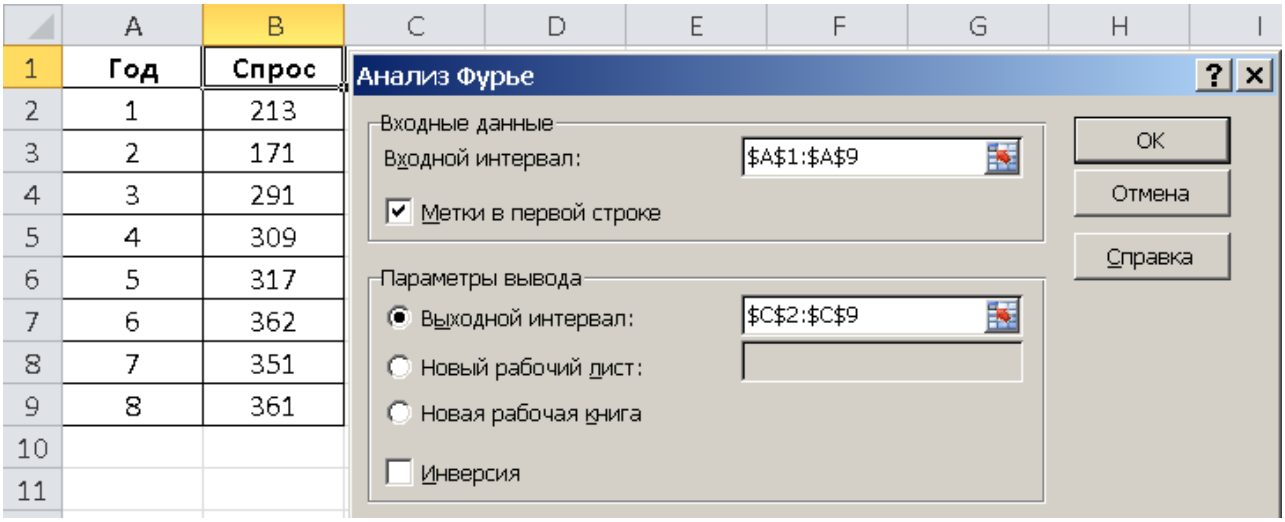


Рис. 7.13. Диалоговое окно режима «Анализ Фурье»

Результат показан на рис. 7.14. Так как вычисленные коэффициенты ДПФ будут комплексными числами вида $a_k + i\,b_k$, то для нахождения отдельно вещественной и мнимой части можно использовать следующие функции рабочего листа MS Excel (категория Инженерные): МНИМ.ВЕЩ(), МНИМ.ЧАСТЬ().

	A	B	C	D	E	F
1	**Год**	**Спрос**				
2	1	213	36			
3	2	171	-3,99999999999999+9,65685424949238i			
4	3	291	-4+4i			
5	4	309	-3,99999999999999+1,65685424949238i			
6	5	317	-4			
7	6	362	-4-1,65685424949238i			
8	7	351	-4-4i			
9	8	361	-4,00000000000001-9,65685424949238i			

Рис. 7.14. Результат работы режима Анализ Фурье

Замечание. Используемый для вычисления ДПФ алгоритм, называемый алгоритмом *быстрого преобразования Фурье*, требует, чтобы n (число значений временного ряда) обязательно было равным степени числа 2 (то есть 8, 16, 32, 64, …), что является существенным ограничением.

Контрольные вопросы

1) Дайте определение временного ряда.

2) В чем заключается природа дискретности временного ряда при проведении статистических исследований?

3) Назовите две основные цели анализа временных рядов.

4) Из каких компонент складывается систематическая составляющая временного ряда?

5) Что понимается под трендом временного ряда?

6) Дайте определение модели с мультипликативной сезонностью.

7) Могут ли концы временного ряда служить экстремальными точками?

8) Какой тренд называется монотонным?

9) Для чего применяется процедура сглаживания при выделении тренда?

10) Каким образом осуществляется сглаживание методом скользящего среднего?

11) Какова процедура линейного сглаживания?

12) В чем заключаются выгоды и недостатки сглаживания полиномом степени p?

13) Что собой представляет квадратическое сглаживание?

14) Почему метод скользящего среднего можно рассматривать как фильтр низких частот?

15) Что такое медианное сглаживание?

16) В чем заключается недостаток метода скользящего среднего?

17) К чему приводит неоднократное применение методов сглаживания?

18) В чем заключается суть метода простого экспоненциального сглаживания?

19) Какие существуют варианты выбора параметра α при экспоненциальном сглаживании?

20) В чем смысл сезонной составляющей временного ряда?

21) В чем отличие аддитивной и мультипликативной модели, учитывающей сезонный фактор?

22) Что представляет собой последовательность коэффициентов автокорреляции, лаг?

23) Что показывает коррелограмма?

24) Какова цель спектрального (гармонического) анализа?

Глава 8. В МОРЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Если вернуться на 60 лет назад и заглянуть в первые библиотеки стандартных программ (БСП), то мы обнаружим превосходные по быстродействию и минимуму требований к памяти программы вычисления элементарных и специальных функций, линейной алгебры, решения обыкновенных дифференциальных уравнений (мы не упоминаем о стандартных программах решения уравнений математической физики, не поставляемых пользователям в составе прилагаемого математического обеспечения).

Представители бизнеса постепенно осознали, что внушающие трепет электронные монстры было бы неплохо приспособить к решению рутинных задач учета и планирования. В 1960 году в США ушел в плавание язык программирования COBOL (COmmon Business Oriented Language), предназначенный для разработки бизнес-приложений и скептически принятый в Старом Свете. Постепенно быстродействие ЭВМ, емкость оперативной памяти и скорость доступа к внешней памяти достигли уровня, необходимого для решения громоздких задач бизнеса. Уже в 1964 году разработчики популярного языка программирования ПЛ/1 (Programming Language) предусмотрели работу с массивами и структурами для организации баз данных.

С расширением круга пользователей в БСП новых алгоритмических языков появились и общеизвестные термины статистического анализа (среднее, дисперсия, корреляция и т. п.). Эта терминология проникла и в специализированные системы программирования, ярким представителем которых является доступный пониманию обычного студента язык MatLab (Matrix Laboratory), обеспечивающий работу с массивами, комплексную арифметику, отлично выполненную математику от синуса и корней многочлена до проблемы собственных значений, превосходную графику (60 лет назад мы знали авторов многих СП, сегодня программное обеспечение, увы, стало безымянным).

8.1. Инструмент статистики в MatLab

Мы не ставим целью дать полное представление о функциях Statistics Toolbox и ограничимся упоминанием малой доли возможностей, предоставляемых пользователю, не имеющему опыта работы со статистическими пакетами (не всегда надо пользоваться кувалдой, чтобы забить гвоздь) и знакомого лишь с азами статистического анализа.

Естественно, в Statistics Toolbox предусмотрены функции вычисления плотности для многообразия распределений случайных величин: дискретного и непрерывного равномерного, нормального, логнормаль-

ного, экспоненциального, гамма- и бета-распределений, Рэлея, Вейбулла, биномиального, отрицательного биномиального, Пуассона, геометрического и гипергеометрического, хи-квадрат, Стьюдента, Фишера. Там же содержатся нетривиальные средства вычисления значений функций эмпирического и вышеприведенных распределений, обратных функций распределения, оценки параметров законов распределения по экспериментальным данным, генерации псевдослучайных чисел по заданному закону распределения.

В этом пакете прикладных программ предусмотрены функции оценки математического ожидания и дисперсии по заданному закону распределения и его параметрам, стандартного отклонения, стандартного отклонения среднего значения, медианы, асимметрии, эксцесса, размаха, построения матрицы ковариаций, коэффициентов корреляции, множественной линейной регрессии, доверительных интервалов для линии регрессии, полиномиальной и экспоненциальной регрессии (`polyfit` – построение многочлена заданной степени, `polyval` – прогноз на основе найденного многочлена и `smoothts(Y)` – экспоненциальное сглаживание временного ряда).

Реализованы средства статистической проверки гипотез (по Вилкоксону и Стьюденту), одно-, двух- и многофакторного дисперсионного анализа, снижения размерности задачи на основе метода главных компонент, проверки статистических гипотез о согласии распределения экспериментальным данным (с нормальным распределением, с заданным по тесту Колмогорова – Смирнова), проверки непараметрических гипотез по тесту Вилкоксона и др.

Любопытным элементом пакета Statistics Toolbox являются функции кластерного анализа: `cluster` – деление иерархического дерева кластеров на отдельные кластеры, `clusterdata` – группировка матрицы исходных данных в кластеры, `cophenet` – расчет коэффициента качества разбиения исходных данных на кластеры и др.

Конечно, комментарии и описание упомянутых функций часто требуют от читателя определенного представления о предмете его интереса, о русскоязычной терминологии, чуть-чуть представления о методах достижения цели и азбуке MatLab.

Можно обратиться к какому-либо популярному статистическому пакету, в соответствии с кнопочной технологией безошибочно ввести исходные данные и представить полученные результаты с англоязычными комментариями заказчику, верящему всему, что «посчитал компьютер». Однако именно интерпретация полученных результатов –

весьма нетривиальная задача, требующая определенных знаний терминологии и особенностей исследуемой предметной области.

8.2. Пакеты прикладных программ

На рынке статистического программного обеспечения существует множество пакетов прикладных программ, которые в той или иной степени призваны решать задачи статистического анализа данных. Все их многообразие можно условно разделить на специализированные пакеты и пакеты общего назначения.

Специализированные пакеты ориентированы на использование в специфической предметной области (страховое дело, маркетинг, статистика ценных бумаг и др.). В них, наряду с традиционными методами математической статистики, анализа и прогноза временных рядов, реализованы интересные эвристические алгоритмы. Среди этих пакетов есть как зарубежные, так и отечественные: Forecast Expert, Stat-Media и т. п. Довольно много на рынке программного обеспечения и специализированных статистических экспертных систем.

Самыми популярными на мировом рынке статистических программ являются пакеты общего назначения, большинство из которых обладают продуманным, дружественным к пользователю интерфейсом и относительно подробной документацией, широким спектром статистических функций, что привлекает как специалистов, так и новичков в статистике. Но одним из мировых лидеров, широко известным в России, остается пакет Statistica (разработка фирмы StatSoft Inc., США) – интегрированная система для комплексного статистического анализа и обработки данных в среде MS Windows.

Примечателен комментарий фирмы о причинах разработки и целях использования пакета. «Жесткая конкуренция ставит перед современным бизнесом проблемы, которые могут быть эффективно решены средствами анализа данных. Основа любого бизнеса – постоянно обновляемые, живые базы данных, содержащие множество разнообразной информации. Однако зачастую эти данные лежат мертвым грузом вместо того, чтобы приносить прибыль. Статистические методы позволяют анализировать информацию, прогнозировать развитие и минимизировать риски при принятии решений, что позволяет строить бизнес на строго научной основе.

Мы поставили и развили в России современные методы анализа данных. Теперь самые эффективные технологии анализа: описательный анализ, визуализация, построение объяснительных моделей, классификация, нейронные сети, добыча данных и многие другие доступны полностью на русском языке. Эти методы применимы во всех областях че-

ловеческой деятельности: бизнесе, маркетинге, экономике, промышленности, медицине и др.[1]

В финансовой сфере статистические методы позволяют строить прогнозы продаж, проводить анализ банковских остатков, эффективности выдачи кредитов и т. д.

В маркетинге наши методы позволяют организовать управление клиентами (CRM-management), скоринг, классификацию, прогнозирование, анализ анкет, количественные исследования и др.

В промышленности с помощью Statistica можно эффективно обеспечить всесторонний контроль качества, включая текущий менеджмент качества, поиск причин потери качества, анализ процессов и др. Уникальные технологии планирования экспериментов позволяют создавать новые материалы с заданными свойствами, используя минимальное количество опытов.

В медицине и фармакологии статистические методы позволяют проводить клинические испытания лекарственных препаратов, разрабатывать новые технологии диагностики и методики лечения, подтверждать значимость полученных эффектов».

Базовый комплект поставки Statistica Base предоставляет широкий набор основных статистик в понятном интерфейсе (описательные и внутригрупповые статистики, разведочный анализ данных, корреляция и множественная регрессия, критерий Стьюдента и другие критерии групповых различий, непараметрические статистики, дисперсионный анализ (ANOVA / MANOVA), подгонка распределений.

Система Statistica содержит новейшие компьютерные и математические методы анализа данных и объединяет в себе электронные таблицы для ввода, задания и преобразования исходных данных, мощную графическую систему визуализации данных и результатов анализа, набор специализированных статистических модулей, встроенные языки программирования Statistica Command Language и Statistica BASIC, позволяющие пользователю расширить стандартные возможности системы.

Пакет действительно обладает колоссальными возможностями и потому требует от пользователя владения статистической терминологией (рядовой пользователь, которому не нужна такая мощь, пробившись через систему подготовки данных, получает избыток англоязычной информации, которая ему непонятна из-за отсутствия соответствующего математического образования).

---

[1] Некоторое традиционное преувеличение роли «цивилизаторской миссии» США.

Глава 9. ПРАКТИКУМ ПО СТАТИСТИЧЕСКОМУ АНАЛИЗУ

Эта глава содержит методические указания и задания к практическим занятиям для студентов направления подготовки «Прикладная информатика». Цель представленного здесь руководства – дать знания и выработать навыки для последующей работы в практике применения информационных технологий при статистическом анализе и моделировании в экономике и смежных сферах, имеющих дело с обработкой результатов эксперимента. Здесь мы ограничиваемся рассмотрением трех традиционных базовых задач прикладного статистического анализа.

Цель первой из них – оценка основных параметров эмпирического распределения вероятностей для некоторой выборки данных по некоторому фактору, подверженному случайности. Следующим этапом работы является выбор наиболее близкого к нему теоретического распределения из небольшого списка известных, чтобы обеспечить простоту моделирования значений этого фактора при возникновении необходимости в этом.

Вторая задача связана с установлением степени линейной или нелинейной связи (корреляции) значений некоторого признака от совокупности значений некоторых факторов, построением соответствующих уравнений с целью в той или иной степени надежного предсказания значения признака при изменении значений факторов.

Третья задача связана с обработкой так называемых временных рядов. Из-за нетривиальности используемого математического аппарата мы ограничимся лишь попыткой обнаружения тренда.

Как было отмечено выше, в России наиболее популярен пакет Statistica – система статистического анализа и обработки данных в среде MS Windows, требующая владения англоязычной статистической терминологией (дословный компьютерный перевод не всегда согласуется с общепринятыми в отечественной практике терминами и подчас дает абсурдные высказывания, а выходные документы с англоязычными надписями не внушают положительных эмоций у отечественного потребителя).

Не отвергая кнопочных технологий, мы предлагаем при решении вышеуказанных задач использовать описанные ранее возможности MS Excel и систему MatLab с тем, чтобы на уровне расчетных формул и даже библиотеки программ осознавать смысл, технологию и интерпретацию получаемых оценок. Создаваемые программные модули должны свидетельствовать о способности исполнителя к программированию расчетных задач.

Предлагаемый практикум дублирует многие расчетные формулы и комментарии, представленные в основной части учебного пособия, хотя было бы разумно для читателя хотя бы «по диагонали» получить общее представление об азбуке статистического анализа.

ТЕМА 1. Обработка эмпирических распределений

Этап 1. Выбор данных для анализа

Исполнителю предлагается самостоятельно отыскать в Интернете или других источниках реальную статистику $X\{x_i, i = 1, 2, \ldots, N\}$ какого-либо интересного показателя (фактора). Объем выборки N не обязательно должен быть очень большим, но не меньшим 20–30 наблюдений.

Если показатель связан с каким-то материальным субъектом (выпуск телеаппаратуры, автомобилей, потребление алкоголя или численность населения), то в результате последующего анализа получаемые оценки могут привести к оценкам, вполне объяснимым с точки зрения здравого смысла или истории цивилизации.

Но к показателям, связанным с денежными расчетами, следует относиться с осторожностью. В России до конца VIII века использовались монеты, выполненные из драгоценных металлов. Часто содержание металла уменьшалось при сохранении объявленной стоимости и простой люд бунтовал. Едва ли читатель согласится получать стипендию рублевыми монетами. Так Ломоносов получил в награду за оду к восшествию на престол Екатерины II 2000 рублей медной монетой и имел трудности с транспортировкой (1 тыс. рублей весила около тонны). Ввод в оборот бумажных денег с пометкой о золотом эквиваленте упростил проблему, но продолжительность работы печатного станка регламентировалась…

На «катеньку» (купюру в 100 рублей) даже в 1912 году можно было накормить обедом 667 студентов Петербургского политехникума или нанять трех сельхозрабочих на год. Полное собрание сочинений М. Ю Лермонтова в одном томе (1911 г.) стоило 1 рубль. Революции, войны, кризисы, деноминации меняли покупательную способность.

Примечательно, что в 1965 году яростный патриот и президент Франции генерал Шарль де Голль отправил в США грузы долларов для обмена по объявленному курсу на французское золото, хранившееся там. Американцы были взбешены, но вынуждены были золото отдать (на этом подобные обмены закончились, «золотой стандарт» и реальная стоимость доллара стала просто дешевеющим символом – в 1914 году Генри Форд платил рабочим на конвейере 5 долларов в день и у заводских ворот стояли тысячи претендентов).

Показатели типа заработной платы в течение столетия в статистическом анализе абсолютно бессмысленны и должны приводиться к материальному эквиваленту (золоту, потребительской корзине и т. п.).

Этап 2. Расчет характеристик распределения

Поскольку вероятность появления конкретного значения x_i в выбранной совокупности X значений случайной величины неизвестна, по *принципу недостаточного основания* принимаем вероятность появления каждого из значений равной $1 / N$.

В этом предположении находим следующие характеристики.

Математическое ожидание (среднее значение)

$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,. \tag{9.1}$$

Дисперсия (мера разброса значений относительно среднего)

$$Dx = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2 \,. \tag{9.2}$$

Среднеквадратичное (стандартное) **отклонение**

$$\sigma = \sqrt{Dx} \quad . \tag{9.3}$$

Исходные данные можно преобразовать в более удобные **стандартизованные значения**

$$z_i = \frac{x_i - \mu}{\sigma} \quad . \tag{9.4}$$

с нулевым математическим ожиданием и разбросом, определяемым с высокой степенью вероятности числами из диапазона от 0 до нескольких единиц (как правило, до 3–4).

Стандартизация минимизирует погрешность последующих вычислений, уменьшает шансы на потерю значности при умножении малых или переполнение при умножении больших значений. При желании всегда можно вернуться к исходным значениям $x_i = \mu + \sigma\, z_i$.

Стандартная ошибка среднего

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \quad . \tag{9.5}$$

Асимметрия (мера несимметричности распределения относительно μ, идеальная симметрия – $Ax = 0$)

$$A_x = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3 \quad . \tag{9.6}$$

Эксцесс *Ex* (лат. *куртозис*) (мера сглаженности – остроты пика плотности распределения)

$$E_x = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma}\right)^4 \quad . \tag{9.7}$$

Для известного нормального распределения эксцесс равен 3, $Ex > 3$ свидетельствует об островершинности. Часто в статистических пакетах, по соображениям отличия от нормальности, *Ex* берут уменьшенным на 3.

Коэффициент вариации (при ненулевых μ)

$$V_x = \frac{\mu}{\sigma}. \tag{9.8}$$

Медиана *Me* делит распределение на две равновероятные половины (число попаданий в левый интервал совпадает с числом попаданий в правый). В силу дискретности эмпирических распределений достаточно упорядочить выборку и взять медиану равной *срединному значению* $x_{[N/2]+1}$ при нечетном N или полусумме $x_{[N/2]} + x_{[N/2]+1}$ при четном; например, для совокупности значений [1 2 3 4 5] медианой будет 3, для [1 2 2 7 9 13] медиана – 4,5.

Мода *Mo* соответствует значению случайной величины, дающему максимум функции плотности. Для дискретного распределения мода находится однозначно лишь в случае **унимодальных распределений** (плотность распределения имеет единственную точку максимума). Если максимум не единственный, то могут возникнуть сомнения в неоднородности выборки (алгоритм поиска моды см. ниже).

Вывод найденных характеристик сопровождаем смысловыми комментариями по тематике рассматриваемой выборки.

Этап 3. Построение эмпирического распределения

Для облегчения последующего расчета исходную последовательность $\{x_i, i = 1, 2, \ldots, N\}$ упорядочивают по возрастанию. Результат такого упорядочивания принято называть вариационным рядом.

Немного увеличенный исходный интервал $[x_{\min} - \varepsilon, x_{\max} + \varepsilon]$ делят на k подынтервалов длиной Δ, где k выбирается интуитивно, как $\sqrt{N}$ или по формуле Стерджесса

$$k = 1 + \log_2 N \equiv 1 + 3{,}322 \lg(N) \equiv 1{,}446 \ln(N) + 3{,}5, \tag{9.9}$$

и подсчитывают число элементов выборки, попавших в каждый из подынтервалов, $\{n_j, j = 1, 2, \ldots, k\}$ и соответствующие частоты $\{w_j = n_j / N, j = 1, 2, \ldots, k\}$, служащие основой для построения эмпирической плотности распределения $p_j = w_j / \Delta, j = 1, 2, \ldots, k$.

Поскольку соответствующая гистограмма (рис. 9.1) существенно зависит от выбора N и k, устанавливаем минимальный порог для значений n_j (не менее 3) и подынтервал с числом попаданий, меньшим порога, объединяем с соседним подынтервалом.

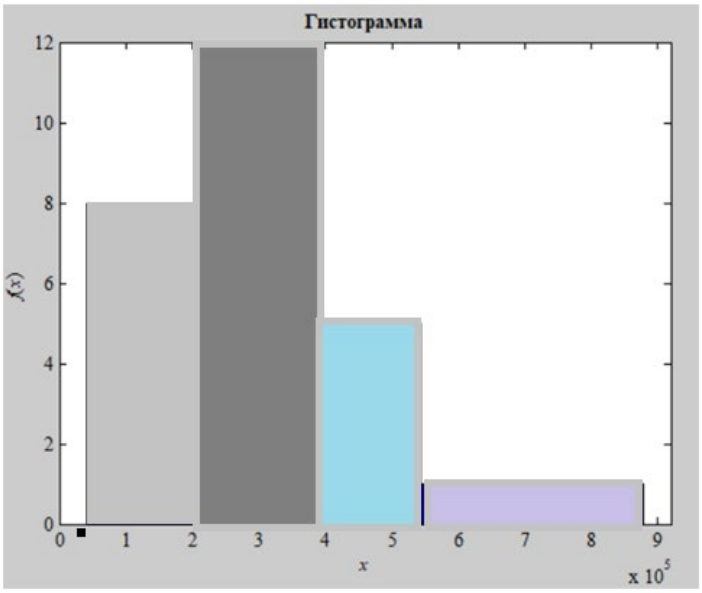


Рис. 9.1. Гистограмма эмпирического распределения

Таким образом уточняется число подынтервалов $k > 2$, их границы $[G_i, i = 1, 2, \ldots, k + 1)]$, число попаданий в подынтервалы n_i и соответствующие эмпирические вероятности $p_i^{\text{эмп}} = n_i / N$. На основе последних и строится окончательная гистограмма распределения (эмпирическая функция плотности распределения вероятностей).

На этом этапе легко найти наиболее вероятную оценку моды:

$$Mo = x_{l-1} + \Delta_l \frac{p_l - p_{l-1}}{(p_l - p_{l-1}) + (p_l - p_{l+1})}, \tag{9.10}$$

где максимум достигается на l-м подынтервале.

Этап 4. Выбор теоретического распределения

Часто с ростом N найденное эмпирическое распределение оказывается близким к какому-то из известных теоретических распределений с плотностью $p_{\text{теор}}(t)$, $t \in [\alpha, \beta]$. Для оценки близости эмпирического и теоретического распределений берется величина хи-квадрат:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - N\, p_i^{\text{теор}}\right)^2}{N\, p_i^{\text{теор}}} = N \sum_{i=1}^{k} \frac{\left(p_i^{\text{эмп}} - p_i^{\text{теор}}\right)^2}{p_i^{\text{теор}}}, \tag{9.11}$$

где

$$p_i^{\text{теор}} = \int_{G_i}^{G_{i+1}} p(t)\, dt \quad . \tag{9.12}$$

Хи-квадрат – статистика, подчиняющаяся χ^2-распределению Пирсона, зависящая от k и задающая предельное значение, при котором гипотеза о близости распределений отвергается с той или иной вероятностью ошибки α %.

Критические области для χ^2-распределения

k / α	0,050	0,010	0,005
3	7,81473	11,34487	12,83816
4	9,48773	13,27670	14,86026
5	11,07050	15,08627	16,74960
10	18,30704	23,20925	25,18818
15	24,99579	30,57791	32,80132
20	31,41043	37,56623	39,99685
25	37,65248	44,31410	46,92789
30	43,77297	46,97924	50,89218

Если при $k = 5$ эта оценка превышает 11, то гипотеза о согласованности распределений отвергается с вероятностью ошибки в 5 %.

Поиск минимума оценок для нескольких известных распределений позволяет сделать соответствующий выбор.

Из-за многообразия известных распределений при выполнении данной работы можно ограничиться 3-4 распределениями, среди которых можно выделить следующие.

Нормальное распределение (гауссовское) задается диапазоном $(-\infty, \infty)$ и определяется плотностью $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$.

Для поиска значений функции распределения $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$ в библиотеке MatLab имеется аппроксимирующая функция `F=normcdf(x,μ,σ)`, позволяющая легко найти $p_i^{\text{теор}} = \int_{G_i}^{G_{i+1}} p(t)\,dt = F(G_{i+1}) - F(G_i)$.

Равномерное распределение $p(x) = \begin{cases} 1/(b-a), & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}$.

Распределение Лапласа (двустороннее показательное)

$$p(x) = \frac{1}{\sigma\sqrt{2}} e^{-\sqrt{2}\left|\frac{x-\mu}{\sigma}\right|}, -\infty < x < \infty, F(x) = \begin{cases} \frac{1}{2} e^{\sqrt{2}\frac{x-\mu}{\sigma}}, x < \mu \\ 1 - \frac{1}{2} e^{\sqrt{2}\frac{\mu-x}{\sigma}}, x > \mu \end{cases}.$$

Логистическое распределение $p(x) = \frac{1}{k}\frac{e^{-z}}{(1+e^{-z})^2}$, $F(x) = \frac{1}{1+e^{-z}}$,

где $k = \frac{\sigma\sqrt{3}}{\pi}$, $z = \frac{x-\mu}{k}$.

Экспоненциальное распределение задается диапазоном $(0, \infty)$ и определяется плотностью $p(x) = \lambda\, e^{\lambda x}$ и функцией распределения $F(x) = 1 - e^{-\lambda x}$, $x > 0$.

Естественно, допускается рассмотрение и других распределений. Итоги анализа требуется иллюстрировать графикой (рис. 9.2).

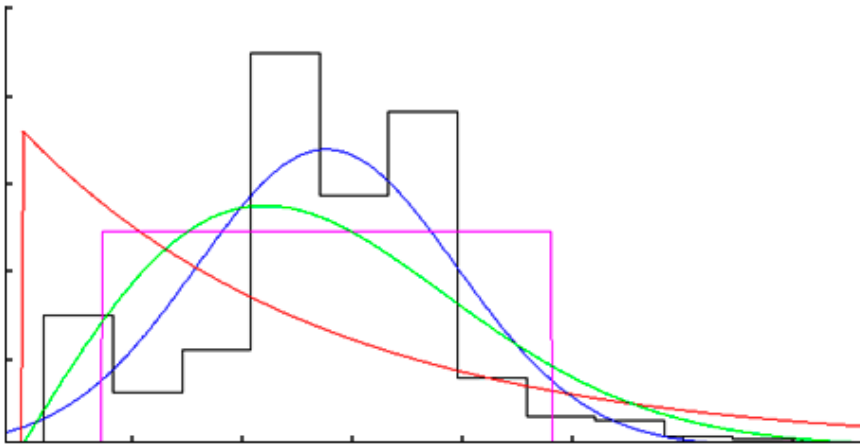


Рис. 9.2. Эмпирическая плотность в сравнении с теоретическими плотностями распределений

Отчет о выполненной работе должен содержать:

1) исходные данные и информацию об их источнике;

2) параметры выборки;

3) число интервалов, границы интервалов, число (процент) попаданий в интервалы, графическое представление эмпирического распределения;

4) рассмотренные теоретические распределения, оценки близости по Пирсону и выводы;

5) листинг программных документов.

Попробуйте сопоставить ваши выводы с интуитивными соображениями, вытекающими из самой природы анализируемой выборки.

Контрольные вопросы

1) Что характеризуют понятия *плотность* и *функция распределения*? Приведите пример их разумного использования.

2) Как найти эмпирическое математическое ожидание?

3) В чем заключается принцип недостаточности оснований?

4) Зачем ищут стандартное отклонение? О чем говорит его обращение в нуль?

5) Что представляет собой правило трех сигм?

6) Что такое медиана распределения?

7) В чем смысл понятия *мода*? Что представляет собой унимодальность?

8) Как искать моду для теоретического и эмпирического распределений?

9) Дайте определения понятий *асимметрии* и *эксцесса*. Что можно сказать о них для вашей выборки?

10) Что такое квантиль и зачем ее искать?

11) Что собой представляет критерий Пирсона и для чего он используется?

ТЕМА 2. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Этап 1. Выбор материала для анализа

Подобно работе по теме 1, исполнителю предлагается отыскать в Интернете или других источниках *реальную* статистику объема N, содержащую значения какого-либо интересного показателя и определяющих его случайных значений m факторов,

$$\left\{ y_i = x_i^0, x_i^1, x_i^2, \ldots, x_i^m,\ i = 1, 2, \ldots, N \right\}. \qquad (9.13)$$

Например, может представлять интерес вопрос о наличии и степени зависимости урожайности от температур, влажности, солнечной активности, площадей, стоимости ГСМ в долларовом эквиваленте, количества органических удобрений, удаленности от элеватора и т. п. Может быть, вам удастся обнаружить значимый фактор, которым можно управлять. Без демагогии, почему урожайность пшеницы в 1956 году в Кузбассе составляла 8–10 центнеров с гектара, а в Акмолинской области Казахстана 35–40 центнеров. Не пытайтесь искать зависимость продолжительности жизни воробья или розового фламинго в природе от природных факторов. Есть связь, но какая – лучше спросите у воробья или специалиста по орнитологии.

Значения выбранных показателя и факторов должны быть случайными, *независимыми* (по крайней мере никак не полученными друг из друга по какой-либо формуле). Нарушение этого требования иногда можно разоблачить в итоге расчета, но иногда оно может привести к иллюзорным умозаключениям и самообману.

Число факторов m должно быть не менее 3. Объем выборки N не обязательно должен быть очень большим, но не меньшим 20–30 наблюдений.

Этап 2. Базовые оценки значений факторов

Математическое ожидание (среднее значение)

$$\mu_k = \bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k,\ k = 0,1,\ldots,m. \qquad (9.14)$$

Дисперсия (мера разброса значений относительно среднего)

$$Dx_k = \sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i^k - \bar{x}_k)^2,\ k = 0,1,\ldots,m. \qquad (9.15)$$

Среднеквадратичное (стандартное) **отклонение**

$$\sigma_k = \sqrt{Dx_k},\ k = 0,1,\ldots,m. \qquad (9.16)$$

Стандартизованные значения

$$z_i^k = \frac{x_i^k - \mu_k}{\sigma_k},\ k = 0,1,\ldots,m;\ i = 1,2,\ldots,N. \qquad (9.17)$$

Матрица парных коэффициентов корреляции

$$r_{kl} = \frac{1}{N}\sum_{i=1}^{N} z_i^k z_i^l;\ k,l = 0,1,\ldots,m; \qquad (9.18)$$

$$D = \begin{bmatrix} 1 & r_{01} & r_{02} & \ldots & r_{0m} \\ r_{10} & 1 & r_{12} & \ldots & r_{1m} \\ r_{20} & r_{21} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{m0} & r_{m1} & r_{m2} & \ldots & 1 \end{bmatrix}.$$

Следует обратить внимание на факт симметричности этой матрицы.

Этап 3. Анализ коэффициентов парной корреляции

Матрица (9.18) определяет показатели *линейной* связи между факторами из диапазона от –1 до 1.

Хорошо, если межфакторные парные коэффициенты корреляции малы (это говорит о независимости влияющих факторов и надежности будущих выводов). Если их абсолютные значения близки к 1, между ними есть *неслучайная* линейная связь и один из них можно удалить за ненадобностью (например, емкость бутылки молока и ее стоимость). В других ситуациях пытаются усмотреть нелинейность связи (обратную, параболическую и др.) пары факторов визуально выводом «облака» их данных на дисплей.

Высказанные замечания касаются и первой строки матрицы, определяющей степень линейной связи между исследуемым показателем и влияющими на него факторами.

При $|r_{0k}| > 0{,}95$ возникает сомнение в пользу случайности зависимости (скорее всего, показатель попросту найден как значение некой линейной функции от k-го фактора). Если $|r_{0k}| > 0{,}99$, то это сомнение превращается в уверенность.

Относительная малость r_{0k} говорит лишь об отсутствии линейной связи, но возможна и какая-то нелинейная связь.

В случае явной нелинейности гиперболического типа $y = a + b / x_k$ замените k-й столбец исходных данных обратными значениями и повторите пересчет этапа 2.

При нелинейности параболического типа $y = a + b\,x_k + c\,x_k^2$ добавьте $(m + 1)$-й фактор значений x_k^2 и выполните пересчет этапа 2.

Указанный анализ в больших прикладных пакетах производится автоматически за счет априорного задания семейства традиционно выбираемых в эконометрике функций и задания пороговых значений $|r_{kl}|$.

Этап 4. Построение уравнения множественной регрессии

После уточнения количества и нелинейности включеня некоторых факторов в общую статистическую модель строится уравнение множественной регрессии *в стандартизованном масштабе*

$$\frac{Y - \mu_0}{\sigma_0} = \sum_{k=1}^{m} \beta_k \frac{x^k - \mu_k}{\sigma_k}, \tag{9.19}$$

которое возникает из требования минимума остаточной дисперсии

$$D_{\text{ост}} = \frac{1}{N - m - 1} \sum_{i=1}^{N} \left(Y_i - y_i\right)^2 . \tag{9.20}$$

Поиск коэффициентов β_k сводится к решению системы из m уравнений

$$\sum_{j=1}^{m} r_{jk}\, \beta_j = r_{0k};\ k = 1, 2, \ldots, m . \tag{9.21}$$

Для решения (9.21) в среде MatLab достаточно в матрице парных коэффициентов корреляции выделить фрагменты

$$A = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & \ldots & 1 \end{bmatrix} \quad ; \quad B = \begin{bmatrix} r_{10} \\ r_{20} \\ \ldots \\ r_{m0} \end{bmatrix} \tag{9.22}$$

и, например, обращением к оператору `inv(A)*B` получить вектор коэффициентов β_k. Эти коэффициенты позволяют оценить и ранжировать относительную их значимость в модели (вклад разброса данного фактора в общий разброс результирующего признака).

Легко получить уравнение множественной регрессии в естественном масштабе

$$Y = a_0 + \sum_{k=1}^{m} a_k \, x^k \,, \tag{9.23}$$

где

$$a_0 = \mu_0 - \sum_{j=1}^{m} \beta_k \frac{\sigma_0}{\sigma_k} \mu_k ; a_k = \beta_k \frac{\sigma_0}{\sigma_k} ; k = 1, 2, \ldots, m \,. \tag{9.24}$$

Уравнение (9.23) удобно использовать для прогнозирования значений итогового показателя, если известна конкретная совокупность значений влияющих факторов. Если этими факторами можно управлять, то имеет смысл постановка задачи оптимизации (9.23) при наличии ограничений.

Этап 5. Коэффициент множественной регрессии и оценки

Квадрат отношения остаточной дисперсии к исходной определяет долю дисперсии итогового показателя *Y*, объясняемую найденной зависимостью от рассматриваемых факторов, и так называемый коэффициент множественной регрессии *R* (есть и синонимы этого названия):

$$R = \sqrt{1 - \frac{D_{\text{ост}}}{D_{\text{исх}}}} \,. \tag{9.25}$$

Величину $D = \sqrt{1 - R^2}$ называют коэффициентом *детерминации* и ее квадрат определяет долю неучтенной в (9.19) дисперсии (при $R = 0{,}8$ остается 60 % необъясненного разброса значений *Y*).

При наличии нелинейных включений в уравнении регрессии вместо термина *коэффициент корреляции* используют термин *корреляционное отношение*.

Ситуация $R = 1$ при случайном характере выборки неправдоподобна. Все другие значения зависят от объема выборки и числа факторов и не позволяют категорически утверждать наличие или отсутствие корреляционной зависимости и надежности этого вывода.

Для проверки гипотезы об отсутствии связи используется критерий Стьюдента, основанный на статистике:

$$t = \frac{R\sqrt{N - m - 1}}{\sqrt{1 - R^2}} \tag{9.26}$$

с $f = N - m - 1$ степенями свободы. Если эмпирическая оценка (9.26) не превышает двустороннюю критическую для уровня значимости $\alpha / 2$, где α – вероятность ошибочного суждения, то нет оснований отвергать гипотезу об отсутствии взаимосвязи*.

Статистика t дает и оценки надежности коэффициентов в (9.23–9.24). Доверительный интервал для коэффициентов a_k

$$\Delta a_k = \pm t \sqrt{\frac{D_{\text{ост}}}{N \sigma_k^2}} . \qquad (9.27)$$

Отчет о выполненной работе должен содержать:

1) исходные данные и информацию об их источнике;
2) параметры выборки;
3) базовые оценки факторов;
4) матрицу парных коэффициентов корреляции;
5) анализ элементов матрицы с графическим представлением для суждений о возможной нелинейности связи;
6) уравнение множественной регрессии в стандартизованном и естественном масштабах;
7) ранжирование влияющих факторов;
8) множественный коэффициент корреляции, оценку по Стьюденту и выводы;
9) листинг программных документов.

Представленная информация должна сопровождаться комментариями, свидетельствующими о ее понимании автором.

Попробуйте сопоставить ваши выводы с интуитивными соображениями, вытекающими из самой природы анализируемой выборки.

Контрольные вопросы

1) Как понимать термин *корреляция*?
2) Как и зачем строится матрица парных коэффициентов корреляции?
3) Поясните формы уравнений регрессии, смысл их коэффициентов и использование.
4) Как вычисляется остаточная дисперсия?
5) Что отражают множественный коэффициент корреляции и коэффициент детерминации?
6) В чем разница между понятиями *коэффициент корреляции* и *корреляционное отношение*?

*На практике выход за критический уровень трактуется более категорично как объявление о наличии связи.

7) В чем смысл критерия Стьюдента, для чего он нужен?
8) Почему случай $N \approx m$ в корреляционном анализе неприемлем?
9) О чем свидетельствует получение $R > 1$?

ТЕМА 3. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

Этап 1. Выбор материала для анализа

Предлагается исполнителю самому отыскать в Интернете или в других источниках реальную статистику наблюдений над определенным явлением, характер которого меняется во «времени», порождая последовательность значений некоторой случайной величины $u_1 = u(t_1)$, $u_2 = u(t_2)$, ..., $u_n = u(t_N)$, называемую *временным* или *динамическим рядом*. В роли параметра t может выступать не только момент или отрезок времени, но и высота от уровня Мирового океана, глубина скважины и т. п. Следует при выборе объекта анализа осторожно выбирать динамику стоимостных характеристик и некоторых других социологических показателей.

Для упрощения расчета шаг по t желательно брать постоянным в соответствии с порядковым номером. Объем выборки N не должен быть меньшим 20–30 наблюдений.

Этап 2. Поиск тренда и сглаживание временного ряда

Под трендом временного ряда понимают некое устойчивое изменение в течение долгого времени.

Для оценки случайности отклонений от тренда (в частности, наличия циклов):

1) отыскиваем ожидаемое количество экстремальных точек («пики» и «ямы») $N_{\text{экстр}} = \frac{2}{3}(N-2)$ и соответствующее стандартное отклонение $\sqrt{(16N-29)/90}$;

2) вылавливаем во временном ряду такие точки (концы ряда не могут служить таковыми) и фиксируем их количество $N_{\text{факт}}$;

3) находим ожидаемые «фазы» длины d (фаза – интервал между двумя равнозначными экстремальными точками)

$$N_d = \frac{2(n-d-2)(d^2+3d+1)}{(d+3)!}, (d = 1, 2, ..., N-3) \qquad (9.28)$$

(так при $N = 10$ значения N_d равны 2,6667; 2,9167; 1,1000; 0,2639; 0,0460; 0,0061; 0,0006; 0,00004);

4) обращаясь к рассматриваемому ряду, оцениваем указанные фазы по восходящей и нисходящей (длина принимается без учета крайних точек).

Для минимизации влияния «белого шума» (случайных отклонений) на оценки искомых тренда и циклов выполняют процедуру *сглаживания* – локального усреднения данных, взаимного погашения случайных компонент.

Для реализации сглаживания воплощается *идея скользящего среднего*. Во временном ряде последовательно, начиная с первого элемента, выбирается окно длины $2m + 1$ (обычно $m = 1, 2$ или 3) в порядке

$$u_{-m}, u_{-m+1}, \ldots, u_{-1}, u_0, u_1, \ldots, u_{m-1}, u_m \qquad (9.29)$$

и для центрального его узла устанавливается значение, получаемое из значений ряда в выбранном окне. Крайние слева и справа m элементов временного ряда корректуре не подвергаются. Чем длина окна больше, тем сильнее сглаживание.

В примитивном варианте *линейного сглаживания* можно взять

$$U_0 = \frac{1}{2m+1} \sum_{t=-m}^{m} u_t \, . \qquad (9.30)$$

Так для ряда (1 4 10 16 25) при $m = 1$ первое сглаживание дает $U_2 = (1 + 4 + 10) / 3 = 5$; $U_3 = (5 + 10 + 16) / 3 = 10{,}3$;
$U_4 = (10{,}3 + 16 + 25) / 3 = 17{,}1$. Первый и последний элементы остаются неизменными. В итоге получаем сглаженный ряд (1 5 10,3 17,1 25).

Подобное сглаживание может выполняться неоднократно, неизменность концевых значений сгладит все нелинейности, но едва ли полученную закономерность можно экстраполировать на будущее ряда.

Иногда используется *медианное сглаживание*: значения элементов в окне выстраиваются по возрастанию (убыванию) и медиана этих значений принимается как новое значение центрального элемента окна.

Другой способ (*полиномиальное сглаживание*) связан с обычной среднеквадратической аппроксимацией полиномом степени p, коэффициенты которого можно найти из системы

$$\frac{\partial \left(\sum_{i=-m}^{m} a_0 + a_1 t_i + a_2 t_i^2 + \ldots + a_p t_i^p - u_i \right)^2}{\partial a_k} = 0; k = 0, 1, 2, \ldots, p \, . \qquad (9.31)$$

Например, при $m = 2$, $p = 2$

$$U_0 \equiv a_0 = \frac{1}{35}\left[-3u_{-2} + 12u_{-1} + 17u_0 + u_1 - 3u_2\right], \qquad (9.32)$$

а при $m = 3$ и $p = 3$

$$U_0 \equiv a_0 = \frac{1}{21}\left[-2u_{-3} + 3u_{-2} + 6u_{-1} + 7u_0 + 6u_1 - 3u_2 - 2u_3\right]. \qquad (9.33)$$

Недостаток метода скользящего среднего в том, что он не дает тренда для конца временного ряда, не предоставляя возможности экстраполяции в будущее.

Популярным методом прогнозирования многих временных рядов является *экспоненциальное сглаживание*. При *простом экспоненциальном сглаживании* используется сглаживание скользящим средним, в котором очередному наблюдению приписывается больший вес (большее доверие), чем совокупной оценке всех предшествующих. Формула такого сглаживания имеет вид

$$S_t = \alpha\, u_t + (1 - \alpha)\, S_{t-1};\ 0 < \alpha < 1. \qquad (9.34)$$

Результат сглаживания зависит от параметра α. Некоторые авторы предлагают (на основе практики) брать $\alpha < 0{,}30$. Другие предлагают осуществить подбор сравнением u_t и S_t для некоторого t или с помощью других оценок.

Выполните однократную процедуру

1) линейного сглаживания (9.30) при $m = 1$;
2) сглаживания полиномом второй степени (9.32);
3) экспоненциального сглаживания (9.34) при $\alpha = 0{,}30$.

Выведите на дисплей исходный временной ряд и результаты вышеуказанных сглаживаний.

Можете использовать процедуру аппроксимации массива значений функции алгебраическим многочленом k-й степени `P=polifit(X,Y,k)`, где `X` массив значений от 1 до n, `Y` – составляющие временного ряда. Найдите аппроксимации при значениях k, изменяющихся от 1 до 4.

Используя процедуру `F=polival(P,X)`, найдите соответствующие многочлены и выведите полученные значения и на дисплей на фоне исходного ряда. Оцените тренд временного ряда и свое подозрение (утверждение) о наличии сезонных колебаний.

Отчет о выполненной работе должен содержать:

1) исходные данные и информацию об их источнике;

2) оценки, полученные на этапе 2, и допустимые выводы;

3) таблицу итогов всех использованных сглаживаний и соответствующие графические иллюстрации;

4) итоги полиномиальной аппроксимации (уравнения и иллюстрации) и выводы о характере тренда и наличии сезонных колебаний;

5) листинг программных документов.

Попробуйте сопоставить ваши выводы с интуитивными соображениями, вытекающими из самой природы анализируемой выборки.

Контрольные вопросы

1) Дайте определение временного ряда. Назовите две основные цели анализа временных рядов.

2) Что понимается под трендом временного ряда?

3) Могут ли концы временного ряда служить экстремальными точками?

4) Какой тренд называется монотонным?

5) Для чего применяется процедура сглаживания при выделении тренда?

6) Каким образом осуществляется сглаживание методом скользящего среднего?

7) Какова процедура линейного сглаживания?

8) В чем заключаются выгоды и недостатки сглаживания полиномом степени p?

9) Что собой представляет квадратическое сглаживание?

10) Что такое медианное сглаживание?

11) В чем заключается недостаток метода скользящего среднего?

12) К чему приводит неоднократное применение методов сглаживания?

13) В чем суть метода простого экспоненциального сглаживания?

ПРИЛОЖЕНИЕ 1. Творцы методов статистического анализа (ВОЗДАЙТЕ КАЖДОМУ ПО ЗАСЛУГАМ)

Бернулли (Bernoulli*)* Я́коб (1655–1705) – швейцарский математик, один из основателей теории вероятностей и математического анализа. Доказал частный случай закона больших чисел – теорему Бернулли. Иностранный член Парижской (1699 г.) и Берлинской АН (1702 г.).

Галилей Галилео (1564–1642) – итальянский физик, механик, астроном, философ, математик. Создал теорию ошибок измерения, разделив их на систематические, обусловленные методами и средствами измерения, и случайные (непредсказуемые), заложив основы статистики.

Гаусс (Gauss) Карл Фридрих (1777–1855) – немецкий математик, с 1807 г. возглавлял кафедру математики и астрономии в Геттингенском университете, иностранный член-корреспондент (1802 г.) и почетный член (1824 г.) Петербургской АН. Его труды оказали влияние на развитие алгебры, теории чисел, классической физики, астрономии, геодезии. В 1795 г. (в возрасте 18 лет) разработал знаменитый метод наименьших квадратов для обработки неравноценных наблюдательных данных.

Гюйгенс Христиан (1629–1695) – нидерландский механик, физик, математик, астроном и изобретатель. Один из основоположников теоретической механики и теории вероятностей. В книге «О расчетах при азартных играх» ввел понятие «математическое ожидание» на примере переменной с конечным числом значений.

Кардано Джероламо (1501–1576) – итальянский математик, автор первой книги о случайности «Книга об азартных играх», первым стал присваивать событию с неизвестным исходом число (вероятность) в интервале от 0 до 1.

Кендалл (Kendall) Дэвид Джордж (1918–2007) – английский математик и статистик, член Лондонского королевского общества (1964 г.). Окончил Оксфордский университет, в 1962 – 1985 гг. профессор Кембриджского университета. Основные труды по теории вероятностей и математической статистике.

Кетле Адольф (1796–1874) – бельгийский математик, применил нормальное распределение в различных сферах жизни (число умерших, родившихся, преступлений, рост мужчин).

Колмогоров Андрей Николаевич (1903–1987) – математик, методист, академик АН СССР (1939 г.), действительный член АПН СССР (1968 г.), с 1931 г. профессор МГУ. Основатель научной школы по теории вероятностей и теории функций. В начале 60-х организовал физико-математическую школу-интернат при МГУ для математически одарен-

ных учащихся и до конца жизни читал в ней лекции не только по математике, но и по истории музыки, искусства, литературы. Организовывал летние математические школы. Инициатор создания специализированного физико-математического журнала «Квант» (1970 г.).

Лаплас (Laplace) Пьер Симон (1749–1827) – французский астроном, математик и физик, иностранный почетный член (1802 г.) Петербургской АН. Создал математическую теорию вероятностей, систематизировав выводы Б. Паскаля, П. Ферма, Я. Бернулли; доказал (1812 г.) теорему Лапласа, развил теорию ошибок, опубликовал «Опыт философии теории вероятностей» (1824 г.).

Линник Юрий Владимирович (1914–1972) – советский математик, академик АН СССР (1964 г.), с 1942 г. работал в Ленинградском отделении Математического института им. В. А. Стеклова АН СССР. В теории вероятностей и математической статистике – предельные теоремы для независимых величин и цепей Маркова, проверка сложных гипотез и теория оценивания.

Ляпунов Александр Михайлович (1857–1918) – русский математик и механик, академик Петербургской АН (1901 г.). Автор современной теории устойчивости. В теории вероятностей доказал центральную предельную теорему.

Марков Андрей Андреевич (1856–1922) – русский математик, академик Петербургской АН (1890 г.). В теории вероятностей дал полное доказательство центральной предельной теоремы, дал вероятностное обоснование метода наименьших квадратов, положил начало теории марковских процессов. Автор блестящего учебника «Исчисление вероятностей» (1900 г.).

Муавр Абрахам де (1667–1754) – французский математик, после отмены Нантского эдикта, поставившей гугенотов вне закона, переехал в Англию. Установил приближенную связь биномиального распределения с нормальным.

Нейман (Neyman) Ежи (1894–1981) – американский математик и статистик, член АН США (1963 г.), профессор Калифорнийского университета, автор трудов по теории статистических выводов.

Паскаль (Pascal) Блез (1623–1662) – французский математик, механик, физик, литератор и философ. Один из основателей математического анализа, теории вероятностей.

Пирсон (Pearson) Чарлз (1857–1936) – английский математик, биолог, философ, член Лондонского королевского общества (1896 г.), профессор Лондонского университета (1884 г.). Труды по математической статистике (кривые Пирсона, распределение Пирсона, создание теории корреляции, статистических тестов и критериев согласия).

Пирсон (Pearson) Эгон Шарп (1895–1980) – английский математик, член Лондонского королевского общества (1966 г.), профессор Лондонского университета (1933 г.). Совместно с Е. Нейманом создал общую теорию проверки статистических гипотез, занимался статистическим контролем качества массовой продукции, составитель статистических таблиц.

Пойя Полиа (Polya) Дьердь (1887–1985) – американский математик, окончил Будапештский университет (1912 г.), в 1918-1940 гг. работал в Цюрихской высшей технической школе. В статистике – распределение Пойа. Автор изумительных книг «Как решать задачу» (Москва, 1959 г.), «Математика и правдоподобные рассуждения» (Москва, 1975 г.), «Математическое открытие» (Москва, 1976 г.).

Пуассон (Poisson) Симеон Дени (1781–1840) – французский механик, физик, математик, иностранный почетный член (1826 г.) Петербургской АН, член Парижской АН (1812 г.). В теории вероятностей доказал частный случай закона больших чисел (теорема Пуассона, распределение Пуассона).

Слуцкий Евгений Евгеньевич (1880–1948) – советский математик, статистик и экономист. Учился в Киевском университете, в Мюнхенском политехникуме. С 1926 г. работал в ЦСУ, с 1934 г. – в МГУ, с 1938 г. – в Математическом институте им. В. А. Стеклова АН СССР. Один из создателей теории случайных функций. Занимался оценкой параметров по рядам связанных наблюдений.

Смирнов Николай Васильевич (1900–1966) – советский математик, член-корр. АН СССР, с 1938 г. работал в Математическом институте им. В. А. Стеклова АН СССР. Один из создателей непараметрических методов математической статистики и теории предельных распределений порядковых статистик.

Спирмен (Spearman) Чарлз Эдуард (1863–1945) – английский психолог, с 1923 г. профессор Лондонского университета. Предложил первый метод оценки надежности психологических тестов. Пытаясь соединить психологическую теорию интеллекта с теорией физических измерений и корреляционными методами, заложил основы факторного анализа.

Стьюдент (псевдоним Уильяма Сили Госсета) (1876–1937) – английский математик и статистик. Один из основателей теории статистических оценок и проверки гипотез (критерий Стьюдента, распределение Стьюдента).

Уилкоксон (Wilcoxon) Фрэнк (1892–1965) – американский химик, статистик, разработал несколько непараметрических статистических критериев.

Феллер (Feller) Уильям (1906–1970) – американский математик, член АН США, окончил Загребский университет, профессор Принстонского университета (1950 г.), автор трудов по теории диффузионных процессов, приложений в генетике, физике и экономике.

Ферма (Pierre de Fermat) Пьер (1601–1665) – французский математик, один из создателей математического анализа, теории вероятностей, аналитической геометрии и теории чисел. Наиболее известен формулировкой Великой теоремы Ферма – самой знаменитой математической загадки всех времен.

Фишер (Fisher) Роналд Эймлер (1890–1962) – английский статистик и генетик, член Лондонского королевского общества (1929 г.), один из творцов математической статистики. Ввел понятие достаточной статистики, создал теорию точечных и интервальных оценок, методику планирования эксперимента и критерий проверки статистических гипотез.

Хинчин Александр Яковлевич (1894–1959) – советский математик, член-корр. АН СССР (1939 г.), с 1922 г. – профессор МГУ, автор ряда предельных теорем теории вероятностей, основоположник теории стационарных случайных процессов и ряда методов теории массового обслуживания.

Чебышёв Пафнутий Львович (1821–1894) – русский математик и механик, академик Петербургской АН (1856 г.). Окончил Московский университет (1846 г.), защитив магистерскую диссертацию «Опыт элементарного анализа теории вероятностей». В теории вероятностей – закон больших чисел в общей форме и ряд уточнений предельной теоремы для суммы случайных величин в форме асимптотических разложений функции распределения.

Шарлье (Charlier) Карл Вильгельм Людвиг (1862–1934) – шведский астроном, один из основоположников фрактальной космологии. Успешно применил статистические методы для изучения пространственного распределения звезд в Галактике и движений звезд в окрестностях Солнца.

Эрланг (Erlang) Агнер Краруп (1878–1929) – датский математик, статистик и инженер, в теории массового обслуживания им получена формула для расчета доли вызовов, получающих обслуживание на телефонной станции и ожидающих окончания внешних вызовов. Автор признанной в мире (1909 г.) работы «Теория вероятностей и телефонные разговоры». Распределение Эрланга часто используется в задачах телекоммуникации для моделирования входящего потока вызовов.

ПРИЛОЖЕНИЕ 2. Функция нормального распределения

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,72	0,2642	1,44	0,4251	2,16	0,4846	2,90	0,4981
0,02	0,0080	0,74	0,2703	1,46	0,4279	2,18	0,4854	2,92	0,4982
0,04	0,0160	0,76	0,2764	1,48	0,4306	2,20	0,4861	2,94	0,4984
0,06	0,0239	0,78	0,2823	1,50	0,4332	2,22	0,4868	2,96	0,4985
0,08	0,0319	0,80	0,2881	1,52	0,4357	2,24	0,4875	2,98	0,4986
0,10	0,0398	0,82	0,2939	1,54	0,4382	2,26	0,4881	3,00	0,49865
0,12	0,0478	0,84	0,2995	1,56	0,4406	2,28	0,4887	3,20	0,49931
0,14	0,0557	0,86	0,3051	1,58	0,4429	2,30	0,4893	3,40	0,49966
0,16	0,0636	0,88	0,3106	1,60	0,4452	2,32	0,4898	3,60	0,499841
0,18	0,0714	0,90	0,3159	1,62	0,4474	2,34	0,4904	3,80	0,499928
0,20	0,0793	0,92	0,3212	1,64	0,4495	2,36	0,4909	4,00	0,499968
0,22	0,0871	0,94	0,3264	1,66	0,4515	2,38	0,4913	4,50	0,499997
0,24	0,0948	0,96	0,3315	1,68	0,4535	2,40	0,4918	5,00	0,499997
0,26	0,1026	0,98	0,3365	1,70	0,4554	2,42	0,4922		
0,28	0,1103	1,00	0,3413	1,72	0,4573	2,44	0,4927		
0,30	0,1179	1,02	0,3461	1,74	0,4591	2,46	0,4931		
0,32	0,1255	1,04	0,3508	1,76	0,4608	2,48	0,4934		
0,34	0,1331	1,06	0,3554	1,78	0,4625	2,50	0,4938		
0,36	0,1406	1,08	0,3599	1,80	0,4641	2,52	0,4941		
0,38	0,1480	1,10	0,3643	1,82	0,4656	2,54	0,4945		
0,40	0,1554	1,12	0,3686	1,84	0,4671	2,56	0,4948		
0,42	0,1628	1,14	0,3729	1,86	0,4686	2,58	0,4951		
0,44	0,1700	1,16	0,3770	1,88	0,4699	2,60	0,4953		
0,46	0,1772	1,18	0,3810	1,90	0,4713	2,62	0,4956		
0,48	0,1844	1,20	0,3849	1,92	0,4726	2,64	0,4959		
0,50	0,1915	1,22	0,3883	1,94	0,4738	2,66	0,4961		
0,52	0,1985	1,24	0,3925	1,96	0,4750	2,68	0,4963		
0,54	0,2054	1,26	0,3962	1,98	0,4761	2,70	0,4965		
0,56	0,2123	1,28	0,3997	2,00	0,4772	2,72	0,4967		
0,58	0,2190	1,30	0,4032	2,02	0,4783	2,74	0,4969		
0,60	0,2257	1,32	0,4066	2,04	0,4793	2,78	0,4973		
0,62	0,2324	1,34	0,4099	2,06	0,4803	2,80	0,4974		
0,64	0,2389	1,36	0,4131	2,08	0,4812	2,82	0,4976		
0,66	0,2454	1,38	0,4162	2,10	0,4821	2,84	0,4977		
0,68	0,2517	1,40	0,4192	2,12	0,4830	2,86	0,4979		
0,70	0,2580	1,42	0,4222	2,14	0,4838	2,88	0,4980		

ПРИЛОЖЕНИЕ 3. Критические области распределения Стьюдента

$k \setminus \alpha$	0,10	0,05	0,025	0,01	0,005
1	3,077684	6,313752	12,70620	31,82052	63,65674
2	1,885618	2,919986	4,30265	6,96456	9,92484
3	1,637744	2,353363	3,18245	4,54070	5,84091
4	1,533206	2,131847	2,77645	3,74695	4,60409
5	1,475884	2,015048	2,57058	3,36493	4,03214
6	1,439756	1,943180	2,44691	3,14267	3,70743
7	1,414924	1,894579	2,36462	2,99795	3,49948
8	1,396815	1,859548	2,30600	2,89646	3,35539
9	1,383029	1,833113	2,26216	2,82144	3,24984
10	1,372184	1,812461	2,22814	2,76377	3,16927
11	1,363430	1,795885	2,20099	2,71808	3,10581
12	1,356217	1,782288	2,17881	2,68100	3,05454
13	1,350171	1,770933	2,16037	2,65031	3,01228
14	1,345030	1,761310	2,14479	2,62449	2,97684
15	1,340606	1,753050	2,13145	2,60248	2,94671
16	1,336757	1,745884	2,11991	2,58349	2,92078
17	1,333379	1,739607	2,10982	2,56693	2,89823
18	1,330391	1,734064	2,10092	2,55238	2,87844
19	1,327728	1,729133	2,09302	2,53948	2,86093
20	1,325341	1,724718	2,08596	2,52798	2,84534
21	1,323188	1,720743	2,07961	2,51765	2,83136
22	1,321237	1,717144	2,07387	2,50832	2,81876
23	1,319460	1,713872	2,06866	2,49987	2,80734
24	1,317836	1,710882	2,06390	2,49216	2,79694
25	1,316345	1,708141	2,05954	2,48511	2,78744
26	1,314972	1,705618	2,05553	2,47863	2,77871
27	1,313703	1,703288	2,05183	2,47266	2,77068
28	1,312527	1,701131	2,04841	2,46714	2,76326
29	1,311434	1,699127	2,04523	2,46202	2,75639
30	1,310415	1,697261	2,04227	2,45726	2,75000
Inf	1,281552	1,644854	1,95996	2,32635	2,57583

$$p(x) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

ПРИЛОЖЕНИЕ 4. Критические области распределения Пирсона

$f \setminus \alpha$	0,100	0,050	0,025	0,010	0,005
1	2,70554	3,84146	5,02389	6,63490	7,87944
2	4,60517	5,99146	7,37776	9,21034	10,59663
3	6,25139	7,81473	9,34840	11,34487	12,83816
4	7,77944	9,48773	11,14329	13,27670	14,86026
5	9,23636	11,07050	12,83250	15,08627	16,74960
6	10,64464	12,59159	14,44938	16,81189	18,54758
7	12,01704	14,06714	16,01276	18,47531	20,27774
8	13,36157	15,50731	17,53455	20,09024	21,95495
9	14,68366	16,91898	19,02277	21,66599	23,58935
10	15,98718	18,30704	20,48318	23,20925	25,18818
11	17,27501	19,67514	21,92005	24,72497	26,75685
12	18,54935	21,02607	23,33666	26,21697	28,29952
13	19,81193	22,36203	24,73560	27,68825	29,81947
14	21,06414	23,68479	26,11895	29,14124	31,31935
15	22,30713	24,99579	27,48839	30,57791	32,80132
16	23,54183	26,29623	28,84535	31,99993	34,26719
17	24,76904	27,58711	30,19101	33,40866	35,71847
18	25,98942	28,86930	31,52638	34,80531	37,15645
19	27,20357	30,14353	32,85233	36,19087	38,58226
20	28,41198	31,41043	34,16961	37,56623	39,99685
21	29,61509	32,67057	35,47888	38,93217	41,40106
22	30,81328	33,92444	36,78071	40,28936	42,79565
23	32,00690	35,17246	38,07563	41,63840	44,18128
24	33,19624	36,41503	39,36408	42,97982	45,55851
25	34,38159	37,65248	40,64647	44,31410	46,92789
26	35,56317	38,88514	41,92317	45,64168	48,28988
27	36,74122	40,11327	43,19451	46,96294	49,64492
28	37,91592	41,33714	44,46079	48,27824	50,99338
29	39,08747	42,55697	45,72229	49,58788	52,33562
30	40,25602	43,77297	46,97924	50,89218	53,67196

$$p(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$$

ПРИЛОЖЕНИЕ 5. Критические области распределения Фишера при $\alpha = 0{,}05$

$f_2 \setminus f_1$	1	2	3	5	10	15	20	30	120	inf
1	161,4	199,500	215,707	230,162	241,882	245,950	248,013	250,095	253,2	254,3
2	18,51	19,0000	19,1643	19,2964	19,3959	19,4291	19,4458	19,4624	19,48	19,49
3	10,12	9,5521	9,2766	9,0135	8,7855	8,7029	8,6602	8,6166	8,549	8,526
5	6,607	5,7861	5,4095	5,0503	4,7351	4,6188	4,5581	4,4957	4,398	4,365
10	4,964	4,1028	3,7083	3,3258	2,9782	2,8450	2,7740	2,6996	2,580	2,537
15	4,543	3,6823	3,2874	2,9013	2,5437	2,4034	2,3275	2,2468	2,114	2,065
16	4,494	3,6337	3,2389	2,8524	2,4935	2,3522	2,2756	2,1938	2,058	2,009
17	4,451	3,5915	3,1968	2,8100	2,4499	2,3077	2,2304	2,1477	2,010	1,960
18	4,413	3,5546	3,1599	2,7729	2,4117	2,2686	2,1906	2,1071	1,968	1,916
19	4,380	3,5219	3,1274	2,7401	2,3779	2,2341	2,1555	2,0712	1,930	1,878
20	4,351	3,4928	3,0984	2,7109	2,3479	2,2033	2,1242	2,0391	1,896	1,843
21	4,324	3,4668	3,0725	2,6848	2,3210	2,1757	2,0960	2,0102	1,865	1,811
22	4,300	3,4434	3,0491	2,6613	2,2967	2,1508	2,0707	1,9842	1,838	1,783
23	4,279	3,4221	3,0280	2,6400	2,2747	2,1282	2,0476	1,9605	1,812	1,757
24	4,259	3,4028	3,0088	2,6207	2,2547	2,1077	2,0267	1,9390	1,789	1,733
25	4,241	3,3852	2,9912	2,6030	2,2365	2,0889	2,0075	1,9192	1,7684	1,7110
26	4,225	3,369	2,9752	2,5868	2,2197	2,0716	1,9898	1,9010	1,748	**1,690**
27	4,210	3,354	2,9604	2,5719	2,2043	2,0558	1,9736	1,8842	1,730	1,671
28	4,196	3,3404	2,9467	2,5581	2,1900	2,0411	1,9586	1,8687	1,713	1,654
29	4,183	3,3277	2,9340	2,5454	2,1768	2,0275	1,9446	1,8543	1,698	1,637
30	4,170	3,3158	2,9223	2,5336	2,1646	2,0148	1,9317	1,8409	1,683	1,622
40	4,084	3,2317	2,8387	2,4495	2,0772	1,9245	1,8389	1,7444	1,576	1,508
60	4,001	3,1504	2,7581	2,3683	1,9926	1,8364	1,7480	1,6491	1,467	1,389
120	3,920	3,0718	2,6802	2,2899	1,9105	1,7505	1,6587	1,5543	1,351	1,253
inf	3,841	2,9957	2,6049	2,2141	1,8307	1,6664	1,5705	1,4591	1,221	1,000

$$p(x) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma(k_1/2)\,\Gamma(k_2/2)} \left(\frac{k_1}{k_2}\right)^{k_1/2} x^{\frac{k_1}{2}-1} \left(1 + \frac{k_1}{k_2} x\right)^{-\frac{k_1+k_2}{2}}$$

ЦИТИРОВАННАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Цейтлин, Н. А. Из опыта аналитического статистика. – Москва : Солар, 2007. – 904 с.
2. Феллер, В. Введение в теорию вероятностей и ее приложения. Т. 1, 2. – Москва : Мир, 1984. Т. 1. – 528 с. Т. 2. – 738 с.
3. Корн, Г. Справочник по математике для научных работников и инженеров / Г. Корн, Т. Корн. – Москва : Наука, 1984. – 832 с.
4. Кендалл, М. Дж. Многомерный статистический анализ и временные ряды / М. Дж. Кендалл, А. Стьюарт. – Москва : Наука, 1976. – 736 с.
5. Поллард, Дж. Справочник по вычислительным методам статистики. – Москва : Финансы и статистика, 1982. – 344 с.
6. Химмельблау, Д. Анализ процессов статистическими методами. – Москва : Мир, 1973. – 957 с.
7. Афифи, А. Статистический анализ (подход с использованием ЭВМ) / А. Афифи, С. Эйзен. – Москва : Мир, 1982. – 488 с.
8. Математическая энциклопедия. Т. 5. – Москва : Советская энциклопедия, 1985. – 1246 с.
9. Андронов, А. М. Теория вероятностей и математическая статистика : учебник для вузов / А. М. Андронов, Е. А. Копытов, Л. Я. Гринглаз. – Санкт-Петербург : Питер, 2004. – 461 с.
10. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников. – Москва : Физматлит, 2006. – 816 с.
11. Валландер, С. С. Лекции по статистике и эконометрике. – Санкт-Петербург : Издательство Европейского университета в Санкт-Петербурге, 2005. – 248 с.
12. Моргенштерн, О. О точности экономико-статистических наблюдений. – Москва : Экономика, 1968. – 293 с.
13. Коуден, Д. Статистические методы контроля качества. – Москва : Физматгиз, 1961. – 623 с.
14. Алгоритмы и программы восстановления зависимостей, под ред. В. Н. Ванника. – Москва : Наука, 1984. – 816 с.
15. Иглин, С. П. Теория вероятностей и математическая статистика на базе MATLAB : учебное пособие. – Харьков : НТУ «ХПИ», 2006. – 612 с.
16. Плис, А. И. MathCad 2000. Математический практикум для экономистов и инженеров / А. И. Плис, Н. А. Сливина. – Москва : Финансы и статистика, 2000. – 656 с.

17. Тюрин, Ю. Н. Анализ данных на компьютере : учебное пособие / Ю. Н. Тюрин, А. А. Макаров. – Москва : МЦНМО, 216. – 368 с.

18. Чен, К. MatLab в математических исследованиях / К. Чен, П. Джиблин, А. Ирвинг. – Москва : Мир, 2001. – 346 с.

19. Вальд, А. Последовательный анализ. – Москва : Физматгиз, 1960. – 328 с.

20. Ван дер Варден, Б. Л. Математическая статистика. – Москва : Книга по требованию, 2012. – 435 с.

21. Крамер, Г. Математические методы статистики. – Москва : Мир, 1975. – 648 с.

22. Козлов, А. Статистический анализ данных в MS Excel : учебное пособие. – Москва : ИНФРА-М, 2012. – 312 с.

23. Тынкевич, М. А. Информационная система статистической обработки экономических данных (СТЭК) / М. А. Тынкевич, О. С. Болотова, Е. И. Латышева // Вестник КузГТУ. – 2004. – № 4. – С. 120–125.

24. Кирина, Ю. С. Развитие информационной системы для статистического анализа экономических данных / Ю. С. Кирина, М. А. Тынкевич // Сборник лучших докладов студентов и аспирантов КузГТУ по результатам 51-й студенческой научно-практической конференции, 17–21 апреля 2006. – Кемерово, 2006. – С. 256–257.

25. Тынкевич, М. А. Введение в численный анализ : учебное пособие / М. А. Тынкевич, А. Г. Пимонов. – Кемерово, 2017. – 176 с.

26. Горбунов, В. М. Практикум по дисциплине «Теория принятия решений» / В. М. Горбунов, Е. А. Синюкова. – Томск, 2014. – 125 с.

27. Корбалан, Ф. Укрощение случайности. Теория вероятностей / Ф. Корбалан, Х. Санц. – Москва : DeAgostini, 2014. – 150 с.

28. Боровиков, В. П. Популярное введение в современный анализ данных и машинное обучение на STATISTICA. – Москва : Горячая линия – Телеком, 2018. – 354 с.

29. Электронный учебник по статистике – Начальная страница [Электронный ресурс]. – Режим доступа: http://statsoft.ru/home/textbook/default.htm, свободный.

30. НОУ ИНТУИТ. Лекция. Статистическая обработка данных [Электронный ресурс]. – Режим доступа: https://intuit.ru/studies/courses/3632/874/lecture/14309?page=4, свободный.

31. Гренандер, У. Лекции по теории образов. В 3 т. – Москва : Мир. Т. 1, 1979. – 383 с. Т. 2, 1981. – 446 с. Т. 3, 1983. – 430 с.

Тынкевич Моисей Аронович
Пимонов Александр Григорьевич
Славолюбова Ярославна Викторовна

Введение
в статистический анализ данных
(теория и практика)

Учебное пособие

Редактор З. М. Савина

Подписано в печать 03.08.2021. Формат 60×84/16
Бумага офсетная. Гарнитура «Times New Roman». Уч.-изд. л. 12,0
Тираж 500 экз. Заказ № 32

Кузбасский государственный технический университет
имени Т. Ф. Горбачева
650000, Кемерово, ул. Весенняя, 28

Издательский центр Кузбасского государственного технического
университета имени Т. Ф. Горбачева
650000, Кемерово, ул. Д. Бедного, 4а

Тынкевич Моисей Аронович – (28.02.1937–31.10.2020) родился 28 февраля 1937 г. в г. Новосибирске, кандидат физико-математических наук, доцент, до последних дней жизни работал профессором кафедры прикладных информационных технологий. В 1959 г. окончил механико-математический факультет Томского государственного университета в группе двадцати четырех первых за Уралом выпускников по новой специальности «Вычислительная математика». В 1959–1966 гг. работал на кафедре вычислительной математики Томского государственного университета. С 1966 г. работал в Кузбасском государственном техническом университете. Внес значительный вклад в становление и развитие кафедры прикладных информационных технологий. Подготовил более 80 научных работ и учебных пособий, несколько циклов методических разработок. Почетный работник высшего профессионального образования Российской Федерации. Вел занятия по дисциплинам «Статистический анализ данных», «Численные методы», «Исследование операций и методы оптимизации».

Пимонов Александр Григорьевич – родился 23 ноября 1959 г. в селе Чапаево Хакасской автономной области, доктор технических наук, профессор, заведующий кафедрой прикладных информационных технологий. В 1981 г. с отличием окончил факультет прикладной математики и кибернетики Томского государственного университета. В Кузбасском государственном техническом университете работает с 1985 г. Подготовил более 230 научных работ и учебно-методических разработок. Почетный работник высшего профессионального образования Российской Федерации. Научный руководитель магистерской программы по направлению подготовки «Прикладная информатика» и программы подготовки аспирантов по направлению «Информатика и вычислительная техника». Ведет занятия по дисциплинам «Статистический анализ результатов вычислительных экспериментов», «Теория систем и системный анализ», «Математические методы и модели поддержки принятия решений», «Математическое и имитационное моделирование».

Славолюбова Ярославна Викторовна – родилась 18 июля 1984 г. в г. Кемерово, кандидат физико-математических наук, доцент, доцент кафедры прикладных информационных технологий. В 2005 г. с отличием окончила математический факультет Кемеровского государственного университета по специальности «Математика». В 2009 г. окончила аспирантуру по специальности «Математическое моделирование, численные методы и комплексы программ». С 2019 г. работает в Кузбасском государственном техническом университете. Подготовила более 80 научных работ и учебно-методических разработок. Ведет занятия по дисциплинам «Системы статистического анализа данных», «Математическое моделирование», «Информатика».

300 ЛЕТ КУЗБАСС

ISBN 978-5-00137-246-2