Серия: **СТАТИСТИЧЕСКИЕ МЕТОДЫ**

А.И.Орлов

МАТЕМАТИКА СЛУЧАЯ

Вероятность и статистика – основные факты

Учебное пособие

МЗ-Пресс

Москва 2004

Серия «Статистические методы»

Рецензенты:
Доктор физико-математических наук, профессор Я.Ю.Никитин
Кафедра «Анализ стохастических процессов в экономике»
Российской экономической академии им. Г.В.Плеханова

Орлов А.И.

Математика случая: Вероятность и статистика – основные факты: Учебное пособие / А.И.Орлов. – М.: МЗ-Пресс, 2004. – 110 с.

Сжато, но строго рассмотрены вероятностно-статистические основы современных статистических методов. Изложение доведено до переднего края научных исследований и практических разработок. Рассмотрены все основные понятия, используемые при применении современных статистических методов. Особое внимание уделено непараметрическим подходам, статистике нечисловых данных и другим перспективным элементам высоких статистических технологий. Учебное пособие рекомендовано Всероссийской ассоциацией статистических методов.

Для инженеров, менеджеров, экономистов, специалистов различных отраслей народного хозяйства, научных работников, студентов, слушателей, аспирантов и преподавателей, для всех, кому необходимо в сжатые сроки овладеть понятийной базой статистических методов.

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Статистика – это наука о том, как обрабатывать данные. Статистические методы основаны на вероятностных моделях. Они активно применяются в технических исследованиях, экономике, теории и практике управления (менеджмента). А также в социологии, медицине, геологии, истории и т.д. С обработкой результатов наблюдений, измерений, испытаний, опытов, анализов имеют дело специалисты во всех отраслях практической деятельности, почти во всех областях научных исследований.

Развитие наукоемких технологий, как правило, основано на применении высоких статистических технологий организации и управления производством. Особенно активно они используются в высокотехнологичных отраслях промышленности. Без вероятностно-статистических методов немыслимы оценка и анализ риска, страхование, финансовая деятельность. Инженеры, менеджеры, экономисты, социологи, врачи, психологи, историки успешно применяют интеллектуальные инструменты принятия решений, основанные на вероятности и статистике.

Статистические методы и модели и их база - теория вероятностей - активно развиваются во всем мире. Американская статистическая ассоциация насчитывает более двадцати тысяч членов, Королевское статистическое общество – более десяти тысяч. Статьи по вероятности и статистике постоянно публикуются более чем в пятистах научных журналах. В университетах США статистических факультетов больше, чем математических и физических. Шесть нобелевских премий получены эконометриками (специалистами по статистическим методам в экономике).

Современная теория вероятностей основана на аксиоматике академика АН СССР А.Н.Колмогорова. Однако в нашей стране специалисты и научные работники, студенты и преподаватели пока

еще недостаточно знакомы с последними достижениями в области вероятностно-статистических методов, хотя ссылки на них постоянно встречаются в научно-технической, деловой и учебной литературе.

Цель этой книги – кратко, но на современном научном уровне рассказать об основных вероятностно-статистических понятиях и фактах. Те, кто еще не знаком с этой ведущей областью современной науки, смогут быстро добраться до переднего фронта исследований. Те же, кто уже изучал вводные курсы теории вероятностей и математической статистики, быстро восстановят свои знания и расширят их до уровня, позволяющего квалифицированно использовать статистические методы в своей научной и практической работе. В частности, применять профессиональные статистические программные продукты, нормативно-техническую и инструктивно-методическую документацию,

Кому нужна эта книга?

Специалисту. В своей профессиональной деятельности инженеру, менеджеру, экономисту, научному работнику, практически любому специалисту приходится сталкиваться с необходимостью осознанно и квалифицированно применять методы, основанные на теории вероятностей и статистике. Но почти у всех при столкновении с такими методами возникают проблемы. Очень просто их описать – термины и подходы плохо понятны. Но освоить надо.

Когда-то давно, в вузовском курсе высшей математики, разбирались основы теории вероятностей и математической статистики. Казалось бы, надо взять учебники и изучить заново. Но эти книги - такие толстые. И к тому же в них нет многих понятий и концепций, нужных для практического использования

вероятностно-статистических методов. Ведь вузовский курс – это только введение в предмет.

Поэтому необходима книга, позволяющая быстро выйти на современный уровень развития статистических методов, достаточно краткая, но содержащая разбор всех необходимых понятий. Она перед Вами.

Студенту. В специальных дисциплинах часто используются вероятностно-статистические методы и модели. Значит, надо уметь в них разобраться. То, что было сдано годы назад, уже забыто, да и недостаточно для решения новых задач.

Не стоит искать старые конспекты и заново читать толстые учебники. Сейчас надо быстро освежить свои знания или заново познакомиться с основными фактами теории вероятностей и статистики. Эта книга – для Вас!

Профессионалу. Вы постоянно обрабатываете данные с помощью статистических методов. Но вероятностно-статистические методы и модели – очень быстро развивающаяся область. Отслеживаете ли Вы изменения? Вы знаете, что критерий Стьюдента остался в прошлом, применять его нецелесообразно? Вам известно, какие методы надо использовать вместо критерия Стьюдента? Вы хорошо знакомы со статистикой нечисловых данных? Если Ваш ответ – «да», то эта книга для Вас слишком элементарна. Если же «нет» - познакомьтесь с современным взглядом на теорию вероятностей и статистику!

Сравнение с аналогами

Как познакомиться с терминологией незнакомой области? Естественная мысль – обратиться к энциклопедии, например, к наиболее солидной под названием «Вероятность и математическая статистика» (см. ссылку [1] в списке цитированной литературы в конце книги). Однако толщина энциклопедии впечатляет, а

большинство статей в ней доступны лишь математикам-профессионалам.

Делались попытки составлять более или менее полные сводки терминов, определений и обозначений. Например, в учебник [2] по статистическим методам в экономике (т.е. по эконометрике) нами включена такая сводка в качестве приложения. Однако получить целостное представление о необходимой для освоения учебника базовой области знания таким образом невозможно.

Конечно, аналогами являются многочисленные учебники и учебные пособия по теории вероятностей и математической статистике (как части типового курса высшей математики) и по общей теории статистики (как части экономического образования). Однако все эти издания страдают двумя недостатками. Во-первых, они содержат много информации, которая в дальнейшем не используется в практической работе (хотя и полезна при первоначальном изучении предмета). Во-вторых, в них нет необходимых сведений о современных статистических методах. Например, типовые учебники и учебные пособия по теории вероятностей и математической статистике не содержат информации о методах, которым посвящена существенная часть распространенных программных продуктов по статистическим методам, таких, как SPSS или Statistica.

Замысел книги

Первоначальный вариант книги, которую вы держите в руках, был написан с целью преодоления разрыва между типовыми курсами по теории вероятностей и математической статистике и государственными стандартами по статистическим методам управления качеством промышленной продукции. Хотя эти стандарты содержали широко распространенные методы, не существовало (и не существует) учебно-методической литературы, заполняющей разрыв между вводными курсами и практически

используемыми в технических исследованиях статистическими методами.

Похожие проблемы имеются и в других направлениях, в которых работал автор – в социально-экономической области (в экономике, менеджменте, социологии), в научных медицинских исследованиях.

Стала очевидной необходимость создания нового типа книг, предназначенных для информационной поддержки современных разработок с использованием статистических методов. Такие книги должны давать краткое, но на современном научном уровне введение в используемые в настоящее время статистические методы.

Структура книги

Книга, которую Вы держите в руках, дает такое введение. Подробное оглавление по существу представляет собой сводку основных понятий в области теории вероятностей и статистики. По ходу изложения постоянно отмечаются возможности применения рассматриваемых концепций при решении практических задач. Конкретные методы обработки данных здесь почти не разбираются. Однако дается вся необходимая база для восприятия описаний таких методов – это и есть основная задача книги.

О содержании книги исчерпывающее представление дает оглавление. В соответствии с направленностью книги доказательства теорем не приводятся. Исключением является глава 2, посвященная опытам с конечным числом исходов. В этом случае доказательства проводятся элементарно. Автор неоднократно проводил занятия для школьников и студентов по материалам этой главы.

Замечание для математиков-профессионалов. В изложении удалось обойти ряд математических сложностей. Хотя математические основы теории вероятностей предполагают

использование у-алгебр событий (измеримых множеств) и интеграла Лебега, прикладникам эти понятия вряд ли нужны, и в книге им внимания не уделяется. Точно также не акцентируется внимание на условиях справедливости Центральной Предельной Теоремы, и т.д.

Нумерация формул, теорем, примеров, рисунков, таблиц – своя в каждой главе. Список литературы содержит только процитированные в книге источники (всего же по теории вероятностей и статистике напечатано больше миллиона статей и книг). Для облегчения труда преподавателей и обучающихся приведены контрольные вопросы и задачи, а также примерные темы докладов, рефератов и исследовательских работ. В приложении дан краткий перечень основных типов постановок задач прикладной статистики, широко используемых в практической деятельности и в научных исследованиях. Обширность этого перечня показывает, что конкретным статистическим методам должны быть посвящены отдельные издания достаточно большого объема.

Включенные в книгу материалы прошли многолетнюю и всестороннюю проверку. Они использовались во многих других отечественных и зарубежных образовательных структурах, а также организациях, занимающихся научной и практической деятельностью. Автор благодарен своим многочисленным коллегам, слушателям и студентам, прежде всего различных образовательных структур Московского государственного технического университета им. Н.Э.Баумана, за полезные обсуждения. Особую благодарность хочу выразить З.А. Отарашвили за плодотворные дискуссии при подготовке настоящего издания.

С текущей научной информацией по статистическим методам можно познакомиться на сайтах автора www.antorlov.nm.ru, www.antorlov.chat.ru, www.newtech.ru/~orlov, www.antorlov.euro.ru.

Достаточно большой объем информации содержит еженедельная рассылка "Эконометрика", выпускаемая с июля 2000 г. (о ней рассказано на указанных выше сайтах). Автор искренне благодарен редактору этого электронного издания А.А. Орлову за многолетний энтузиазм по выпуску еженедельника.

В книге раскрыто представление о случае, вероятности и статистике, соответствующее общепринятому в мире. Сделана попытка довести рассказ до современного уровня научных исследований в этой области. Конечно, возможны различные точки зрения по тем или иным частным вопросам. Автор будет благодарен читателям, если они сообщат свои вопросы и замечания по адресу издательства или непосредственно автору по электронной почте E-mail: orlov@professor.ru .

1. Вероятность и статистика нужны всем

Теория вероятностей и математическая статистика – основа вероятностно-статистических методов обработки данных. А данные мы обрабатываем и анализируем прежде всего для принятия решений. Чтобы воспользоваться современным математическим аппаратом, необходимо рассматриваемые задачи выразить в терминах вероятностно-статистических моделей.

Применение конкретного вероятностно-статистического метода состоит из трех этапов:

- переход от экономической, управленческой, технологической реальности к абстрактной математико-статистической схеме, т.е. построение вероятностной модели системы управления, технологического процесса, процедуры принятия решений, в частности по результатам статистического контроля, и т.п.

- проведение расчетов и получение выводов чисто математическими средствами в рамках вероятностной модели;

- интерпретация математико-статистических выводов применительно к реальной ситуации и принятие соответствующего решения (например, о соответствии или несоответствии качества продукции установленным требованиям, необходимости наладки технологического процесса и т.п.), в частности, заключения (о доле дефектных единиц продукции в партии, о конкретном виде законов распределения контролируемых параметров технологического процесса и др.).

Математическая статистика использует понятия, методы и результаты теории вероятностей. Далее рассматриваем основные вопросы построения вероятностных моделей в экономических, управленческих, технологических и иных ситуациях. Подчеркнем, что для активного и правильного использования нормативно-

технических и инструктивно-методических документов по вероятностно-статистическим методам нужны предварительные знания. Так, необходимо знать, при каких условиях следует применять тот или иной документ, какую исходную информацию необходимо иметь для его выбора и применения, какие решения должны быть приняты по результатам обработки данных и т.д.

Примеры применения теории вероятностей и математической статистики. Рассмотрим несколько примеров, когда вероятностно-статистические модели являются хорошим инструментом для решения управленческих, производственных, экономических, народнохозяйственных задач. Так, например, в романе А.Н.Толстого «Хождение по мукам» (т.1) говорится: «мастерская дает двадцать три процента брака, этой цифры вы и держитесь, - сказал Струков Ивану Ильичу».

Как понимать эти слова в разговоре заводских менеджеров? Одна единица продукции не может быть дефектна на 23%. Она может быть либо годной, либо дефектной. Наверно, Струков имел в виду, что в партии большого объема содержится примерно 23% дефектных единиц продукции. Тогда возникает вопрос, а что значит «примерно»? Пусть из 100 проверенных единиц продукции 30 окажутся дефектными, или из 1000 – 300, или из 100000 – 30000 и т.д., надо ли обвинять Струкова во лжи?

Или другой пример. Монетка, которую используют как жребий, должна быть «симметричной». При ее бросании в среднем в половине случаев должен выпадать герб (орел), а в половине случаев – решетка (решка, цифра). Но что означает «в среднем»? Если провести много серий по 10 бросаний в каждой серии, то часто будут встречаться серии, в которых монетка 4 раза выпадает гербом. Для симметричной монеты это будет происходить в 20,5% серий. А если на 100000 бросаний окажется 40000 гербов, то можно ли считать монету симметричной? Процедура принятия решений

строится на основе теории вероятностей и математической статистики.

Пример может показаться недостаточно серьезным. Однако это не так. Жеребьевка широко используется при организации промышленных технико-экономических экспериментов. Например, при обработке результатов измерения показателя качества (момента трения) подшипников в зависимости от различных технологических факторов (влияния консервационной среды, методов подготовки подшипников перед измерением, влияния нагрузки подшипников в процессе измерения и т.п.). Допустим, необходимо сравнить качество подшипников в зависимости от результатов хранения их в разных консервационных маслах, т.е. в маслах состава *A* и *B*. При планировании такого эксперимента возникает вопрос, какие подшипники следует поместить в масло состава *A*, а какие – в масло состава *B*, но так, чтобы избежать субъективизма и обеспечить объективность принимаемого решения. Ответ на этот вопрос может быть получен с помощью жребия.

Аналогичный пример можно привести и с контролем качества любой продукции. Чтобы решить, соответствует или не соответствует контролируемая партия продукции установленным требованиям, из нее отбирается выборка. По результатам контроля выборки делается заключение о всей партии. В этом случае очень важно избежать субъективизма при формировании выборки, т.е. необходимо, чтобы каждая единица продукции в контролируемой партии имела одинаковую вероятность быть отобранной в выборку. В производственных условиях отбор единиц продукции в выборку обычно осуществляют не с помощью жребия, а по специальным таблицам случайных чисел или с помощью компьютерных датчиков случайных чисел.

Похожие проблемы обеспечения объективности сравнения возникают при сопоставлении различных схем организации

производства, оплаты труда, при проведении тендеров и конкурсов, подбора кандидатов на вакантные должности и т.п. Всюду нужна жеребьевка или подобные ей процедуры.

Пусть надо выявить наиболее сильную и вторую по силе команду при организации турнира по олимпийской системе (проигравший выбывает). Допустим, что более сильная команда всегда побеждает более слабую. Ясно, что самая сильная команда однозначно станет чемпионом. Вторая по силе команда выйдет в финал тогда и только тогда, когда до финала у нее не будет игр с будущим чемпионом. Если такая игра запланирована, то вторая по силе команда в финал не попадет. Тот, кто планирует турнир, может либо досрочно «выбить» вторую по силе команду из турнира, сведя ее в первой же встрече с лидером, либо обеспечить ей второе место, обеспечив встречи с более слабыми командами вплоть до финала. Чтобы избежать субъективизма, проводят жеребьевку. Для турнира из 8 команд вероятность того, что в финале встретятся две самые сильные команды, равна 4/7. Соответственно с вероятностью 3/7 вторая по силе команда покинет турнир досрочно.

При любом измерении единиц продукции (с помощью штангенциркуля, микрометра, амперметра и т.п.) имеются погрешности. Чтобы выяснить, есть ли систематические погрешности, необходимо сделать многократные измерения единицы продукции, характеристики которой известны (например, стандартного образца). При этом следует помнить, что кроме систематической погрешности присутствует и случайная погрешность.

Поэтому встает вопрос, как по результатам измерений узнать, есть л систематическая погрешность. Если отмечать только, является ли полученная при очередном измерении погрешность положительной или отрицательной, то эту задачу можно свести к уже рассмотренной. Действительно, сопоставим измерение с

бросанием монеты, положительную погрешность – с выпадением герба, отрицательную – решетки (нулевая погрешность при достаточном числе делений шкалы практически никогда не встречается). Тогда проверка отсутствия систематической погрешности эквивалентна проверке симметричности монеты.

Итак, задача проверки отсутствия систематической погрешности сведена к задаче проверки симметричности монеты. Проведенные рассуждения приводят к так называемому «критерию знаков» в математической статистике.

При статистическом регулировании технологических процессов на основе методов математической статистики разрабатываются правила и планы статистического контроля процессов, направленные на своевременное обнаружение разладки технологических процессов и принятия мер к их наладке и предотвращению выпуска продукции, не соответствующей установленным требованиям. Эти меры нацелены на сокращение издержек производства и потерь от поставки некачественных единиц продукции. При статистическом приемочном контроле на основе методов математической статистики разрабатываются планы контроля качества путем анализа выборок из партий продукции. Сложность заключается в том, чтобы уметь правильно строить вероятностно-статистические модели принятия решений. В математической статистике для этого разработаны вероятностные модели и методы проверки гипотез, в частности, гипотез о том, что доля дефектных единиц продукции равна определенному числу p_0, например, $p_0 = 0{,}23$ (вспомните слова Струкова из романа А.Н.Толстого).

Задачи оценивания. В ряде управленческих, производственных, экономических, народнохозяйственных ситуаций возникают задачи другого типа – задачи оценки характеристик и параметров распределений вероятностей.

Рассмотрим пример. Пусть на контроль поступила партия из N электроламп. Из этой партии случайным образом отобрана выборка объемом n электроламп. Возникает ряд естественных вопросов. Как по результатам испытаний элементов выборки определить средний срок службы электроламп, с какой точностью можно оценить эту характеристику? Как изменится точность, если взять выборку большего объема? При каком числе часов T можно гарантировать, что не менее 90% электроламп прослужат T и более часов?

Предположим, что при испытании выборки объемом n электроламп дефектными оказались X электроламп. Какие границы можно указать для числа D дефектных электроламп в партии, для уровня дефектности D/N и т.п.?

Или при статистическом анализе точности и стабильности технологических процессов надлежит оценить такие показатели качества, как среднее значение контролируемого параметра и степень его разброса в рассматриваемом процессе. Согласно теории вероятностей в качестве среднего значения случайной величины целесообразно использовать ее математическое ожидание, а в качестве статистической характеристики разброса – дисперсию, среднее квадратическое отклонение или коэффициент вариации. Возникают вопросы: как оценить эти статистические характеристики по выборочным данным, с какой точностью это удается сделать?

Аналогичных примеров можно привести очень много. Здесь важно было показать, как теория вероятностей и математическая статистика могут быть использованы в инженерных и управленческих задачах.

Современное представление о математической статистике. Под математической статистикой понимают «раздел математики, посвященный математическим методам сбора,

систематизации, обработки и интерпретации статистических данных, а также использование их для научных или практических выводов. Правила и процедуры математической статистики опираются на теорию вероятностей, позволяющую оценить точность и надежность выводов, получаемых в каждой задаче на основании имеющегося статистического материала» [1, с.326]. При этом статистическими данными называются сведения о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками.

По типу решаемых задач математическая статистика обычно делится на три раздела: описание данных, оценивание и проверка гипотез.

По виду обрабатываемых статистических данных математическая статистика делится на четыре направления:

- одномерная статистика (статистика случайных величин), в которой результат наблюдения описывается действительным числом;

- многомерный статистический анализ, где результат наблюдения над объектом описывается несколькими числами (вектором);

- статистика случайных процессов и временных рядов, где результат наблюдения – функция;

- статистика объектов нечисловой природы, в которой результат наблюдения имеет нечисловую природу, например, является множеством (геометрической фигурой), упорядочением или получен в результате измерения по качественному признаку.

Исторически первой появились некоторые области статистики объектов нечисловой природы (в частности, задачи оценивания доли брака и проверки гипотез о ней) и одномерная статистика. Математический аппарат для них проще, поэтому на их

примере обычно демонстрируют основные идеи математической статистики.

Лишь те методы обработки данных, т.е. математической статистики, являются доказательными, которые опираются на вероятностные модели соответствующих реальных явлений и процессов. Речь идет о моделях поведения потребителей, возникновения рисков, функционирования технологического оборудования, получения результатов эксперимента, течения заболевания и т.п. Вероятностную модель реального явления следует считать построенной, если рассматриваемые величины и связи между ними выражены в терминах теории вероятностей. Соответствие вероятностной модели реальности, т.е. ее адекватность, обосновывают, в частности, с помощью статистических методов проверки гипотез.

Невероятностные методы обработки данных являются поисковыми, их можно использовать лишь при предварительном анализе данных, так как они не дают возможности оценить точность и надежность выводов, полученных на основании ограниченного статистического материала.

Вероятностные и статистические методы применимы всюду, где удается построить и обосновать вероятностную модель явления или процесса. Их применение обязательно, когда сделанные на основе выборочных данных выводы переносятся на всю совокупность (например, с выборки на всю партию продукции).

В конкретных областях применений используются как вероятностно-статистические методы широкого применения, так и специфические. Например, в разделе производственного менеджмента, посвященного статистическим методам управления качеством продукции, используют прикладную математическую статистику (включая планирование экспериментов). С помощью ее методов проводится статистический анализ точности и

стабильности технологических процессов и статистическая оценка качества. К специфическим методам относятся методы статистического приемочного контроля качества продукции, статистического регулирования технологических процессов, оценки и контроля надежности и др.

Широко применяются такие прикладные вероятностно-статистические дисциплины, как теория надежности и теория массового обслуживания. Содержание первой из них ясно из названия, вторая занимается изучением систем типа телефонной станции, на которую в случайные моменты времени поступают вызовы - требования абонентов, набирающих номера на своих телефонных аппаратах. Длительность обслуживания этих требований, т.е. длительность разговоров, также моделируется случайными величинами. Большой вклад в развитие этих дисциплин внесли член-корреспондент АН СССР А.Я. Хинчин (1894-1959), академик АН УССР Б.В.Гнеденко (1912-1995) и другие отечественные ученые.

Коротко об истории математической статистики. Математическая статистика как наука начинается с работ знаменитого немецкого математика Карла Фридриха Гаусса (1777-1855), который на основе теории вероятностей исследовал и обосновал метод наименьших квадратов, созданный им в 1795 г. и примененный для обработки астрономических данных (с целью уточнения орбиты малой планеты Церера). Его именем часто называют одно из наиболее популярных распределений вероятностей – нормальное, а в теории случайных процессов основной объект изучения – гауссовские процессы.

В конце XIX в. – начале XX в. крупный вклад в математическую статистику внесли английские исследователи, прежде всего К.Пирсон (1857-1936) и Р.А.Фишер (1890-1962). В частности, Пирсон разработал критерий «хи-квадрат» проверки

статистических гипотез, а Фишер – дисперсионный анализ, теорию планирования эксперимента, метод максимального правдоподобия оценки параметров.

В 30-е годы XX в. поляк Ежи Нейман (1894-1977) и англичанин Э.Пирсон развили общую теорию проверки статистических гипотез, а советские математики академик А.Н. Колмогоров (1903-1987) и член-корреспондент АН СССР Н.В.Смирнов (1900-1966) заложили основы непараметрической статистики. В сороковые годы XX в. румын А. Вальд (1902-1950) построил теорию последовательного статистического анализа.

Математическая статистика бурно развивается и в настоящее время. Так, за последние 40 лет можно выделить четыре принципиально новых направления исследований [2]:

- разработка и внедрение математических методов планирования экспериментов;

- развитие статистики объектов нечисловой природы как самостоятельного направления в прикладной математической статистике;

- развитие статистических методов, устойчивых по отношению к малым отклонениям от используемой вероятностной модели;

- широкое развертывание работ по созданию компьютерных пакетов программ, предназначенных для проведения статистического анализа данных.

Вероятностно-статистические методы и оптимизация. Идея оптимизации пронизывает современную прикладную математическую статистику и иные статистические методы. А именно, методы планирования экспериментов, статистического приемочного контроля, статистического регулирования технологических процессов и др. С другой стороны, оптимизационные постановки в теории принятия решений,

например, прикладная теория оптимизации качества продукции и требований стандартов, предусматривают широкое использование вероятностно-статистических методов, прежде всего прикладной математической статистики.

В производственном менеджменте, в частности, при оптимизации качества продукции и требований стандартов особенно важно применять статистические методы на начальном этапе жизненного цикла продукции, т.е. на этапе научно-исследовательской подготовки опытно-конструкторских разработок (разработка перспективных требований к продукции, аванпроекта, технического задания на опытно-конструкторскую разработку). Это объясняется ограниченностью информации, доступной на начальном этапе жизненного цикла продукции, и необходимостью прогнозирования технических возможностей и экономической ситуации на будущее. Статистические методы должны применяться на всех этапах решения задачи оптимизации – при шкалировании переменных, разработке математических моделей функционирования изделий и систем, проведении технических и экономических экспериментов и т.д.

В задачах оптимизации, в том числе оптимизации качества продукции и требований стандартов, используют все области статистики. А именно, статистику случайных величин, многомерный статистический анализ, статистику случайных процессов и временных рядов, статистику объектов нечисловой природы. Разработаны рекомендации по выбору статистического метода для анализа конкретных данных [3].

2. Основы теории вероятностей

Этот раздел содержит полные доказательства всех рассматриваемых утверждений.

События и множества. Исходное понятие при построении вероятностных моделей в задачах принятия решений – опыт (испытание). Примерами опытов являются проверка качества единицы продукции, бросание трех монет независимо друг от друга и т.д.

Первый шаг при построении вероятностной модели реального явления или процесса – выделение возможных исходов опыта. Их называют элементарными событиями. Обычно считают, что в первом опыте возможны два исхода – «единица продукции годная» и «единица продукции дефектная». Естественно принять, что при бросании монеты осуществляется одно из двух элементарных событий – «выпала решетка (цифра)» и «выпал герб». Таким образом, случаи «монета встала на ребро» или «монету не удалось найти» считаем невозможными.

При бросании трех монет элементарных событий значительно больше. Вот одно из них – «первая монета выпала гербом, вторая – решеткой, третья – снова гербом». Перечислим все элементарные события в этом опыте. Для этого обозначим выпадение герба буквой Г, а решетки – буквой Р. Имеется $2^3=8$ элементарных событий: ГГГ, ГГР, ГРГ, ГРР, РГГ, РГР, РРГ, РРР – в каждой тройке символов первый показывает результат бросания первой модели, второй – второй монеты, третий – третьей монеты.

Совокупность всех возможных исходов опыта, т.е. всех элементарных событий, называется пространством элементарных событий. Вначале мы ограничимся пространством элементарных событий, состоящим из конечного числа элементов.

С математической точки зрения пространство (совокупность) всех элементарных событий, возможных в опыте – это некоторое множество, а элементарные события – его элементы. Однако в теории вероятностей для обозначения используемых понятий по традиции используются свои термины, отличающиеся от терминов

теории множеств. В табл. 1 установлено соответствие между терминологическими рядами этих двух математических дисциплин.

Таблица 1.

Соответствие терминов теории вероятностей и теории множеств

Теория вероятностей	Теория множеств
Пространство элементарных событий	Множество
Элементарное событие	Элемент этого множества
Событие	Подмножество
Достоверное событие	Подмножество, совпадающее с множеством
Невозможное событие	Пустое подмножество $\varnothing$
Сумма $A+B$ событий A и B	Объединение $A \bigcup B$
Произведение AB событий A и B	Пересечение $A \bigcap B$
Событие, противоположное A	Дополнение A
События A и B несовместны	$A \bigcap B$ пусто
События A и B совместны	$A \bigcap B$ не пусто

Как сложились два параллельных терминологических ряда? Основные понятия теории вероятностей и ее терминология сформировались в XVII-XVIII вв. Теория множеств возникла в конце XIX в. независимо от теории вероятностей и получила распространение в XX в.

Принятый в настоящее время аксиоматический подход к теории вероятностей, разработанный академиком АН СССР А.Н. Колмогоровым (1903-1987), дал возможность развивать эту дисциплину на базе теории множеств и теории меры. Этот подход позволил рассматривать теорию вероятностей и математическую статистику как часть математики, проводить рассуждения на математическом уровне строгости. В частности, было введено четкое различие между частотой и вероятностью, случайная величина стала рассматриваться как функция от элементарного исхода, и т.д. За основу методов статистического анализа данных стало возможным брать вероятностно-статистические модели, сформулированные в математических терминах. В результате удалось четко отделить строгие утверждения от обсуждения

философских вопросов случайности, преодолеть подход на основе понятия равновозможности, имеющий ограниченное практическое значение. Наиболее существенно, что после работ А.Н.Колмогорова нет необходимости связывать вероятности тех или иных событий с пределами частот. Так называемые «субъективные вероятности» получили смысл экспертных оценок вероятностей.

После выхода (в 1933 г. на немецком языке и в 1936 г. – на русском) основополагающей монографии [4] аксиоматический подход к теории вероятностей стал общепринятым в научных исследованиях в этой области. Во многом перестроилось преподавание. Повысился научный уровень многих прикладных работ. Однако традиционный подход оказался живучим. Распространены устаревшие и во многом неверные представления о теории вероятностей и математической статистике. Поэтому в настоящей главе рассматриваем основные понятия, подходы, идеи, методы и результаты в этих областях, необходимые для их квалифицированного применения в задачах различных областей знаний и практической деятельности.

В послевоенные годы А.Н.Колмогоров формализовал понятие случайности на основе теории информации [5]. Грубо говоря, числовая последовательность является случайной, если ее нельзя заметно сжать без потери информации. Однако этот подход не был предназначен для использования в прикладных работах и преподавании. Он представляет собой важное методологическое и теоретическое продвижение.

Вероятность события. Перейдем к основному понятию теории вероятностей – понятию вероятности события. В методологических терминах можно сказать, что вероятность события является мерой возможности осуществления события. В ряде случаев естественно считать, что вероятность события A – это число, к которому приближается отношение количества

осуществлений события A к общему числу всех опытов (т.е. частота осуществления события A) – при увеличении числа опытов, проводящихся независимо друг от друга. Иногда можно предсказать это число из соображений равновозможности. Так, при бросании симметричной монеты и герб, и решетка имеют одинаковые шансы оказаться сверху, а именно, 1 шанс из 2, а потому вероятности выпадения герба и решетки равны 1/2.

Однако этих соображений недостаточно для развития теории. Методологическое определение не дает численных значений. Не все вероятности можно оценивать как пределы частот, и неясно, сколько опытов надо брать. На основе идеи равновозможности можно решить ряд задач, но в большинстве практических ситуаций применить ее нельзя. Например, для оценки вероятности дефектности единицы продукции. Поэтому перейдем к определениям в рамках аксиоматического подхода на базе математической модели, предложенной А.Н.Колмогоровым (1933).

Определение 1. Пусть конечное множество $\Omega = \{\omega\}$ является пространством элементарных событий, соответствующим некоторому опыту. Пусть каждому $\omega \in \Omega$ поставлено в соответствие неотрицательное число $P(\omega)$, называемое вероятностью элементарного события ω, причем сумма вероятностей всех элементарных событий равна 1, т.е.

$$\sum_{\omega \in \Omega} P(\omega) = 1. \quad (1)$$

Тогда пара $\{\Omega, P\}$, состоящая из конечного множества Ω и неотрицательной функции P, определенной на Ω и удовлетворяющей условию (1), называется *вероятностным пространством*. Вероятность события A равна сумме вероятностей элементарных событий, входящих в А, т.е. определяется равенством

$$P(A) = \sum_{\omega \in A} P(\omega). \quad (2)$$

Сконструирован математический объект, основной при построении вероятностных моделей. Рассмотрим примеры.

Пример 1. Бросанию монеты соответствует вероятностное пространство с $\Omega = \{Г, Р\}$ и $P(Г) = P(Р) = S$; здесь обозначено: Г – выпал герб, Р – выпала решетка.

Пример 2. Проверке качества одной единицы продукции (в ситуации, описанной в романе А.Н.Толстого «Хождение по мукам» - см. выше) соответствует вероятностное пространство с $\Omega = \{Б, Г\}$ и $Р(Б) = 0{,}23$, $Р(Г) = 0{,}77$; здесь обозначено: Б - дефектная единица продукции, Г – годная единица продукции; значение вероятности 0,23 взято из слов Струкова.

Отметим, что приведенное выше определение вероятности *P(A)* согласуется с интуитивным представлением о связи вероятностей события и входящих в него элементарных событий, а также с распространенным мнением, согласно которому «вероятность события A – число от 0 до 1, которое представляет собой предел частоты реализации события A при неограниченном числе повторений одного и того же комплекса условий».

Из определения вероятности события, свойств символа суммирования и равенства (1) вытекает, что

$$а) P(\Omega) = 1,\ \ б) P(\varnothing) = 0,\ \ в) P(A + B) = P(A) + P(B) - P(AB). \quad (3)$$

Для несовместных событий A и B согласно формуле (3) $P(A+B) = P(A)+P(B)$. Последнее утверждение называют также теоремой сложения вероятностей.

Независимые события. При практическом применении вероятностно-статистических методов принятия решений постоянно используется понятие независимости. Например, при применении статистических методов управления качеством продукции говорят о независимых измерениях значений контролируемых параметров у включенных в выборку единиц

продукции, о независимости появления дефектов одного вида от появления дефектов другого вида, и т.д. Независимость случайных событий понимается в вероятностных моделях в следующем смысле.

Определение 2. События A и B называются независимыми, если $P(AB) = P(A)\,P(B)$. Несколько событий A, B, C,… называются независимыми, если вероятность их совместного осуществления равна произведению вероятностей осуществления каждого из них в отдельности: $P(ABC\ldots) = P(A)P(B)P(C)\ldots$

Это определение соответствует интуитивному представлению о независимости: осуществление или неосуществление одного события не должно влиять на осуществление или неосуществление другого. Иногда соотношение $P(AB) = P(A)\,P(B|A) = P(B)P(A|B)$, справедливое при $P(A)P(\mathrm{B}) > 0$, называют также теоремой умножения вероятностей.

Утверждение 1. Пусть события A и B независимы. Тогда события $\overline{A}$ и $\overline{B}$ независимы, события $\overline{A}$ и B независимы, события A и $\overline{B}$ независимы (здесь $\overline{A}$ - событие, противоположное A, и $\overline{B}$ - событие, противоположное B).

Действительно, из свойства в) в (3) следует, что для событий C и D, произведение которых пусто, $P(C+D) = P(C) + P(D)$. Поскольку пересечение AB и $\overline{A}\,B$ пусто, а объединение есть B, то $P(AB) + P(\overline{A}\,B) = P(B)$. Так как А и В независимы, то $P(\overline{A}\,B) = P(B) - P(AB) = P(B) - P(A)P(B) = P(B)(1 - P(A))$. Заметим теперь, что из соотношений (1) и (2) следует, что $P(\overline{A}) = 1 - P(A)$. Значит, $P(\overline{A}\,B) = P(\overline{A})P(B)$.

Вывод равенства $P(A\,\overline{B}) = P(A)P(\overline{B})$ отличается от предыдущего лишь заменой всюду A на B, а B на A.

Для доказательства независимости $\overline{A}$ и $\overline{B}$ воспользуемся тем, что события AB, $\overline{A}\,B$, $A\,\overline{B}$, $\overline{A}\;\overline{B}$ не имеют попарно общих элементов, а в сумме составляют все пространство элементарных событий.

Следовательно, $P(AB) + P(\overline{A}B) + P(A\overline{B}) + P(\overline{A}\,\overline{B}) = 1$. Воспользовавшись ранее доказанными соотношениями, получаем, что $P(\overline{A}B) = 1 - P(AB) - P(B)(1 - P(A)) - P(A)(1 - P(B)) = (1 - P(A))(1 - P(B)) = P(\overline{A})P(\overline{B})$, что и требовалось доказать.

Пример 3. Рассмотрим опыт, состоящий в бросании игрального кубика, на гранях которого написаны числа 1, 2, 3, 4, 5,6. Считаем, что все грани имеют одинаковые шансы оказаться наверху. Построим соответствующее вероятностное пространство. Покажем, что события «наверху – грань с четным номером» и «наверху – грань с числом, делящимся на 3» являются независимыми.

Разбор примера. Пространство элементарных исходов состоит из 6 элементов: «наверху – грань с 1», «наверху – грань с 2»,…, «наверху – грань с 6». Событие «наверху – грань с четным номером» состоит из трех элементарных событий – когда наверху оказывается 2, 4 или 6. Событие «наверху – грань с числом, делящимся на 3» состоит из двух элементарных событий – когда наверху оказывается 3 или 6. Поскольку все грани имеют одинаковые шансы оказаться наверху, то все элементарные события должны иметь одинаковую вероятность. Поскольку всего имеется 6 элементарных событий, то каждое из них имеет вероятность 1/6. По определению 1событие «наверху – грань с четным номером» имеет вероятность S, а событие «наверху – грань с числом, делящимся на 3» - вероятность 1/3. Произведение этих событий состоит из одного элементарного события «наверху – грань с 6», а потому имеет вероятность 1/6. Поскольку 1/6 = S x 1/3, то рассматриваемые события являются независимыми в соответствии с определением независимости.

Независимые испытания. В вероятностных моделях процедур принятия решений с помощью понятия независимости событий можно придать точный смысл понятию «независимые

испытания». Для этого рассмотрим сложный опыт, состоящий в проведении двух испытаний. Эти испытания называются независимыми, если любые два события A и B, из которых A определяется по исходу первого испытания, а B – по исходу второго, являются независимыми.

Пример 4. Опишем вероятностное пространство, соответствующее бросанию двух монет независимо друг от друга.

Разбор примера. Пространство элементарных событий состоит из четырех элементов: ГГ, ГР, РГ, РР (запись ГГ означает, что первая монета выпала гербом и вторая – тоже гербом; запись РГ – первая – решеткой, а вторая – гербом, и т.д.). Поскольку события «первая монета выпала решеткой» и «вторая монета выпала гербом» являются независимыми по определению независимых испытаний и вероятность каждого из них равна S, то вероятность РГ равна j. Аналогично вероятность каждого из остальных элементарных событий также равна j.

Пример 5. Опишем вероятностное пространство, соответствующее проверке качества двух единиц продукции независимо друг от друга, если вероятность дефектности равна x.

Разбор примера. Пространство элементарных событий состоит из четырех элементов:

ω_1 - обе единицы продукции годны;

ω_2 - первая единица продукции годна, а вторая – дефектна;

ω_3 - первая единица продукции дефектна, а вторая – годна;

ω_4 - обе единицы продукции являются дефектными.

Вероятность того, что единица продукции дефектна, есть x, а потому вероятность того, что имеет место противоположное событие, т.е. единица продукции годна, есть $1 - x$. Поскольку результат проверки первой единицы продукции не зависит от

такового для второй, то $P(\omega_1) = (1-x)^2,\ P(\omega_2) = P(\omega_3) = x(1-x),\ P(\omega_4) = x^2.$

Условные вероятности. В некоторых задачах прикладной статистики оказывается полезным такое понятие, как условная вероятность $P(B|A)$ – вероятность осуществления события B при условии, что событие A произошло. При $P(A)>0$ по определению

$$P(B \mid A) = \frac{P(AB)}{P(A)}.$$

Для независимых событий A и B, очевидно, $P(B|A)= P(B)$. Это равенство эквивалентно определению независимости. Понятия условной вероятности и независимости введены А.Муавром в 1718 г.

Необходимо иметь в виду, что для независимости в совокупности нескольких событий недостаточно их попарной независимости. Рассмотрим классический пример [6, с.46]. Пусть одна грань тетраэдра окрашена в красный цвет, вторая - в зеленый. Третья грань окрашена в синий цвет и четвертая – во все эти три цвета. Пусть событие A состоит в том, что грань, на которую упал тетраэдр при бросании, окрашена красным (полностью или частично), событие B – зеленым, событие C – синим. Пусть при бросании все четыре грани тетраэдра имеют одинаковые шансы оказаться внизу. Поскольку граней четыре и две из них имеют в окраске красный цвет, то $P(A) = 1/2$. Легко подсчитать, что

$$P(B) = P(C) = P(A|B) = P(B|C) = P(C|A) = P(B|A)$$
$$= P(C|A) = P(A|C) = \mathrm{S}.$$

События A, B и C, таким образом, попарно независимы. Однако если известно, что осуществились одновременно события B и C, то это значит, что тетраэдр встал на грань, содержащую все три цвета, т.е. осуществилось и событие A. Следовательно, $P(ABC)$ = j, в то время как для независимых событий должно быть $P(A)P(B)P(C)$ =

1/8. Следовательно, события A, B и C в совокупности зависимы, хотя попарно независимы.

Формула полной вероятности. Предположим, что событие B может осуществиться с одним и только с одним из n попарно несовместных событий $A_1, A_2,\ldots, A_k$. Тогда

$$B = \sum_{j=1}^{k} BA_j,$$

где события BA_i и BA_j с разными индексами i и j несовместны. По теореме сложения вероятностей

$$P(B) = \sum_{j=1}^{k} P(BA_j).$$

Воспользовавшись теоремой умножения, находим, что

$$P(B) = \sum_{j=1}^{k} P(A_j)P(B \mid A_j).$$

Получена т.н. «формула полной вероятности». Она широко использовалась математиками при конкретных расчетах еще в начале 18 века, но впервые была сформулирована как одно из основных утверждений теории вероятностей П.Лапласом лишь в конце этого века. Она применяется, в частности, при нахождении среднего выходного уровня дефектности в задачах статистического обеспечения качества продукции.

Формулы Байеса. Применим формулу полной вероятности для вывода т.н. «формул Байеса», которые иногда используют при проверке статистических гипотез. Требуется найти вероятность события A_i, если известно, что событие B произошло. Согласно теореме умножения

$$P(A_iB) = P(B)P(A_i|B) = P(A_i)\ P(B|A_i).$$

Следовательно,

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(B)}.$$

Используя формулу полной вероятности для знаменателя, находим, что

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j)P(B \mid A_j)}.$$

Две последние формулы и называют обычно формулами Байеса. Общая схема их использования такова. Пусть событие B может протекать в различных условиях, относительно которых может быть сделано k гипотез A_1, A_2,…, A_k. Априорные (от *a priori* (лат.) – до опыта) вероятности этих гипотез есть $P(A_1)$, $P(A_2)$,…, $P(A_k)$. Известно также, что при справедливости гипотезы A_i вероятность осуществления события B равна $P(B|A_i)$. Произведен опыт, в результате которого событие B наступило. Естественно после этого уточнить оценки вероятностей гипотез. Апостериорные (от *a posteriori* (лат.) – на основе опыта) оценки вероятностей гипотез $P(A_1|B)$, $P(A_2|B)$,…, $P(A_k|B)$ даются формулами Байеса. В прикладной статистике существует направление «байесовская статистика», в которой, в частности, на основе априорного распределения параметров после проведения измерений, наблюдений, испытаний, опытов анализов вычисляют уточненные оценки параметров.

Случайные величины. Случайная величина – это величина, значение которой зависит от случая, т.е. от элементарного события ω. Таким образом, случайная величина – это функция, определенная на пространстве элементарных событий Ω. Примеры случайных величин: количество гербов, выпавших при независимом бросании двух монет; число, выпавшее на верхней грани игрального кубика; число дефектных единиц продукции среди проверенных.

Определение случайной величины X как функции от элементарного события ω, т.е. функции $X:\Omega \to H$, отображающей пространство элементарных событий Ω в некоторое множество H, казалось бы, содержит в себе противоречие. О чем идет речь – о

величине или о функции? Дело в том, что наблюдается всегда лишь т.н. «реализация случайной величины», т.е. ее значение, соответствующее именно тому элементарному исходу опыта (элементарному событию), которое осуществилось в конкретной реальной ситуации. Т.е. наблюдается именно «величина». А функция от элементарного события – это теоретическое понятие, основа вероятностной модели реального явления или процесса.

Отметим, что элементы H – это не обязательно числа. Ими могут быть и последовательности чисел (вектора), и функции, и математические объекты иной природы, в частности, нечисловой (упорядочения и другие бинарные отношения, множества, нечеткие множества и др.) [2]. Однако наиболее часто рассматриваются вероятностные модели, в которых элементы H – числа, т.е. $H = R^1$. В иных случаях обычно используют термины «случайный вектор», «случайное множество», «случайное упорядочение», «случайный элемент» и др.

Математическое ожидание. Рассмотрим случайную величину с числовыми значениями. Часто оказывается полезным связать с этой функцией число – ее «среднее значение» или, как говорят, «среднюю величину», «показатель центральной тенденции». По ряду причин, некоторые из которых будут ясны из дальнейшего, в качестве «среднего значения» обычно используют математическое ожидание.

Определение 3. Математическим ожиданием случайной величины X называется число

$$M(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega), \quad (4)$$

т.е. математическое ожидание случайной величины – это взвешенная сумма значений случайной величины с весами, равными вероятностям соответствующих элементарных событий.

Пример 6. Вычислим математическое ожидание числа, выпавшего на верхней грани игрального кубика. Непосредственно из определения 3 следует, что

$$M(X) = 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + 4\times\frac{1}{6} + 5\times\frac{1}{6} + 6\times\frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3{,}5.$$

Утверждение 2. Пусть случайная величина X принимает значения $x_1, x_2, \ldots, x_m$. Тогда справедливо равенство

$$M(X) = \sum_{1\le i\le m} x_i P(X = x_i), \quad (5)$$

т.е. математическое ожидание случайной величины – это взвешенная сумма значений случайной величины с весами, равными вероятностям того, что случайная величина принимает определенные значения.

В отличие от (4), где суммирование проводится непосредственно по элементарным событиям, случайное событие $\{X = x_i\} = \{\omega : X(\omega) = x_i\}$ может состоять из нескольких элементарных событий.

Иногда соотношение (5) принимают как определение математического ожидания. Однако с помощью определения 3, как показано далее, более легко установить свойства математического ожидания, нужные для построения вероятностных моделей реальных явлений, чем с помощью соотношения (5).

Для доказательства соотношения (5) сгруппируем в (4) члены с одинаковыми значениями случайной величины $X(\omega)$:

$$M(X) = \sum_{1\le i\le m} \left(\sum_{\omega: X(\omega)=x_i} X(\omega) P(\omega) \right).$$

Поскольку постоянный множитель можно вынести за знак суммы, то

$$\sum_{\omega: X(\omega)=x_i} X(\omega) P(\omega) = \sum_{\omega: X(\omega)=x_i} x_i P(\omega) = x_i \sum_{\omega: X(\omega)=x_i} P(\omega).$$

По определению вероятности события

$$\sum_{\omega: X(\omega)=x_i} P(\omega) = P(X = x_i).$$

С помощью двух последних соотношений получаем требуемое:

$$M(X) = \sum_{1\le i\le m}\left(x_i \sum_{\omega: X(\omega)=x_i} P(\omega)\right) = \sum_{1\le i\le m}\left(x_i P(X = x_i\right).$$

Понятие математического ожидания в вероятностно-статистической теории соответствует понятию центра тяжести в механике. Поместим в точки $x_1, x_2, \ldots, x_m$ на числовой оси массы $P(X=x_1), P(X=x_2), \ldots, P(X=x_m)$ соответственно. Тогда равенство (5) показывает, что центр тяжести этой системы материальных точек совпадает с математическим ожиданием, что показывает естественность определения 3.

Утверждение 3. Пусть X – случайная величина, $M(X)$ – ее математическое ожидание, a – некоторое число. Тогда

1) $M(a)=a$; 2) $M(X-M(X))=0$; 3) $M[(X-a)^2]=M[(X-M(X))^2]+(a-M(X))^2$.

Для доказательства рассмотрим сначала случайную величину, являющуюся постоянной, $X(\omega) = a$, т.е. функция $X(\omega)$ отображает пространство элементарных событий Ω в единственную точку a. Поскольку постоянный множитель можно выносить за знак суммы, то

$$M(X) = \sum_{\omega\in\Omega} aP(\omega) = a\sum_{\omega\in\Omega} P(\omega) = a.$$

Если каждый член суммы разбивается на два слагаемых, то и вся сумма разбивается на две суммы, из которых первая составлена из первых слагаемых, а вторая – из вторых. Следовательно, математическое ожидание суммы двух случайных величин $X+Y$, определенных на одном и том же пространстве элементарных событий, равно сумме математических ожиданий $M(X)$ и $M(Y)$ этих случайных величин:

$$M(X+Y) = M(X) + M(Y).$$

А потому $M(X-M(X)) = M(X) - M(M(X))$. Как показано выше, $M(M(X)) = M(X)$. Следовательно, $M(X-M(X)) = M(X) - M(X) = 0$.

Поскольку $(X - a)^2 = \{(X – M(X)) + (M(X) - a)\}^2 = (X - M(X))^2 + 2(X - M(X))(M(X) - a) + (M(X) – a)^2$, то $M[(X - a)^2] = M(X - M(X))^2 + M\{2(X - M(X))(M(X) - a)\} + M[(M(X) – a)^2]$. Упростим последнее равенство. Как показано в начале доказательства утверждения 3, математическое ожидание константы – сама эта константа, а потому $M[(M(X) – a)^2] = (M(X) – a)^2$. Поскольку постоянный множитель можно выносить за знак суммы, то $M\{2(X - M(X))(M(X) - a)\} = 2(M(X) - a)M(X - M(X))$. Правая часть последнего равенства равна 0, поскольку, как показано выше, $M(X-M(X))=0$. Следовательно, $M[(X-a)^2]=M[(X-M(X))^2]+(a-M(X))^2$, что и требовалось доказать.

Из сказанного вытекает, что $M[(X-a)^2]$ достигает минимума по a, равного $M[(X-M(X))^2]$, при $a = M(X)$, поскольку второе слагаемое в равенстве 3) всегда неотрицательно и равно 0 только при указанном значении a.

Утверждение 4. Пусть случайная величина X принимает значения $x_1, x_2, \ldots, x_m$, а f – некоторая функция числового аргумента. Тогда

$$M[f(X)] = \sum_{1 \le i \le m} f(x_i) P(X = x_i).$$

Для доказательства сгруппируем в правой части равенства (4), определяющего математическое ожидание, члены с одинаковыми значениями $X(\omega)$:

$$M[f(X)] = \sum_{1 \le i \le m} \left(\sum_{\omega : X(\omega) = x_i} f(X(\omega)) P(\omega) \right).$$

Пользуясь тем, что постоянный множитель можно выносить за знак суммы, и определением вероятности случайного события (2), получаем

$$M[f(X)] = \sum_{1 \le i \le m} \left(f(x_i) \sum_{\omega : X(\omega) = x_i} P(\omega) \right) = \sum_{1 \le i \le m} f(x_i) P(X = x_i),$$

что и требовалось доказать.

Утверждение 5. Пусть *X* и *У* – случайные величины, определенные на одном и том же пространстве элементарных событий, *a* и *b* – некоторые числа. Тогда *M(aX+bY)=aM(X)+bM(Y)*.

С помощью определения математического ожидания и свойств символа суммирования получаем цепочку равенств:

$$aM(X)+bM(Y)=a\sum_{\omega\in\Omega}X(\omega)P(\omega)+b\sum_{\omega\in\Omega}Y(\omega)P(\omega)=$$
$$=\sum_{\omega\in\Omega}(aX(\omega)+bY(\omega))P(\omega)=M(aX+bY).$$

Требуемое доказано.

Выше показано, как зависит математическое ожидание от перехода к другому началу отсчета и к другой единице измерения (переход *Y=aX+b*), а также к функциям от случайных величин. Полученные результаты постоянно используются в технико-экономическом анализе, при оценке финансово-хозяйственной деятельности предприятия, при переходе от одной валюты к другой во внешнеэкономических расчетах, в нормативно-технической документации и др. Рассматриваемые результаты позволяют применять одни и те же расчетные формулы при различных параметрах масштаба и сдвига.

Независимость случайных величин – одно из базовых понятий теории вероятностей, лежащее в основе практических всех вероятностно-статистических методов принятия решений.

Определение 4. Случайные величины *X* и *У*, определенные на одном и том же пространстве элементарных событий, называются независимыми, если для любых чисел *a* и *b* независимы события $\{X=a\}$ и $\{Y=b\}$.

Утверждение 6. Если случайные величины *X* и *У* независимы, *a* и *b* – некоторые числа, то случайные величины *X+a* и *Y+b* также независимы.

Действительно, события {X+a=c} и {$Y+b=d$} совпадают с событиями {$X=c-a$} и {$Y=d-b$} соответственно, а потому независимы.

Пример 7. Случайные величины, определенные по результатам различных испытаний в схеме независимых испытаний, сами независимы. Это вытекает из того, что события, с помощью которых определяется независимость случайных величин, определяются по результатам различных испытаний, а потому независимы по определению независимых испытаний.

В вероятностно-статистических методах принятия решений постоянно используется следующий факт: если X и Y – независимые случайные величины, $f(X)$ и $g(Y)$ – случайные величины, полученные из X и Y с помощью некоторых функций f и g, то $f(X)$ и $g(Y)$ – также независимые случайные величины. Например, если X и Y независимы, то X^2 и $2Y+3$ независимы, $\log X$ и $\log Y$ независимы. Доказательство рассматриваемого факта – тема одной из контрольных задач в конце главы.

Подавляющее большинство вероятностно-статистических моделей, используемых на практике, основывается на понятии независимых случайных величин. Так, результаты наблюдений, измерений, испытаний, анализов, опытов обычно моделируются независимыми случайными величинами. Часто считают, что наблюдения проводятся согласно схеме независимых испытаний. Например, результаты финансово-хозяйственной деятельности предприятий, выработка рабочих, результаты (данные) измерений контролируемого параметра у изделий, отобранных в выборку при статистическом регулировании технологического процесса, ответы потребителей при маркетинговом опросе и другие типы данных, используемых при принятии решений, обычно рассматриваются как независимые случайные величины, вектора или элементы. Причина такой популярности понятия независимости случайных величин

состоит в том, что к настоящему времени теория продвинута существенно дальше для независимых случайных величин, чем для зависимых.

Часто используется следующее свойство независимых случайных величин.

Утверждение 7. Если случайные величины *X* и *У* независимы, то математическое ожидание произведения ХУ равно произведению математических ожиданий *X* и *У*, т.е. *M(XУ)=M(X)M(У).*

Доказательство. Пусть X принимает значения $x_1, x_2, \ldots, x_m$, в то время как *У* принимает значения $y_1, y_2, \ldots, y_k$. Сгруппируем в задающей *M(XУ)* сумме члены, в которых X и *У* принимают фиксированные значения:

$$M(XY) = \sum_{1\le i\le m, 1\le j\le k} \left(\sum_{\omega: X(\omega)=x_i, Y(\omega)=y_j} X(\omega)Y(\omega)P(\omega) \right). \quad (6)$$

Поскольку постоянный множитель можно вынести за знак суммы, то

$$\sum_{\omega: X(\omega)=x_i, Y(\omega)=y_j} X(\omega)Y(\omega)P(\omega) = x_i y_j \sum_{\omega: X(\omega)=x_i, Y(\omega)=y_j} P(\omega).$$

Из последнего равенства и определения вероятности события заключаем, что равенство (6) можно преобразовать к виду

$$M(XY) = \sum_{1\le i\le m, 1\le j\le k} x_i y_j P(X=x_i, Y=y_j).$$

Так как *X* и *У* независимы, то $P(X=x_i, Y=y_j) = P(X=x_i)P(Y=y_j)$. Воспользовавшись этим равенством и свойством символа суммирования

$$\sum_{1\le i\le m, 1\le j\le k} c_i d_j = \left(\sum_{1\le i\le m} c_i \right) \left(\sum_{1\le j\le k} d_j \right),$$

заключаем, что

$$M(XY) = \left(\sum_{1\le i\le m} x_i P(X=x_i) \right) \left(\sum_{1\le j\le k} y_j P(Y=y_j) \right). \quad (7)$$

Из равенства (5) следует, что первый сомножитель в правой части (7) есть *M(X),* а второй – *M(У),* что и требовалось доказать.

Пример 8. Построим пример, показывающий, что из равенства *M(XУ)=M(X)M(У)* не следует независимость случайных величин *X* и *У*. Пусть вероятностное пространство состоит из трех равновероятных элементов $\omega_1, \omega_2, \omega_3$. Пусть

$$X(\omega_1)=1, X(\omega_2)=0, X(\omega_3)=-1, Y(\omega_1)=Y(\omega_3)=1, Y(\omega_2)=0.$$

Тогда *XУ = X, M(X) = M(XУ)* = 0, следовательно, *M(XУ) = M(X)M(У)*. Однако при этом *P(X=0) = P(У=0) = P(X=0, У=0)* = $P(\omega_2)=1/3$, в то время как вероятность события {*X*=0, *Y*=0} в случае независимых X и У должна была равняться $\frac{1}{3}\times\frac{1}{3}=\frac{1}{9}$.

Независимость нескольких случайных величин *X, Y, Z,…* означает по определению, что для любых чисел *x, y, z,…* справедливо равенство

P(X=x, Y=y, Z=z,…) = P(X=x) P(Y=y) P(Z=z)…

Например, если случайные величины определяются по результатам различных испытаний в схеме независимых испытаний, то они независимы.

Дисперсия случайной величины. Математическое ожидание показывает, вокруг какой точки группируются значения случайной величины. Необходимо также уметь измерить изменчивость случайной величины относительно математического ожидания. Выше показано, что $M[(X-a)^2]$ достигает минимума по *a* при *a = M(X)*. Поэтому за показатель изменчивости случайной величины естественно взять именно $M[(X-M(X))^2]$.

Определение 5. Дисперсией случайной величины *X* называется число $\sigma^2 = D(X) = M[(X - M(X))^2].$

Установим ряд свойств дисперсии случайной величины, постоянно используемых в вероятностно-статистических методах принятия решений.

Утверждение 8. Пусть X – случайная величина, a и b – некоторые числа, $Y = aX + b$. Тогда $D(Y) = a^2D(X)$.

Как следует из утверждений 3 и 5, $M(Y) = aM(X) + b$. Следовательно, $D(Y) = M[(Y - M(Y))^2] = M[(aX + b - aM(X) - b)^2] = M[a^2(X - M(X))^2]$. Поскольку постоянный множитель можно выносить за знак суммы, то $M[a^2(X - M(X))^2] = a^2 M[(X - M(X))^2] = a^2 D(X)$.

Утверждение 8 показывает, в частности, как меняется дисперсия результата наблюдений при изменении начала отсчета и единицы измерения. Оно дает правило преобразования расчетных формул при переходе к другим значениям параметров сдвига и масштаба.

Утверждение 9. Если случайные величины X и $У$ независимы, то дисперсия их суммы $X+У$ равна сумме дисперсий: $D(X+Y) = D(X) + D(Y)$.

Для доказательства воспользуемся тождеством

$$(X + У – (M(X)+M(У))^2 = (X–M(X))^2$$
$$+ 2(X–M(X))(У–M(У)) + (У–M(У))^2,$$

которое вытекает из известной формулы элементарной алгебры $(a+b)^2 = a^2 + 2ab + b^2$ при подстановке $a = X\text{-}M(X)$ и $b = Y\text{-}M(Y)$. Из утверждений 3 и 5 и определения дисперсии следует, что

$$D(X+Y) = D(X) + D(Y) + 2M\{(X–M(X))(У–M(У))\}.$$

Согласно утверждению 6 из независимости X и У вытекает независимость $X\text{-}M(X)$ и $У\text{-}M(У)$. Из утверждения 7 следует, что

$$M\{(X–M(X))(У–M(У))\}= M(X–M(X))M(У–M(У)).$$

Поскольку $M(X–M(X)) = 0$ (см. утверждение 3), то правая часть последнего равенства равна 0, откуда с учетом двух предыдущих равенств и следует заключение утверждения 9.

Утверждение 10. Пусть $X_1, X_2, \ldots, X_k$ – попарно независимые случайные величины (т.е. X_i и X_j независимы, если $i \neq j$). Пусть Y_k – их сумма, $Y_k = X_1 + X_2 + \ldots + X_k$. Тогда математическое ожидание

суммы равно сумме математических ожиданий слагаемых, $M(Y_k) = M(X_1) + M(X_2) + \ldots + M(X_k)$, дисперсия суммы равна сумме дисперсий слагаемых, $D(Y_k) = D(X_1) + D(X_2) + \ldots + D(X_k)$.

Соотношения, сформулированные в утверждении 10, являются основными при изучении выборочных характеристик, поскольку результаты наблюдений или измерений, включенные в выборку, обычно рассматриваются в математической статистике, теории принятия решений и эконометрике как реализации независимых случайных величин.

Для любого набора числовых случайных величин (не только независимых) математическое ожидание их суммы равно сумме их математических ожиданий. Это утверждение является обобщением утверждения 5. Строгое доказательство легко проводится методом математической индукции.

При выводе формулы для дисперсии $D(Y_k)$ воспользуемся следующим свойством символа суммирования:

$$\left(\sum_{1\le i\le k} a_i\right)^2 = \left(\sum_{1\le i\le k} a_i\right)\left(\sum_{1\le j\le k} a_j\right) = \sum_{1\le i\le k,\,1\le j\le k} a_i a_j .$$

Положим $a_i = X_i - M(X_i)$, получим

$$(X_1 + X_2 + \ldots + X_k - M(X_1) - M(X_2) - \ldots - M(X_k))^2 =$$
$$= \sum_{1\le i\le k,\,1\le j\le k} (X_i - M(X_i))(X_j - M(X_j)).$$

Воспользуемся теперь тем, что математическое ожидание суммы равно сумме математических ожиданий:

$$D(Y_k) = \sum_{1\le i\le k,\,1\le j\le k} M\{(X_i - M(X_i))(X_j - M(X_j))\}. \quad (8)$$

Как показано при доказательстве утверждения 9, из попарной независимости рассматриваемых случайных величин следует, что $M\{(X_i - M(X_i))(X_j - M(X_j))\} = 0$ при $i \ne j$. Следовательно, в сумме (8) остаются только члены с $i=j$, а они равны как раз $D(X_i)$.

Полученные в утверждениях 8-10 фундаментальные свойства таких характеристик случайных величин, как математическое

ожидание и дисперсия, постоянно используются практически во всех вероятностно-статистических моделях реальных явлений и процессов.

Пример 9. Рассмотрим событие A и случайную величину X такую, что $X(\omega)=1$, если $\omega \in A$, и $X(\omega)=0$ в противном случае, т.е. если $\omega \in \Omega \setminus A$. Покажем, что $M(X) = P(A)$, $D(X) = P(A)(1 - P(A))$.

Воспользуемся формулой (5) для математического ожидания. Случайная величина X принимает два значения – 0 и 1, значение 1 с вероятностью $P(A)$ и значение 0 с вероятностью $1 - P(A)$, а потому $M(X) = 1 \times P(A) + 0 \times (1 - P(A)) = P(A)$. Аналогично $(X - M(X))^2 = (1 - P(A))^2$ с вероятностью $P(A)$ и $(X - M(X))^2 = (0 - P(A))^2$ с вероятностью $1 - P(A)$, а потому $D(A) = (1 - P(A))^2 P(A) + (P(A))^2(1 - P(A))$. Вынося общий множитель, получаем, что $D(A) = P(A)(1 - P(A))$.

Пример 10. Рассмотрим k независимых испытаний, в каждом из которых некоторое событие A может наступить, а может и не наступить. Введем случайные величины $X_1, X_2, \ldots, X_k$ следующим образом: $X_i(\omega) = 1$, если в i-ом испытании событие A наступило, и $X_i(\omega) = 0$ в противном случае. Тогда случайные величины $X_1, X_2, \ldots, X_k$ попарно независимы (см. пример 7). Как показано в примере 9, $M(X_i) = p$, $D(X_i) = p(1 - p)$, где $p = P(A)$. Иногда p называют «вероятностью успеха» - в случае, если наступление события A рассматривается как «успех».

Биномиальное распределение. Случайная величина $B = X_1 + X_2 + \ldots + X_k$ называется биномиальной. Ясно, что $0 \le B \le k$ при всех возможных исходах опытов. Чтобы найти распределение B, т.е. вероятности $P(B = a)$ при $a = 0, 1, \ldots, k$, достаточно знать p – вероятность наступления рассматриваемого события в каждом из опытов. Действительно, случайное событие $B = a$ осуществляется тогда и только тогда, когда событие A наступает ровно при a

испытаниях. Если известны номера всех этих испытаний (т.е. номера в последовательности испытаний), то вероятность одновременного осуществления в *a* опытах события *A* и в *k-a* опытах противоположного ему – это вероятность произведения *k* независимых событий. Вероятность произведения равна произведению вероятностей, т.е. $p^a(1 - p)^{k-a}$. Сколькими способами можно задать номера *a* испытаний из *k*? Это $\binom{k}{a}$ - число сочетаний из *k* элементов по *a*, рассматриваемое в комбинаторике. Как известно,

$$\binom{k}{a} = \frac{k!}{a!(k-a)!},$$

где символом *k!* обозначено произведение всех натуральных чисел от 1 до *k*, т.е. $k! = 1 \cdot 2 \cdot \ldots \cdot k$ (дополнительно принимают, что 0! = 1). Из сказанного следует, что биномиальное распределение, т.е. распределение биномиальной случайной величины, имеет вид

$$P(B = a) = \binom{k}{a} p^a (1-p)^{k-a}.$$

Название «биномиальное распределение» основано на том, что *P(B = a)* является членом с номером (*a*+1) в разложении по биному Ньютона

$$(A + C)^k = \sum_{0 \le j \le k} \binom{k}{j} A^{k-j} C^j,$$

если положить *A* = 1 – *p*, *C* = *p*. Тогда при *j* = *a* получим

$$\binom{k}{j} A^{k-j} C^j = P(B = a).$$

Для числа сочетаний из *k* элементов по *a*, кроме $\binom{k}{a}$, используют более распространенное в отечественной литературе обозначение C_k^a.

Из утверждения 10 и расчетов примера 9 следует, что для случайной величины B, имеющей биномиальное распределение, математическое ожидание и дисперсия выражаются формулами

$$M(B) = kp,\ \ D(B) = kp(1-p),$$

поскольку B является суммой k независимых случайных величин с одинаковыми математическими ожиданиями и дисперсиями, найденными в примере 9.

Неравенства Чебышёва. Во введении к разделу обсуждалась задача проверки того, что доля дефектной продукции в партии равна определенному числу. Для демонстрации вероятностно-статистического подхода к проверке подобных утверждений являются полезными неравенства, впервые примененные в теории вероятностей великим русским математиком Пафнутием Львовичем Чебышёвым (1821-1894) и потому носящие его имя. Эти неравенства широко используются в теории математической статистики, а также непосредственно применяются в ряде практических задач принятия решении. Например, в задачах статистического анализа технологических процессов и качества продукции в случаях, когда явный вид функции распределения результатов наблюдений не известен. Они применяются также в задаче исключения резко отклоняющихся результатов наблюдений.

Первое неравенство Чебышева. Пусть X – неотрицательная случайная величина (т.е. $X(\omega) \ge 0$ для любого $\omega \in \Omega$). Тогда для любого положительного числа a справедливо неравенство

$$P(X \ge a) \le \frac{M(X)}{a}.$$

Доказательство. Все слагаемые в правой части формулы (4), определяющей математическое ожидание, в рассматриваемом случае неотрицательны. Поэтому при отбрасывании некоторых слагаемых сумма не увеличивается. Оставим в сумме только те члены, для которых $X(\omega) \ge a$. Получим, что

$$M(X) \geq \sum_{\omega: X(\omega) \geq a} X(\omega) P(\omega). \quad (9)$$

Для всех слагаемых в правой части (9) $X(\omega) \geq a$, поэтому

$$\sum_{\omega: X(\omega) \geq a} X(\omega) P(\omega) \geq a \sum_{\omega: X(\omega) \geq a} P(\omega) = aP(X \geq a). \quad (10)$$

Из (9) и (10) следует требуемое.

Второе неравенство Чебышева. Пусть X – случайная величина. Для любого положительного числа *a* справедливо неравенство

$$P(|X - M(X)| \geq a) \leq \frac{D(X)}{a^2}.$$

Это неравенство содержалось в работе П.Л.Чебышёва «О средних величинах», доложенной Российской академии наук 17 декабря 1866 г. и опубликованной в следующем году.

Для доказательства второго неравенства Чебышёва рассмотрим случайную величину $У = (X - M(X))^2$. Она неотрицательна, и потому для любого положительного числа *b*, как следует из первого неравенства Чебышёва, справедливо неравенство

$$P(Y \geq b) \leq \frac{M(Y)}{b} = \frac{D(X)}{b}.$$

Положим $b = a^2$. Событие $\{Y \geq b\}$ совпадает с событием $\{|X - M(X)| \geq a\}$, а потому

$$P(|X - M(X)| \geq a) = P(Y \geq a^2) \leq \frac{D(X)}{a},$$

что и требовалось доказать.

Пример 11. Можно указать неотрицательную случайную величину *X* и положительное число *a* такие, что первое неравенство Чебышёва обращается в равенство.

Достаточно рассмотреть $X(\omega) = a$. Тогда $M(X) = a$, $M(X)/a = 1$ и $P(a \geq a) = 1$, т.е. $P(X \geq a) = M(X)|a = 1$.

Следовательно, первое неравенство Чебышёва в его общей формулировке не может быть усилено. Однако для подавляющего

большинства случайных величин, используемых при вероятностно-статистическом моделировании реальных явлений и процессов, левые части неравенств Чебышёва много меньше соответствующих правых частей.

Пример 12. Может ли первое неравенство Чебышёва обращаться в равенство при всех a? Оказывается, нет. Покажем, что для любой неотрицательной случайной величины с ненулевым математическим ожиданием можно найти такое положительное число a, что первое неравенство Чебышёва является строгим.

Действительно, математическое ожидание неотрицательной случайной величины либо положительно, либо равно 0. В первом случае возьмем положительное a, меньшее положительного числа $M(X)$, например, положим $a = M(X)/2$. Тогда $M(X)/a$ больше 1, в то время как вероятность события не может превышать 1, а потому первое неравенство Чебышева является для этого a строгим. Второй случай исключается условиями примера 11.

Отметим, что во втором случае равенство 0 математического ожидания влечет тождественное равенство 0 случайной величины. Для такой случайной величины левая и правая части первого неравенства Чебышёва равны 0 при любом положительном a.

Можно ли в формулировке первого неравенства Чебышева отбросить требование неотрицательности случайной величины X? А требование положительности a? Легко видеть, что ни одно из двух требований не может быть отброшено, поскольку иначе правая часть первого неравенства Чебышева может стать отрицательной.

Закон больших чисел. Неравенство Чебышёва позволяет доказать замечательный результат, лежащий в основе математической статистики – закон больших чисел. Из него вытекает, что выборочные характеристики при возрастании числа опытов приближаются к теоретическим, а это дает возможность оценивать параметры вероятностных моделей по опытным данным.

Без закона больших чисел не было бы *бóльшей* части прикладной математической статистики.

Теорема Чебышёва. Пусть случайные величины $X_1, X_2, \ldots, X_k$ попарно независимы и существует число C такое, что $D(X_i) \le C$ при всех $i = 1, 2, \ldots, k$. Тогда для любого положительного ε выполнено неравенство

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_k}{k} - \frac{M(X_1) + M(X_2) + \ldots + M(X_k)}{k}\right| \ge \varepsilon\right\} \le \frac{C}{k\varepsilon^2}. \quad (11)$$

Доказательство. Рассмотрим случайные величины $Y_k = X_1 + X_2 + \ldots, + X_k$ и $Z_k = Y_k/k$. Тогда согласно утверждению 10

$$M(Y_k) = M(X_1) + M(X_2) + \ldots + M(X_k),\ D(Y_k) = D(X_1) + D(X_2) + \ldots + D(X_k).$$

Из свойств математического ожидания следует, что $M(Z_k) = M(Y_k)/k$, а из свойств дисперсии - что $D(Z_k) = D(Y_k)/k^2$. Таким образом,

$$M(Z_k) = \{M(X_1) + M(X_2) + \ldots + M(X_k)\}/k,$$

$$D(Z_k) = \{D(X_1) + D(X_2) + \ldots + D(X_k)\}/k^2.$$

Из условия теоремы Чебышёва, что

$$D(Z_k) \le \frac{Ck}{k^2} = \frac{C}{k}.$$

Применим к Z_k второе неравенство Чебышёва. Получим для стоящей в левой части неравенства (11) вероятности оценку

$$P\{|Z_k - M(Z_k)| \ge \varepsilon\} \le \frac{D(Z_k)}{\varepsilon^2} \le \frac{C}{k\varepsilon^2},$$

что и требовалось доказать.

Эта теорема была получена П.Л.Чебышёвым в той же работе 1867 г. «О средних величинах», что и неравенства Чебышёва.

Пример 13. Пусть $C = 1$, $\varepsilon = 0{,}1$. При каких k правая часть неравенства (11) не превосходит 0,1? 0,05? 0,00001?

В рассматриваемом случае правая часть неравенства (11) равно $100/k$. Она не превосходит 0,1, если k не меньше 1000, не превосходит 0,05, если k не меньше 2000, не превосходит 0,00001, если k не меньше 10 000 000.

Правая часть неравенства (11), а вместе с ней и левая, при возрастании k и фиксированных C и ε убывает, приближаясь к 0. Следовательно, вероятность того, что среднее арифметическое независимых случайных величин отличается от своего математического ожидания менее чем на ε, приближается к 1 при возрастании числа случайных величин, причем при любом ε. Это утверждение называют ЗАКОНОМ БОЛЬШИХ ЧИСЕЛ.

Наиболее важен для вероятностно-статистических методов принятия решений (и для математической статистики в целом) случай, когда все X_i , i = 1, 2, …, имеют одно и то же математическое ожидание $M(X_1)$ и одну и ту же дисперсию $\sigma^2 = D(X_1)$. В качестве замены (оценки) неизвестного исследователю математического ожидания используют выборочное среднее арифметическое

$$\overline{X} = \frac{X_1 + X_2 + ... + X_k}{k}.$$

Из закона больших чисел следует, что $\overline{X}$ при увеличении числа опытов (испытаний, измерений) сколь угодно близко приближается к $M(X_1)$, что записывают так:

$$\overline{X} \xrightarrow{P} M(X_1).$$

Здесь знак $\xrightarrow{P}$ означает «сходимость по вероятности». Обратим внимание, что понятие «сходимость по вероятности» отличается от понятия «переход к пределу» в математическом анализе. Напомним, что последовательность b_n имеет предел b при $n \to \infty$, если для любого сколь угодно малого $\delta > 0$ существует число $n(\delta)$ такое, что при любом $n > n(\delta)$ справедливо утверждение: $b_n \in (b - \delta; b + \delta)$. При использовании понятия «сходимость по вероятности» элементы последовательности предполагаются случайными, вводится еще одно сколь угодно малое число $\varepsilon > 0$ и

утверждение $b_n \in (b-\delta; b+\delta)$ предполагается выполненным не наверняка, а с вероятностью не менее $1-\varepsilon$.

Сходимость частот к вероятностям. Уже отмечалось, что с точки зрения ряда естествоиспытателей вероятность события A – это число, к которому приближается отношение количества осуществлений события A к количеству всех опытов при безграничном увеличении числа опытов. Известный математики Якоб Бернулли (1654-1705), живший в городе Базель в Швейцарии, в самом конце XVII века доказал это утверждение в рамках математической модели (опубликовано доказательство было лишь после его смерти, в 1713 году). Современная формулировка теоремы Бернулли такова.

Теорема Бернулли. Пусть m – число наступлений события A в k независимых (попарно) испытаниях, и p есть вероятность наступления события A в каждом из испытаний. Тогда при любом $\varepsilon > 0$ справедливо неравенство

$$P\left\{\left|\frac{m}{k} - p\right| \geq \varepsilon\right\} \leq \frac{p(1-p)}{k\varepsilon^2}. \quad (12)$$

Доказательство. Как показано в примере 10, случайная величина m имеет биномиальное распределение с вероятностью успеха p и является суммой k независимых случайных величин X_i, i = 1, 2. …, k, каждое из которых равно 1 с вероятностью p и 0 с вероятностью 1-p, т.е. $m = X_1 + X_2 + \ldots + X_k$.Применим к $X_1, X_2, \ldots, X_k$ теорему Чебышёва с $C = p(1 - p)$ и получим требуемое неравенство (12).

Теорема Бернулли дает возможность связать математическое определение вероятности (по А.Н.Колмогорову) с определением ряда естествоиспытателей (по Р. Мизесу (1883-1953)), согласно которому вероятность есть предел частоты в бесконечной последовательности испытаний. Продемонстрируем эту связь. Для этого сначала отметим, что

$$p(1-p) \le 1/4$$

при всех p. Действительно,

$$1/4 - p(1-p) = (p - 1/2)^2 \ge 0.$$

Следовательно, в теореме Чебышёва можно использовать C = j. Тогда при любом p и фиксированном ε правая часть неравенства (12) при возрастании k приближается к 0, что и доказывает согласие математического определения в рамках вероятностной модели с мнением естествоиспытателей.

Есть и прямые экспериментальные подтверждения того, что частота осуществления определенных событий близка к вероятности, определенной из теоретических соображений. Рассмотрим бросания монеты. Поскольку и герб, и решетка имеют одинаковые шансы оказаться сверху, то вероятность выпадения герба равна S из соображений равновозможности. Французский естествоиспытатель XVIII века Бюффон бросил монету 4040 раз, герб выпал при этом 2048 раз. Частота появления герба в опыте Бюффона равна 0,507. Английский статистик К.Пирсон бросил монету 12000 раз и при этом наблюдал 6019 выпадений герба – частота 0,5016. В другой раз он бросил монету 24000 раз, герб выпал 12012 раз – частота 0,5005. Как видим, во всех этих случаях частоты лишь незначительно отличаются от теоретической вероятности 0,5 [6, с.148].

О проверке статистических гипотез. С помощью неравенства (12) можно кое-что сказать по поводу проверки соответствия качества продукции заданным требованиям.

Пусть из 100000 единиц продукции 30000 оказались дефектными. Согласуется ли это с гипотезой о том, что вероятность дефектности равна 0,23? Прежде всего, какую вероятностную модель целесообразно использовать? Принимаем, что проводится сложный опыт, состоящий из 100000 испытаний 100000 единиц

продукции на годность. Считаем, что испытания (попарно) независимы и что в каждом испытании вероятность того, что единица продукции является дефектной, равна p. В реальном опыте получено, что событие «единица продукции не является годной» осуществилось 30000 раз при 100000 испытаниях. Согласуется ли это с гипотезой о том, что вероятность дефектности $p = 0{,}23$?

Для проверки гипотезы воспользуемся неравенством (12). В рассматриваемом случае $k = 100000$, $m = 30000$, $m/k = 0{,}3$, $p = 0{,}23$, $m/k - p = 0{,}07$. Для проверки гипотезы поступают так. Оценим вероятность того, что m/k отличается от p так же, как в рассматриваемом случае, или больше, т.е. оценим вероятность выполнения неравенства $|m/k - 0{,}23| \geq 0{,}07$. Положим в неравенстве (12) $p = 0{,}23$, $\varepsilon = 0{,}07$. Тогда

$$P\left\{\left|\frac{m}{k} - 0{,}23\right| \geq 0{,}07\right\} \leq \frac{0{,}23 \cdot 0{,}77}{0{,}0049k} \approx \frac{36{,}11}{k}. \quad (13)$$

При $k = 100000$ правая часть (13) меньше 1/2500. Значит, вероятность того, что отклонение будет не меньше наблюдаемого, весьма мала. Следовательно, если исходная гипотеза верна, то в рассматриваемом опыте осуществилось событие, вероятность которого меньше 1/2500. Поскольку 1/2500 – очень маленькое число, то исходную гипотезу надо отвергнуть.

Подробнее методы проверки статистических гипотез будут рассмотрены ниже. Здесь отметим, что одна из основных характеристик метода проверки гипотезы – уровень значимости, т.е. вероятность отвергнуть проверяемую гипотезу (ее в математической статистике называют нулевой и обозначают H_0), когда она верна. Для проверки статистической гипотезы часто поступают так. Выбирают уровень значимости - малое число α. Если описанная в предыдущем абзаце вероятность меньше α, то гипотезу отвергают, как говорят, на уровне значимости α. Если эта вероятность больше или равна α, то гипотезу принимают. Обычно

в вероятностно-статистических методах принятия решений выбирают α = 0,05, значительно реже α = 0,01 или α = 0,1, в зависимости от конкретной практической ситуации. В рассматриваемом случае α, напомним, та доля опытов (т.е. проверок партий по 100000 единиц продукции), в которой мы отвергаем гипотезу H_0: p = 0,23, хотя она верна.

Насколько результат проверки гипотезы H_0 зависит от числа испытаний k? Пусть при k = 100, k = 1000, k = 10000 оказалось, что m = 30, m = 300, m = 3000 соответственно, так что во всех случаях m/k = 0,3. Какие значения принимает вероятность

$$P_k = P\left\{\left|\frac{m}{k} - 0{,}23\right| \geq 0{,}07\right\}$$

и ее оценка – правая часть формулы (13)?

При k = 100 правая часть (13) равна приблизительно 0,36, что не дает оснований отвергнуть гипотезу. При k = 1000 правая часть (13) равна примерно 0,036. Гипотеза отвергается на уровне значимости α = 0,05 (и б = 0,1), но на основе оценки вероятности с помощью правой части формулы (13) не удается отвергнуть гипотезу на уровне значимости б = 0,01. При k = 10000 правая часть (13) меньше 1/250, и гипотеза отвергается на всех обычно используемых уровнях значимости.

Более точные расчеты, основанные на применении центральной предельной теоремы теории вероятностей (см. ниже), дают P_{100} = 0,095, P_{1000} = 0,0000005, так что оценка (13) является в рассматриваемом случае весьма завышенной. Причина в том, что получена она из наиболее общих соображений, применительно ко всем возможным случайным величинам улучшить ее нельзя (см. пример 11 выше), но применительно к конкретному биномиальному распределению – можно.

Ясно, что без введения уровня значимости не обойтись, ибо даже очень большие отклонения m/k от p имеют положительную

вероятность осуществления. Так, при справедливости гипотезы H_0 событие «все 100000 единиц продукции являются дефектными» отнюдь не является невозможным с математической точки зрения, оно имеет положительную вероятность осуществления, равную $0{,}23^{100000}$, хотя эта вероятность и невообразимо мала.

Аналогично разберем проверку гипотезы о симметричности монеты.

Пример 14. Если монета симметрична, то p = S, где p – вероятность выпадения герба. Согласуется ли с этой гипотезой результат эксперимента, в котором при 10000 бросаниях выпало 4000 гербов?

В рассматриваемом случае $m/k = 0{,}4$. Положим в неравенстве (12) $p = 0{,}5$, $e = 0{,}1$:

$$P\left\{\left|\frac{m}{k} - 0{,}5\right| \geq 0{,}1\right\} \leq \frac{0{,}5 \cdot 0{,}5}{0{,}01k} = \frac{25}{k}.$$

При k = 10000 правая часть последнего неравенства равна 1/400. Значит, если исходная гипотеза верна, то в нашем единственном эксперименте осуществилось событие, вероятность которого весьма мала – меньше 1/400. Поэтому исходную гипотезу необходимо отвергнуть.

Если из 1000 бросаний монеты гербы выпали в 400 случаях, то правая часть выписанного выше неравенства равна 1/40. Гипотеза симметричности отклоняется на уровне значимости 0,05 (и 0,1), но рассматриваемые методы не дают возможности отвергнуть ее на уровне значимости 0,01.

Если k = 100, а m = 40, то правая часть неравенства равна j. Оснований для отклонения гипотезы нет. С помощью более тонких методов, основанных на центральной предельной теореме теории вероятностей, можно показать, что левая часть неравенства равна приблизительно 0,05. Это показывает, как важно правильно выбрать метод проверки гипотезы или оценивания параметров.

Следовательно, целесообразна стандартизация подобных методов, позволяющая сэкономить усилия, необходимые для сравнения и выбора наилучшего метода, а также избежать устаревших, неверных или неэффективных методов.

Ясно, что даже по нескольким сотням опытов нельзя достоверно отличить абсолютно симметричную монету (p = S) от несколько несимметричной монеты (для которой, скажем, p = 0,49). Более того, любая реальная монета несколько несимметрична, так что монета с p = S - математическая абстракция. Между тем в ряде управленческих и производственных ситуаций необходимо осуществить справедливую жеребьевку, а для этого требуется абсолютно симметричная монета. Например, речь может идти об очередности рассмотрения инвестиционных проектов комиссией экспертов, о порядке вызова для собеседования кандидатов на должность, об отборе единиц продукции из партии в выборку для контроля и т.п.

Пример 15. Можно ли с помощью несимметричной монеты получить последовательность испытаний с двумя исходами, каждый из которых имеет вероятность 1/2 ?

Ответ: да, можно. Приведем способ, предложенный видным польским математиком Гуго Штейнгаузом (1887-1972).

Будем бросать монету два раза подряд и записывать исходы бросаний так (Г – герб, Р – решетка, на первом месте стоит результат первого бросания, на втором – второго): ГР запишем как Г, в то время РГ запишем как Р, а ГГ и РР вообще не станем записывать. Например, если исходы бросаний окажутся такими:

ГР, РГ, ГР, РР, ГР, РГ, ГГ, РГ, РР, РГ,

то запишем их в виде:

Г, Р, Г, Г, Р, Р, Р.

Сконструированная таким образом последовательность обладает теми же свойствами, что и полученная при бросании идеально

симметричной монеты, поскольку даже у несимметричной монеты последовательность ГР встречается столь же часто, как и последовательность РГ.

Применим теорему Бернулли и неравенство (12) к обработке реальных данных.

Пример 16. С 1871 г. по 1900 г. в Швейцарии родились 1359671 мальчик и 1285086 девочек. Совместимы ли эти данные с предположением о том, что вероятность рождения мальчика равна 0,5? А с предположением, что она равна 0,515? Другими словами, требуется проверить нулевые гипотезы H_0: $p = 0{,}5$ и H_0: $p = 0{,}515$ с помощью неравенства (12).

Число испытаний равно общему числу рождений, т.е. 1359671 + 1285086 = 2644757. Есть все основания считать испытания независимыми. Число рождений мальчиков составляет приблизительно 0,514 всех рождений. В случае $p = S$ имеем $e = 0{,}014$, и правая часть неравенства (12) имеет вид

$$\frac{0{,}5 \times 0{,}5}{0{,}014 \times 0{,}014 \times 2644757} \approx 0{,}00001.$$

Таким образом, гипотезу $p = 0{,}5$ следует считать несовместимой с приведенными в условии данными. В случае $p = 0{,}515$ имеем $e = 0{,}001$, и правая часть (12) равна приблизительно 0,1, так что с помощью неравенства (12) отклонить гипотезу H_0: $p = 0{,}515$ нельзя.

Итак, здесь на основе элементарной теории вероятностей (с конечным пространством элементарных событий) мы сумели построить вероятностные модели для описания проверки качества деталей (единиц продукции) и бросания монет и предложить методы проверки гипотез, относящихся к этим явлениям. В математической статистике есть более тонкие и сложные методы проверки описанных выше гипотез, которыми и пользуются в практических расчетах.

Можно спросить: «В рассмотренных выше моделях вероятности были известны заранее – со слов Струкова или же из-за того, что мы предположили симметричность монеты. А как строить модели, если вероятности неизвестны? Как оценить неизвестные вероятности?» Теорема Бернулли – результат, с помощью которого дается ответ на этот вопрос. Именно, оценкой неизвестной вероятности p является число m/k, поскольку доказано, что при возрастании k вероятность того, что m/k отличается от p более чем на какое-либо фиксированное число, приближается к 0. Оценка будет тем точнее, чем больше k. Более того, можно доказать, что с некоторой точки зрения (см. далее) оценка m/k для вероятности p является наилучшей из возможных (в терминах математической статистики – состоятельной, несмещенной и эффективной).

3. Суть вероятностно-статистических методов

Как подходы, идеи и результаты теории вероятностей и математической статистики используются при обработке данных – результатов наблюдений, измерений, испытаний, анализов, опытов с целью принятия практически важных решений?

Базой является вероятностная модель реального явления или процесса, т.е. математическая модель, в которой объективные соотношения выражены в терминах теории вероятностей. Вероятности используются прежде всего для описания неопределенностей, которые необходимо учитывать при принятии решений. Имеются в виду как нежелательные возможности (риски), так и привлекательные («счастливый случай»). Иногда случайность вносится в ситуацию сознательно, например, при жеребьевке, случайном отборе единиц для контроля, проведении лотерей или опросов потребителей.

Теория вероятностей позволяет по одним вероятностям рассчитать другие, интересующие исследователя. Например, по вероятности выпадения герба можно рассчитать вероятность того, что при 10 бросаниях монет выпадет не менее 3 гербов. Подобный расчет опирается на вероятностную модель, согласно которой бросания монет описываются схемой независимых испытаний, кроме того, выпадения герба и решетки равновозможны, а потому вероятность каждого из этих событий равна S. Более сложной является модель, в которой вместо бросания монеты рассматривается проверка качества единицы продукции. Соответствующая вероятностная модель опирается на предположение о том, что контроль качества различных единиц продукции описывается схемой независимых испытаний. В отличие от модели с бросанием монет необходимо ввести новый параметр – вероятность p того, что единица продукции является дефектной. Модель будет полностью описана, если принять, что все единицы продукции имеют одинаковую вероятность оказаться дефектными. Если последнее предположение неверно, то число параметров модели возрастает. Например, можно принять, что каждая единица продукции имеет свою вероятность оказаться дефектной.

Обсудим модель контроля качества с общей для всех единиц продукции вероятностью дефектности p. Чтобы при анализе модели «дойти до числа», необходимо заменить p на некоторое конкретное значение. Для этого необходимо выйти из рамок вероятностной модели и обратиться к данным, полученным при контроле качества. Математическая статистика решает обратную задачу по отношению к теории вероятностей. Ее цель – на основе результатов наблюдений (измерений, анализов, испытаний, опытов) получить выводы о вероятностях, лежащих в основе вероятностной модели. Например, на основе частоты появления дефектных изделий при контроле можно сделать выводы о вероятности дефектности (см.

обсуждение выше сиспользованием теоремы Бернулли). На основе неравенства Чебышева делались выводы о соответствии частоты появления дефектных изделий гипотезе о том, что вероятность дефектности принимает определенное значение.

Таким образом, применение математической статистики опирается на вероятностную модель явления или процесса. Используются два параллельных ряда понятий – относящиеся к теории (вероятностной модели) и относящиеся к практике (выборке результатов наблюдений). Например, теоретической вероятности соответствует частота, найденная по выборке. Математическому ожиданию (теоретический ряд) соответствует выборочное среднее арифметическое (практический ряд). Как правило, выборочные характеристики являются оценками теоретических. При этом величины, относящиеся к теоретическому ряду, «находятся в головах исследователей», относятся к миру идей (по древнегреческому философу Платону), недоступны для непосредственного измерения. Исследователи располагают лишь выборочными данными, с помощью которых они стараются установить интересующие их свойства теоретической вероятностной модели.

Зачем же нужна вероятностная модель? Дело в том, что только с ее помощью можно перенести свойства, установленные по результатам анализа конкретной выборки, на другие выборки, а также на всю так называемую генеральную совокупность. Термин «генеральная совокупность» используется, когда речь идет о большой, но конечной совокупности изучаемых единиц. Например, о совокупности всех жителей России или совокупности всех потребителей растворимого кофе в Москве. Цель маркетинговых или социологических опросов состоит в том, чтобы утверждения, полученные по выборке из сотен или тысяч человек, перенести на генеральные совокупности в несколько миллионов человек. При

контроле качества в роли генеральной совокупности выступает партия продукции.

Чтобы перенести выводы с выборки на более обширную совокупность, необходимы те или иные предположения о связи выборочных характеристик с характеристиками этой более обширной совокупности. Эти предположения основаны на соответствующей вероятностной модели.

Конечно, можно обрабатывать выборочные данные, не используя ту или иную вероятностную модель. Например, можно рассчитывать выборочное среднее арифметическое, подсчитывать частоту выполнения тех или иных условий и т.п. Однако результаты расчетов будут относиться только к конкретной выборке, перенос полученных с их помощью выводов на какую-либо иную совокупность некорректен. Иногда подобную деятельность называют «анализ данных». По сравнению с вероятностно-статистическими методами анализ данных имеет ограниченную познавательную ценность.

Итак, использование вероятностных моделей на основе оценивания и проверки гипотез с помощью выборочных характеристик – вот суть вероятностно-статистических методов принятия решений.

Подчеркнем, что логика использования выборочных характеристик для принятия решений на основе теоретических моделей предполагает одновременное использование двух параллельных рядов понятий, один из которых соответствует вероятностным моделям, а второй – выборочным данным. К сожалению, в ряде литературных источников, обычно устаревших либо написанных в рецептурном духе, не делается различия между выборочными и теоретическими характеристиками, что приводит читателей к недоумениям и ошибкам при практическом использовании статистических методов.

4. Случайные величины и их распределения

Распределения случайных величин и функции распределения. Распределение числовой случайной величины – это функция, которая однозначно определяет вероятность того, что случайная величина принимает заданное значение или принадлежит к некоторому заданному интервалу.

Первое – если случайная величина принимает конечное число значений. Тогда распределение задается функцией $P(X = x)$, ставящей каждому возможному значению x случайной величины X вероятность того, что $X = x$.

Второе – если случайная величина принимает бесконечно много значений. Это возможно лишь тогда, когда вероятностное пространство, на котором определена случайная величина, состоит из бесконечного числа элементарных событий. Тогда распределение задается набором вероятностей $P(a \leq X < b)$ для всех пар чисел a, b таких, что $a<b$. Распределение может быть задано с помощью т.н. *функции распределения* $F(x) = P(X<x)$, определяющей для всех действительных x вероятность того, что случайная величина X принимает значения, меньшие x. Ясно, что

$$P(a \leq X < b) = F(b) - F(a).$$

Это соотношение показывает, что как распределение может быть рассчитано по функции распределения, так и, наоборот, функция распределения – по распределению.

Используемые в прикладных исследованиях функции распределения бывают либо дискретными, либо непрерывными, либо их комбинациями.

Дискретные функции распределения соответствуют дискретным случайным величинам, принимающим конечное число значений или же значения из множества, элементы которого можно

перенумеровать натуральными числами (такие множества в математике называют счетными). Их график имеет вид ступенчатой лестницы (рис. 1).

Пример 1. Число X дефектных изделий в партии принимает значение 0 с вероятностью 0,3, значение 1 с вероятностью 0,4, значение 2 с вероятностью 0,2 и значение 3 с вероятностью 0,1. График функции распределения случайной величины X изображен на рис.1.

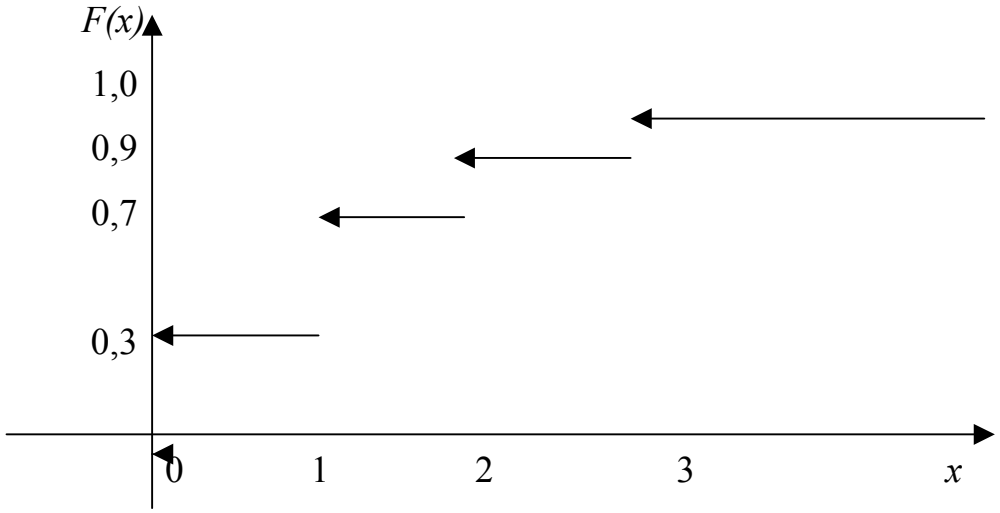

Рис.1. График функции распределения числа дефектных изделий.

Непрерывные функции распределения не имеют скачков. Они монотонно возрастают [1] при увеличении аргумента – от 0 при $x \to -\infty$ до 1 при $x \to +\infty$. Случайные величины, имеющие непрерывные функции распределения, называют непрерывными.

Практически используемые непрерывные функции распределения, как правило, имеют производные. Первая производная *f(x)* функции распределения *F(x)* называется плотностью вероятности,

[1] В некоторых случаях, например, при изучении цен, объемов выпуска или суммарной наработки на отказ в задачах надежности, функции распределения постоянны на некоторых интервалах, в

$$f(x) = \frac{dF(x)}{dx}.$$

По плотности вероятности можно определить функцию распределения:

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

Для любой функции распределения

$$\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to+\infty} F(x) = 1,$$

а потому

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Перечисленные свойства функций распределения постоянно используются в вероятностно-статистических методах принятия решений. В частности, из последнего равенства вытекает конкретный вид констант в формулах для плотностей вероятностей, рассматриваемых ниже.

Пример 2. Часто используется следующая функция распределения:

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

где a и b – некоторые числа, $a<b$. Найдем плотность вероятности этой функции распределения:

$$f(x) = \begin{cases} 0, & x < a, \\ \dfrac{1}{b-a}, & a < x < b, \\ 0, & x > b \end{cases}$$

(в точках $x = a$ и $x = b$ производная функции $F(x)$ не существует).

которые значения исследуемых случайных величин не могут попасть.

Случайная величина с функцией распределения (1) называется «равномерно распределенной на отрезке [*a; b*]».

Смешанные функции распределения встречаются, в частности, тогда, когда наблюдения в какой-то момент прекращаются. Например, при анализе статистических данных, полученных при использовании планов испытаний на надежность, предусматривающих прекращение испытаний по истечении некоторого срока. Или при анализе данных о технических изделиях, потребовавших гарантийного ремонта.

Пример 3. Пусть, например, срок службы электрической лампочки – случайная величина с функцией распределения *F(t),* а испытание проводится до выхода лампочки из строя, если это произойдет менее чем за 100 часов от начала испытаний, или до момента t_0 = 100 часов. Пусть *G(t)* – функция распределения времени эксплуатации лампочки в исправном состоянии при этом испытании. Тогда

$$G(t) = \begin{cases} F(t), & t \le 100 \\ 1, & t > 100. \end{cases}$$

Функция *G(t)* имеет скачок в точке t_0, поскольку соответствующая случайная величина принимает значение t_0 с вероятностью $1-F(t_0)>0$.

Характеристики случайных величин. В вероятностно-статистических методах используется ряд характеристик случайных величин, выражающихся через функции распределения и плотности вероятностей.

Квантили. При описании дифференциации доходов, при нахождении доверительных границ для параметров распределений случайных величин и во многих иных случаях применяется такое понятие, как «квантиль порядка *p*», где $0 < p < 1$ (обозначается x_p). Квантиль порядка *p* – значение случайной величины, для которого

функция распределения принимает значение p или имеет место «скачок» со значения меньше p до значения больше p (рис.2). Может случиться, что это условие выполняется для всех значений x, принадлежащих этому интервалу (т.е. функция распределения постоянна на этом интервале и равна p). Тогда каждое такое значение называется «квантилем порядка p». Для непрерывных функций распределения, как правило, существует единственный квантиль x_p порядка p (рис.2), причем

$$F(x_p) = p. \quad (2)$$

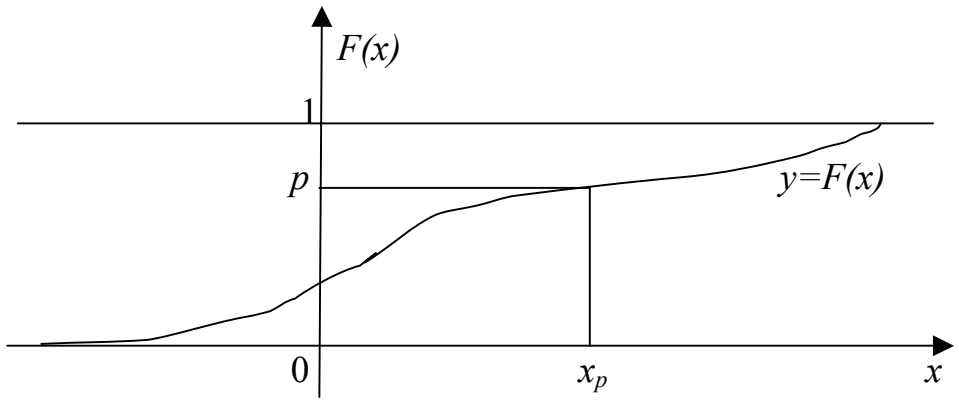

Рис.2. Определение квантиля x_p порядка p.

Пример 4. Найдем квантиль x_p порядка p для функции распределения $F(x)$ из (1).

При $0 < p < 1$ квантиль x_p находится из уравнения

$$\frac{x-a}{b-a} = p\,,$$

т.е. $x_p = a + p(b - a) = a(1- p) + bp$. При $p = 0$ любое $x \leq a$ является квантилем порядка $p = 0$. Квантилем порядка $p = 1$ является любое число $x \geq b$.

Для дискретных распределений, как правило, не существует x_p, удовлетворяющих уравнению (2). Точнее, если распределение случайной величины дается табл.1, где $x_1 < x_2 < \ldots < x_k$, то равенство (2), рассматриваемое как уравнение относительно x_p, имеет решения только для k значений p, а именно,

$p = p_1$,

$p = p_1 + p_2$,

$p = p_1 + p_2 + p_3$,

…

$p = p_1 + p_2 + \ldots + p_m, \qquad 3 < m < k$,

...

$p = p_1 + p_2 + \ldots + p_k$.

Таблица 1.

Распределение дискретной случайной величины

Значения x случайной величины X	x_1	x_2	...	x_k
Вероятности $P(X=x)$	p_1	p_2	...	p_k

Для перечисленных k значений вероятности p решение x_p уравнения (2) неединственно, а именно,

$$F(x) = p_1 + p_2 + \ldots + p_m$$

для всех x таких, что $x_m < x \leq x_{m+1}$. Т.е. x_p – любое число из интервала $(x_m; x_{m+1}]$. Для всех остальных p из промежутка (0;1), не входящих в перечень (3), имеет место «скачок» со значения меньше p до значения больше p. А именно, если

$$p_1 + p_2 + \ldots + p_m < p < p_1 + p_2 + \ldots + p_m + p_{m+1},$$

то $x_p = x_{m+1}$.

Рассмотренное свойство дискретных распределений создает значительные трудности при табулировании и использовании подобных распределений, поскольку невозможным оказывается точно выдержать типовые численные значения характеристик распределения. В частности, это так для критических значений и уровней значимости непараметрических статистических критериев (см. ниже), поскольку распределения статистик этих критериев дискретны.

Характеристики положения указывают на «центр» распределения. Большое значение в статистике имеет квантиль порядка p = S. Он называется медианой (случайной величины X или ее функции распределения $F(x)$) и обозначается $Me(X)$. В геометрии есть понятие «медиана» - прямая, проходящая через вершину треугольника и делящая противоположную его сторону пополам. В математической статистике медиана делит пополам не сторону треугольника, а распределение случайной величины: равенство $F(x_{0,5}) = 0{,}5$ означает, что вероятность попасть левее $x_{0,5}$ и вероятность попасть правее $x_{0,5}$ (или непосредственно в $x_{0,5}$) равны между собой и равны S, т.е.

$$P(X < x_{0,5}) = P(X \geq x_{0,5}) = \text{S}.$$

Медиана указывает «центр» распределения. С точки зрения одной из современных концепций – теории устойчивых статистических процедур – медиана является более хорошей характеристикой случайной величины, чем математическое ожидание [2, 7]. При обработке результатов измерений в порядковой шкале (см. главу о теории измерений) медианой можно пользоваться, а математическим ожиданием – нет.

Ясный смысл имеет такая характеристика случайной величины, как мода – значение (или значения) случайной величины, соответствующее локальному максимуму плотности вероятности для непрерывной случайной величины или локальному максимуму вероятности для дискретной случайной величины.

Если x_0 – мода случайной величины с плотностью $f(x)$, то, как известно из дифференциального исчисления, $\frac{df(x_0)}{dx} = 0$.

У случайной величины может быть много мод. Так, для равномерного распределения (1) каждая точка x такая, что $a < x < b$, является модой. Однако это исключение. Большинство случайных величин, используемых в вероятностно-статистических методах

принятия решений и других прикладных исследованиях, имеют одну моду. Случайные величины, плотности, распределения, имеющие одну моду, называются унимодальными.

Математическое ожидание для дискретных случайных величин с конечным числом значений рассмотрено в главе «События и вероятности». Для непрерывной случайной величины X математическое ожидание $M(X)$ удовлетворяет равенству

$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

являющемуся аналогом формулы (5) из утверждения 2 главы «События и вероятности».

Пример 5. Математическое ожидание для равномерно распределенной случайной величины X равно

$$M(X) = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\frac{x^2}{2}\Bigg|_a^b = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{a+b}{2}.$$

Для рассматриваемых в настоящей главе случайных величин верны все те свойства математических ожиданий и дисперсий, которые были рассмотрены ранее для дискретных случайных величин с конечным числом значений. Однако доказательства этих свойств не приводим, поскольку они требуют углубления в математические тонкости, не являющегося необходимым для понимания и квалифицированного применения вероятностно-статистических методов принятия решений.

Замечание. В настоящей книге сознательно обходятся математические тонкости, связанные, в частности, с понятиями измеримых множеств и измеримых функций, σ-алгебры событий и т.п. Желающим освоить эти понятия необходимо обратиться к специальной литературе, в частности, к энциклопедии [1].

Каждая из трех характеристик – математическое ожидание, медиана, мода – описывает «центр» распределения вероятностей. Понятие «центр» можно определять разными способами – отсюда

три разные характеристики. Однако для важного класса распределений – симметричных унимодальных – все три характеристики совпадают.

Плотность распределения *f(x)* – плотность симметричного распределения, если найдется число x_0 такое, что

$$f(x) = f(2x_0 - x). \quad (3)$$

Равенство (3) означает, что график функции $y = f(x)$ симметричен относительно вертикальной прямой, проходящей через центр симметрии $x = x_0$. Из (3) следует, что функция симметричного распределения удовлетворяет соотношению

$$F(x) = 1 - F(2x_0 - x). \quad (4)$$

Для симметричного распределения с одной модой математическое ожидание, медиана и мода совпадают и равны x_0.

Наиболее важен случай симметрии относительно 0, т.е. $x_0 = 0$. Тогда (3) и (4) переходят в равенства

$$f(x) = f(-x) \quad (5)$$

и

$$F(x) = 1 - F(-x) \quad (6)$$

соответственно. Приведенные соотношения показывают, что симметричные распределения нет необходимости табулировать при всех x, достаточно иметь таблицы при $x \geq x_0$.

Отметим еще одно свойство симметричных распределений, постоянно используемое в вероятностно-статистических методах принятия решений и других прикладных исследованиях. Для непрерывной функции распределения

$$P(|X| \leq a) = P(-a \leq X \leq a) = F(a) - F(-a),$$

где F – функция распределения случайной величины X. Если функция распределения F симметрична относительно 0, т.е. для нее справедлива формула (6), то

$$P(|X| \leq a) = 2F(a) - 1.$$

Часто используют другую формулировку рассматриваемого утверждения: если

$$1 - F(a) = \alpha ,$$

то

$$P(|X| > a) = 2\alpha .$$

Если x_α и $x_{1-\alpha}$ - квантили порядка α и $1-\alpha$ соответственно (см. (2)) функции распределения, симметричной относительно 0, то из (6) следует, что

$$x_\alpha = -x_{1-\alpha} .$$

Характеристики разброса. От характеристик положения – математического ожидания, медианы, моды – перейдем к характеристикам разброса случайной величины *X*: дисперсии $D(X) = \sigma^2$, среднему квадратическому отклонению σ и коэффициенту вариации *v*. Определение и свойства дисперсии для дискретных случайных величин рассмотрены в предыдущей главе. Для непрерывных случайных величин

$$D(X) = M[(X - M(X))^2] = \int_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx .$$

Среднее квадратическое отклонение – это неотрицательное значение квадратного корня из дисперсии:

$$\sigma = +\sqrt{D(X)} .$$

Коэффициент вариации – это отношение среднего квадратического отклонения к математическому ожиданию:

$$v = \frac{\sigma}{M(X)} .$$

Коэффициент вариации применяется при $M(X)>0$. Он измеряет разброс в относительных единицах, в то время как среднее квадратическое отклонение – в абсолютных.

Пример 6. Для равномерно распределенной случайной величины X найдем дисперсию, среднеквадратическое отклонение и коэффициент вариации. Дисперсия равна:

$$D(X) = \int_a^b \frac{1}{b-a}\left(x - \frac{a+b}{2}\right)^2 dx.$$

Замена переменной $y = x - \frac{a+b}{2}$ дает возможность записать:

$$D(X) = \frac{1}{b-a}\int_{-c}^{c} y^2 dy = \frac{1}{b-a}\frac{y^3}{3}\Bigg|_{-c}^{c} = \frac{2c^3}{3(b-a)} = \frac{(b-a)^2}{12},$$

где $c = (b - a)/2$. Следовательно, среднее квадратическое отклонение равно $\sigma = \frac{b-a}{2\sqrt{3}}$, а коэффициент вариации таков: $v = \frac{b-a}{\sqrt{3}(a+b)}$.

Преобразования случайных величин. По каждой случайной величине X определяют еще три величины – центрированную Y, нормированную V и приведенную U. Центрированная случайная величина Y – это разность между данной случайной величиной X и ее математическим ожиданием $M(X)$, т.е. $Y = X - M(X)$. Математическое ожидание центрированной случайной величины Y равно 0, а дисперсия – дисперсии данной случайной величины: $M(Y) = 0$, $D(Y) = D(X)$. Функция распределения $F_Y(x)$ центрированной случайной величины Y связана с функцией распределения $F(x)$ исходной случайной величины X соотношением:

$$F_Y(x) = F(x + M(X)).$$

Для плотностей этих случайных величин справедливо равенство

$$f_Y(x) = f(x + M(X)).$$

Нормированная случайная величина V – это отношение данной случайной величины X к ее среднему квадратическому отклонению σ, т.е. $V = X/\sigma$. Математическое ожидание и дисперсия

нормированной случайной величины V выражаются через характеристики X так:

$$M(V) = \frac{M(X)}{\sigma} = \frac{1}{v}, \quad D(V) = 1,$$

где v – коэффициент вариации исходной случайной величины X. Для функции распределения $F_V(x)$ и плотности $f_V(x)$ нормированной случайной величины V имеем:

$$F_V(x) = F(\sigma x), \quad f_V(x) = \sigma f(\sigma x),$$

где $F(x)$ – функция распределения исходной случайной величины X, а $f(x)$ – ее плотность вероятности.

Приведенная случайная величина U – это центрированная и нормированная случайная величина:

$$U = \frac{X - M(X)}{\sigma}.$$

Для приведенной случайной величины

$$M(U) = 0, \ D(U) = 1, \ F_U(x) = F(\sigma x + M(X)), \ f_U(x) = \sigma f(\sigma x + M(X)). \quad (7)$$

Нормированные, центрированные и приведенные случайные величины постоянно используются как в теоретических исследованиях, так и в алгоритмах, программных продуктах, нормативно-технической и инструктивно-методической документации. В частности, потому, что равенства $M(U) = 0, \ D(U) = 1$ позволяют упростить обоснования методов, формулировки теорем и расчетные формулы.

Используются преобразования случайных величин и более общего плана. Так, если $Y = aX + b$, где a и b – некоторые числа, то

$$M(Y) = aM(X) + b, \ D(Y) = \sigma^2 D(X), \ F_Y(x) = F\left(\frac{x-b}{a}\right), \ f_Y(x) = \frac{1}{a} f\left(\frac{x-b}{a}\right). \quad (8)$$

Пример 7. Если $a = 1/\sigma$, $b = -M(X)/\sigma$, то Y – приведенная случайная величина, и формулы (8) переходят в формулы (7).

С каждой случайной величиной X можно связать множество случайных величин Y, заданных формулой $Y = aX + b$ при различных $a>0$ и b. Это множество называют *масштабно-сдвиговым семейством*, порожденным случайной величиной X. Функции распределения $F_Y(x)$ составляют масштабно сдвиговое семейство распределений, порожденное функцией распределения $F(x)$. Вместо $Y = aX + b$ часто используют запись

$$Y = \frac{X - c}{d}, \quad (9)$$

где

$$d = \frac{1}{a} > 0, \ c = -\frac{b}{a}.$$

Число c называют параметром сдвига, а число d - параметром масштаба. Формула (9) показывает, что X – результат измерения некоторой величины – переходит в $У$ – результат измерения той же величины, если начало измерения перенести в точку c, а затем использовать новую единицу измерения, в d раз большую старой.

Для масштабно-сдвигового семейства (9) распределение X называют стандартным. В вероятностно-статистических методах принятия решений и других прикладных исследованиях используют стандартное нормальное распределение, стандартное распределение Вейбулла-Гнеденко, стандартное гамма-распределение и др. (см. ниже).

Применяют и другие преобразования случайных величин. Например, для положительной случайной величины X рассматривают $Y = \lg X$, где $\lg X$ – десятичный логарифм числа X. Цепочка равенств

$$F_Y(x) = P(\lg X < x) = P(X < 10^x) = F(10^x)$$

связывает функции распределения X и Y.

Моменты случайных величин. При обработке данных используют такие характеристики случайной величины X как

моменты порядка q, т.е. математические ожидания случайной величины X^q, $q = 1, 2, \ldots$ Так, само математическое ожидание – это момент порядка 1. Для дискретной случайной величины момент порядка q может быть рассчитан как

$$m_q = M(X^q) = \sum_i x_i^q P(X = x_i).$$

Для непрерывной случайной величины

$$m_q = M(X^q) = \int_{-\infty}^{+\infty} x^q f(x)dx.$$

Моменты порядка q называют также начальными моментами порядка q, в отличие от родственных характеристик – центральных моментов порядка q, задаваемых формулой

$$\mu_q = M[(X - M(X))^q], \ q = 2,3,...$$

Так, дисперсия – это центральный момент порядка 2.

Стандартное нормальное распределение и центральная предельная теорема. В вероятностно-статистических методах часто идет речь о нормальном распределении. Иногда его пытаются использовать для моделирования распределения исходных данных (эти попытки не всегда являются обоснованными – см. ниже). Более существенно, что многие методы обработки данных основаны на том, что расчетные величины имеют распределения, близкие к нормальному.

Пусть X_1, X_2,..., X_n, …– независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i = 1, 2, \ldots, n, \ldots$ Как следует из результатов предыдущей главы,

$$M(X_1 + X_2 + ... + X_n) = nm, \ D(X_1 + X_2 + ... + X_n) = n\sigma^2.$$

Рассмотрим приведенную случайную величину U_n для суммы $X_1 + X_2 + ... + X_n$, а именно,

$$U_n = \frac{X_1 + X_2 + ... + X_n - nm}{\sigma\sqrt{n}}.$$

Как следует из формул (7), $M(U_n) = 0$, $D(U_n) = 1$.

Центральная предельная теорема (для одинаково распределенных слагаемых). Пусть $X_1, X_2, \ldots, X_n, \ldots$ – независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i = 1, 2, \ldots, n, \ldots$ Тогда для любого x существует предел

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} < x \right) = \Phi(x),$$

где $\Phi(x)$ – функция стандартного нормального распределения.

Подробнее о функции $\Phi(x)$ – ниже (читается «фи от икс», поскольку Φ – греческая прописная буква «фи»).

Центральная предельная теорема (ЦПТ) носит свое название по той причине, что она является центральным, наиболее часто применяющимся математическим результатом теории вероятностей и математической статистики. История ЦПТ занимает около 200 лет – с 1730 г., когда английский математик А.Муавр (1667-1754) опубликовал первый результат, относящийся к ЦПТ (см. ниже о теореме Муавра-Лапласа), до двадцатых – тридцатых годов XX в., когда финн Дж.У. Линдеберг, француз Поль Леви (1886-1971), югослав В. Феллер (1906-1970), русский А.Я. Хинчин (1894-1959) и другие ученые получили необходимые и достаточные условия справедливости классической центральной предельной теоремы.

Развитие рассматриваемой тематики на этом отнюдь не прекратилось – изучали случайные величины, не имеющие дисперсии, т.е. те, для которых

$$\int_{-\infty}^{+\infty} x^2 f(x)dx = +\infty$$

(академик Б.В.Гнеденко и др.), ситуацию, когда суммируются случайные величины (точнее, случайные элементы) более сложной природы, чем числа (академики Ю.В.Прохоров, А.А.Боровков и их соратники), и т.д.

Функция распределения *Ф(x)* задается равенством

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x)dx,$$

где $\varphi(y)$ - плотность стандартного нормального распределения, имеющая довольно сложное выражение:

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

Здесь π=3,1415925… - известное в геометрии число, равное отношению длины окружности к диаметру, e = 2,718281828… - основание натуральных логарифмов (для запоминания этого числа обратите внимание, что 1828 – год рождения писателя Л.Н.Толстого). Как известно из математического анализа,

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n.$$

При обработке результатов наблюдений функцию нормального распределения в настоящее время уже не вычисляют по приведенным формулам, а находят с помощью специальных таблиц или компьютерных программ. Лучшие на русском языке «Таблицы математической статистики» составлены членами-корреспондентами АН СССР Л.Н. Большевым и Н.В.Смирновым [8].

Вид плотности стандартного нормального распределения $\varphi(y)$ вытекает из математической теории, которую не имеем возможности здесь рассматривать, равно как и доказательство ЦПТ.

Для иллюстрации приводим небольшие таблицы функции распределения *Ф(x)* (табл.2) и ее квантилей (табл.3). Функция *Ф(x)* симметрична относительно 0, что отражается в табл.2-3.

Если случайная величина *X* имеет функцию распределения *Ф(x),* то *M(X)* = 0, *D(X)* = 1. Это утверждение доказывается в теории вероятностей, исходя из вида плотности вероятностей $\varphi(y)$. Оно согласуется с аналогичным утверждением для характеристик

приведенной случайной величины U_n, что вполне естественно, поскольку ЦПТ утверждает, что при безграничном возрастании числа слагаемых функция распределения U_n стремится к функции стандартного нормального распределения *Φ(x),* причем этот предельный переход справедлив для любого числа *x*.

Таблица 2.

Функция стандартного нормального распределения.

x	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*
-5,0	0,00000029	-1,0	0,158655	2,0	0,9772499
-4,0	0,00003167	-0,5	0,308538	2,5	0,99379033
-3,0	0,00134990	0,0	0,500000	3,0	0,99865010
-2,5	0,00620967	0,5	0,691462	4,0	0,99996833
-2,0	0,0227501	1,0	0,841345	5,0	0,99999971
-1,5	0,0668072	1,5	0,9331928		

Таблица 3.

Квантили стандартного нормального распределения.

p	Квантиль порядка *p*	*p*	Квантиль порядка *p*
0,01	-2,326348	0,60	0,253347
0,025	-1,959964	0,70	0,524401
0,05	-1,644854	0,80	0,841621
0,10	-1,281552	0,90	1,281552
0,30	-0,524401	0,95	1,644854
0,40	-0,253347	0,975	1,959964
0,50	0,000000	0,99	2,326348

Семейство нормальных распределений. Введем понятие семейства нормальных распределений. По определению нормальным распределением называется распределение случайной величины *X*, для которой распределение приведенной случайной величины есть *Φ(x)*. Как следует из общих свойств масштабно-сдвиговых семейств распределений (см. выше), нормальное распределение – это распределение случайной величины

$$Y = \sigma X + m,$$

где *X* – случайная величина с распределением *Φ(X),* причем *m = M(Y),* σ^2 *= D(Y)*. Нормальное распределение с параметрами сдвига

m и масштаба σ обычно обозначается $N(m, \sigma)$ (иногда используется обозначение $N(m, \sigma^2)$).

Как следует из (8), плотность вероятности нормального распределения $N(m, \sigma)$ есть

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}.$$

Нормальные распределения образуют масштабно-сдвиговое семейство. При этом параметром масштаба является $d = 1/\sigma$, а параметром сдвига $c = - m/\sigma$.

Для центральных моментов третьего и четвертого порядка нормального распределения справедливы равенства

$$\mu_3 = 0, \ \mu_4 = 3\sigma^4.$$

Эти равенства лежат в основе классических методов проверки того, что результаты наблюдений подчиняются нормальному распределению. В настоящее время нормальность обычно рекомендуется проверять по критерию W Шапиро – Уилка. Проблема проверки нормальности обсуждается ниже.

Если случайные величины X_1 и X_2 имеют функции распределения $N(m_1, \sigma_1)$ и $N(m_2, \sigma_2)$ соответственно, то $X_1 + X_2$ имеет распределение $N\left(m_1 + m_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$ Следовательно, если случайные величины $X_1, X_2, \ldots, X_n$ независимы и имеют одно и тоже распределение $N(m, \sigma)$, то их среднее арифметическое

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

имеет распределение $N(m, \frac{\sigma}{\sqrt{n}})$. Эти свойства нормального распределения постоянно используются в различных вероятностно-статистических методах принятия решений, в частности, при статистическом регулировании технологических процессов и в статистическом приемочном контроле по количественному признаку.

Распределения Пирсона (хи – квадрат), Стьюдента и Фишера. С помощью нормального распределения определяются три распределения, которые в настоящее время часто используются при статистической обработке данных. В дальнейших разделах книги много раз встречаются эти распределения.

Распределение Пирсона χ^2 (хи - квадрат) – распределение случайной величины

$$X = X_1^2 + X_2^2 + ... + X_n^2,$$

где случайные величины X_1, X_2,..., X_n независимы и имеют одно и тоже распределение $N(0,1)$. При этом число слагаемых, т.е. n, называется «числом степеней свободы» распределения хи – квадрат.

Распределение хи-квадрат используют при оценивании дисперсии (с помощью доверительного интервала), при проверке гипотез согласия, однородности, независимости, прежде всего для качественных (категоризованных) переменных, принимающих конечное число значений, и во многих других задачах статистического анализа данных [8, 9, 11, 16].

Распределение t Стьюдента – это распределение случайной величины

$$T = \frac{U\sqrt{n}}{\sqrt{X}},$$

где случайные величины U и X независимы, U имеет распределение стандартное нормальное распределение $N(0,1)$, а X – распределение хи – квадрат с n степенями свободы. При этом n называется «числом степеней свободы» распределения Стьюдента.

Распределение Стьюдента было введено в 1908 г. английским статистиком В. Госсетом, работавшем на фабрике, выпускающей пиво. Вероятностно-статистические методы использовались для принятия экономических и технических решений на этой фабрике, поэтому ее руководство запрещало В. Госсету публиковать научные

статьи под своим именем. Таким способом охранялась коммерческая тайна, «ноу-хау» в виде вероятностно-статистических методов, разработанных В. Госсетом. Однако он имел возможность публиковаться под псевдонимом «Стьюдент». История Госсета - Стьюдента показывает, что еще сто лет назад менеджерам Великобритании была очевидна большая экономическая эффективность вероятностно-статистических методов.

В настоящее время распределение Стьюдента – одно из наиболее известных распределений среди используемых при анализе реальных данных. Его применяют при оценивании математического ожидания, прогнозного значения и других характеристик с помощью доверительных интервалов, по проверке гипотез о значениях математических ожиданий, коэффициентов регрессионной зависимости, гипотез однородности выборок и т.д. [8, 9, 11, 16].

Распределение Фишера – это распределение случайной величины

$$F = \frac{\frac{1}{k_1} X_1}{\frac{1}{k_2} X_2},$$

где случайные величины X_1 и X_2 независимы и имеют распределения хи – квадрат с числом степеней свободы k_1 и k_2 соответственно. При этом пара (k_1, k_2) – пара «чисел степеней свободы» распределения Фишера, а именно, k_1 – число степеней свободы числителя, а k_2 – число степеней свободы знаменателя. Распределение случайной величины F названо в честь великого английского статистика Р.Фишера (1890-1962), активно использовавшего его в своих работах.

Распределение Фишера используют при проверке гипотез об адекватности модели в регрессионном анализе, о равенстве дисперсий и в других задачах прикладной статистики [8, 9, 11, 16].

Выражения для функций распределения хи - квадрат, Стьюдента и Фишера, их плотностей и характеристик, а также таблицы, необходимые для их практического использования, можно найти в специальной литературе (см., например, [8]).

Центральная предельная теорема (общий случай). Как уже отмечалось, нормальные распределения в настоящее время часто используют в вероятностных моделях в различных прикладных областях. В чем причина такой широкой распространенности этого двухпараметрического семейства распределений? Она проясняется следующей теоремой.

Центральная предельная теорема (для разнораспределенных слагаемых). Пусть $X_1, X_2, \ldots, X_n, \ldots$ - независимые случайные величины с математическими ожиданиями $M(X_1), M(X_2), \ldots, M(X_n), \ldots$ и дисперсиями $D(X_1), D(X_2), \ldots, D(X_n), \ldots$ соответственно. Пусть

$$U_n = \frac{X_1 + X_2 + \ldots + X_n - M(X_1) - M(X_2) - \ldots - M(X_n)}{\sqrt{D(X_1) + D(X_2) + \ldots + D(X_n)}}.$$

Тогда при справедливости некоторых условий, обеспечивающих малость вклада любого из слагаемых в U_n,

$$\lim_{n \to \infty} P(U_n < x) = \Phi(x)$$

для любого x.

Условия, о которых идет речь, не будем здесь формулировать. Их можно найти в специальной литературе (см., например, [6]). «Выяснение условий, при которых действует ЦПТ, составляет заслугу выдающихся русских ученых А.А.Маркова (1857-1922) и, в особенности, А.М.Ляпунова (1857-1918)» [9, с.197].

Центральная предельная теорема показывает, что в случае, когда результат измерения (наблюдения) складывается под действием многих причин, причем каждая из них вносит лишь малый вклад, а совокупный итог определяется *аддитивно*, т.е.

путем сложения, то распределение результата измерения (наблюдения) близко к нормальному.

Иногда считают, что для нормальности распределения достаточно того, что результат измерения (наблюдения) X формируется под действием многих причин, каждая из которых оказывает малое воздействие. **Это заключение неверно.** Важно, как эти причины действуют. Если аддитивно – то X имеет приближенно нормальное распределение. Если *мультипликативно* (т.е. действия отдельных причин перемножаются, а не складываются), то распределение X близко не к нормальному, а к т.н. логарифмически нормальному, т.е. не X, а lg X имеет приблизительно нормальное распределение. Если же нет оснований считать, что действует один из этих двух механизмов формирования итогового результата (или какой-либо иной вполне определенный механизм), то про распределение X ничего определенного сказать нельзя.

Из сказанного вытекает, что в конкретной прикладной задаче нормальность результатов измерений (наблюдений), как правило, нельзя установить из общих соображений, ее следует проверять с помощью статистических критериев. Или же использовать непараметрические статистические методы, не опирающиеся на предположения о принадлежности функций распределения результатов измерений (наблюдений) к тому или иному параметрическому семейству.

Непрерывные распределения, используемые в вероятностно-статистических методах. Кроме масштабно-сдвигового семейства нормальных распределений, широко используют ряд других семейств распределения – логарифмически нормальных, экспоненциальных, Вейбулла-Гнеденко, гамма-распределений. Рассмотрим эти семейства.

Логарифмически нормальные распределения. Случайная величина X имеет логарифмически нормальное распределение, если случайная величина $Y = \lg X$ имеет нормальное распределение. Тогда $Z = \ln X = 2{,}3026\ldots Y$ также имеет нормальное распределение $N(a_1,y_1)$, где $\ln X$ - натуральный логарифм X. Плотность логарифмически нормального распределения такова:

$$f(x;a_1,\sigma_1) = \begin{cases} \dfrac{1}{\sigma_1\sqrt{2\pi}x}\exp\left[-\dfrac{(\ln x - a_1)^2}{2\sigma_1^2}\right], & x > 0, \\ 0, & x \le 0. \end{cases}$$

Из центральной предельной теоремы следует, что произведение $X = X_1X_2...X_n$ независимых положительных случайных величин X_i, $i = 1, 2,\ldots, n$, при больших n можно аппроксимировать логарифмически нормальным распределением. В частности, мультипликативная модель формирования заработной платы или дохода приводит к рекомендации приближать распределения заработной платы и дохода логарифмически нормальными законами. Для России эта рекомендация оказалась обоснованной - статистические данные подтверждают ее.

Имеются и другие вероятностные модели, приводящие к логарифмически нормальному закону. Классический пример такой модели дан А.Н.Колмогоровым [10], который из физически обоснованной системы постулатов вывел заключение о том, что размеры частиц при дроблении кусков руды, угля и т.п. на шаровых мельницах имеют логарифмически нормальное распределение.

Экспоненциальные распределения. Перейдем к другому семейству распределений, широко используемому в различных вероятностно-статистических методах принятия решений и других прикладных исследованиях, - семейству экспоненциальных распределений. Начнем с вероятностной модели, приводящей к таким распределениям. Для этого рассмотрим "поток событий", т.е. последовательность событий, происходящих одно за другим в

какие-то моменты времени. Примерами могут служить: поток вызовов на телефонной станции; поток отказов оборудования в технологической цепочке; поток отказов изделий при испытаниях продукции; поток обращений клиентов в отделение банка; поток покупателей, обращающихся за товарами и услугами, и т.д. В теории потоков событий справедлива теорема, аналогичная центральной предельной теореме, но в ней речь идет не о суммировании случайных величин, а о суммировании потоков событий. Рассматривается суммарный поток, составленный из большого числа независимых потоков, ни один из которых не оказывает преобладающего влияния на суммарный поток. Например, поток вызовов, поступающих на телефонную станцию, слагается из большого числа независимых потоков вызовов, исходящих от отдельных абонентов. Доказано [6], что в случае, когда характеристики потоков не зависят от времени, суммарный поток полностью описывается одним числом λ - интенсивностью потока. Для суммарного потока рассмотрим случайную величину X - длину промежутка времени между последовательными событиями. Ее функция распределения имеет вид

$$F(x;\lambda) = P(X \le x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases} \quad (10)$$

Это распределение называется экспоненциальным распределением, т.к. в формуле (10) участвует экспоненциальная функция $e^{-лх}$. Величина 1/л - масштабный параметр. Иногда вводят и параметр сдвига c, при этом экспоненциальным распределением называют распределение случайной величины $X + c$, где распределение X задается формулой (10).

В формуле (10) e = 2,718281828… - основание натуральных логарифмов. Функция экспоненциального распределения $F(x$, л) и его плотность $f(x$. л) связаны простым соотношением

$$f(x,\lambda) = \lambda(1 - F(x,\lambda)).$$

Это соотношение имеет простую интерпретацию в терминах теории надежности технических изделий и устройств. Оно означает, что интенсивность отказов (т.е. интенсивность выхода изделий из строя) постоянна, другими словами, не зависит от того, сколько времени изделие уже проработало. Обычно интенсивность отказов постоянна на основном этапе эксплуатации, после того, как на начальном этапе выявлены скрытые дефекты, и до того, как из-за естественного старения материалов начинает происходить ускоренный износ с резким возрастанием интенсивности выхода изделия из строя.

Распределения Вейбулла - Гнеденко. Экспоненциальные распределения - частный случай т. н. распределений Вейбулла - Гнеденко. Они названы по фамилиям инженера В. Вейбулла, введшего эти распределения в практику анализа результатов усталостных испытаний, и математика Б.В.Гнеденко (1912-1995), получившего такие распределения в качестве предельных при изучении максимального из результатов испытаний. Пусть X - случайная величина, характеризующая длительность функционирования изделия, сложной системы, элемента (т.е. ресурс, наработку до предельного состояния и т.п.), длительность функционирования предприятия или жизни живого существа и т.д. Важную роль играет интенсивность отказа

$$\lambda(x) = \frac{f(x)}{1 - F(x)}, \quad (11)$$

где $F(x)$ и $f(x)$ - функция распределения и плотность случайной величины X.

Опишем типичное поведение интенсивности отказа. Весь интервал времени можно разбить на три периода. На первом из них функция $л(x)$ имеет высокие значения и явную тенденцию к убыванию (чаще всего она монотонно убывает). Это можно объяснить наличием в рассматриваемой партии единиц продукции с

явными и скрытыми дефектами, которые приводят к относительно быстрому выходу из строя этих единиц продукции. Первый период называют "периодом приработки" (или "обкатки"). Именно на него обычно распространяется гарантийный срок.

Затем наступает период нормальной эксплуатации, характеризующийся приблизительно постоянной и сравнительно низкой интенсивностью отказов. Природа отказов в этот период носит внезапный характер (аварии, ошибки эксплуатационных работников и т.п.) и не зависит от длительности эксплуатации единицы продукции.

Наконец, последний период эксплуатации - период старения и износа. Природа отказов в этот период - в необратимых физико-механических и химических изменениях материалов, приводящих к прогрессирующему ухудшению качества единицы продукции и окончательному выходу ее из строя.

Каждому периоду соответствует свой вид функции *л(x)*. Рассмотрим класс степенных зависимостей

$$л(x) = л_0 b x^{b-1}, \quad (12)$$

где $л_0 > 0$ и $b > 0$ - некоторые числовые параметры. Значения $b < 1$, $b = 0$ и $b > 1$ отвечают виду интенсивности отказов в периоды приработки, нормальной эксплуатации и старения соответственно.

Соотношение (11) при заданной интенсивности отказа *л(x)* - дифференциальное уравнение относительно функции *F(x)*. Из теории дифференциальных уравнений следует, что

$$F(x) = 1 - \exp\left\{-\int_0^x \lambda(t)dt\right\}. \quad (13)$$

Подставив (12) в (13), получим, что

$$F(x) = \begin{cases} 1 - \exp[-\lambda_0 x^b], & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (14)$$

Распределение, задаваемое формулой (14) называется распределением Вейбулла - Гнеденко. Поскольку

$$\lambda_0 x^b = \left(\frac{x}{a}\right)^b,$$

где

$$a = \lambda_0^{-\frac{1}{b}}, \quad (15)$$

то из формулы (14) следует, что величина a, задаваемая формулой (15), является масштабным параметром. Иногда вводят и параметр сдвига, т.е. функциями распределения Вейбулла - Гнеденко называют $F(x - c)$, где $F(x)$ задается формулой (14) при некоторых $л_0$ и b.

Плотность распределения Вейбулла - Гнеденко имеет вид

$$f(x;a,b,c) = \begin{cases} \frac{b}{a}\left(\frac{x-c}{a}\right)^{b-1} \exp\left[-\left(\frac{x-c}{a}\right)^b\right], & x \ge c, \\ 0, & x < c, \end{cases} \quad (16)$$

где $a > 0$ - параметр масштаба, $b > 0$ - параметр формы, c - параметр сдвига. При этом параметр a из формулы (16) связан с параметром $л_0$ из формулы (14) соотношением, указанным в формуле (15).

Экспоненциальное распределение - весьма частный случай распределения Вейбулла - Гнеденко, соответствующий значению параметра формы $b = 1$.

Распределение Вейбулла - Гнеденко применяется также при построении вероятностных моделей ситуаций, в которых поведение объекта определяется "наиболее слабым звеном". Подразумевается аналогия с цепью, сохранность которой определяется тем ее звеном, которое имеет наименьшую прочность. Другими словами, пусть $X_1, X_2, \ldots, X_n$ - независимые одинаково распределенные случайные величины,

$$X(1) = \min(X_1, X_2, \ldots, X_n), X(n) = \max(X_1, X_2, \ldots, X_n).$$

В ряде прикладных задач большую роль играют *X(1)* и *X(n)*, в частности, при исследовании максимально возможных значений ("рекордов") тех или иных значений, например, страховых выплат или потерь из-за коммерческих рисков, при изучении пределов упругости и выносливости стали, ряда характеристик надежности и т.п. Показано, что при больших n распределения *X(1)* и *X(n)*, как правило, хорошо описываются распределениями Вейбулла - Гнеденко. Основополагающий вклад в изучение распределений *X(1)* и *X(n)* внес советский математик Б.В.Гнеденко. Использованию полученных результатов в экономике, менеджменте, технике и других областях посвящены труды В. Вейбулла, Э. Гумбеля, В.Б. Невзорова, Э.М. Кудлаева и многих иных специалистов.

Гамма-распределения. Перейдем к семейству гамма-распределений. Они широко применяются в экономике и менеджменте, теории и практике надежности и испытаний, в различных областях техники, метеорологии и т.д. В частности, гамма-распределению подчинены во многих ситуациях такие величины, как общий срок службы изделия, длина цепочки токопроводящих пылинок, время достижения изделием предельного состояния при коррозии, время наработки до k-го отказа, $k = 1, 2, \ldots$, и т.д. Продолжительность жизни больных хроническими заболеваниями, время достижения определенного эффекта при лечении в ряде случаев имеют гамма-распределение. Это распределение наиболее адекватно для описания спроса в экономико-математических моделях управления запасами (логистики).

Плотность гамма-распределения имеет вид

$$f(x;a,b,c)=\begin{cases}\dfrac{1}{\Gamma(a)}(x-c)^{a-1}b^{-a}\exp\left[-\dfrac{x-c}{b}\right], & x\ge c,\\ 0, & x<c.\end{cases} \qquad (17)$$

Плотность вероятности в формуле (17) определяется тремя параметрами *a, b, c*, где *a*>0, *b*>0. При этом *a* является параметром формы, *b* - параметром масштаба и *c* - параметром сдвига. Множитель *1/Г(a)* является нормировочным, он введен, чтобы

$$\int_{-\infty}^{+\infty} f(x;a,b,c)dx = 1.$$

Здесь *Г(a)* - одна из используемых в математике специальных функций, так называемая "гамма-функция", по которой названо и распределение, задаваемое формулой (17),

$$\Gamma(a) = \int_{0}^{+\infty} x^{a-1}e^{-x}dx.$$

При фиксированном *a* формула (17) задает масштабно-сдвиговое семейство распределений, порождаемое распределением с плотностью

$$f(x;a) = \begin{cases} \dfrac{1}{\Gamma(a)} x^{a-1}e^{-x}, & x \ge 0, \\ 0, & x < 0. \end{cases} \qquad (18)$$

Распределение вида (18) называется стандартным гамма-распределением. Оно получается из формулы (17) при *b* = 1 и *c* = 0.

Частным случаем гамма-распределений при *a* = 1 являются экспоненциальные распределения (с *л = 1/b*). При натуральном *a* и *c*=0 гамма-распределения называются распределениями Эрланга. С работ датского ученого К.А.Эрланга (1878-1929), сотрудника Копенгагенской телефонной компании, изучавшего в 1908-1922 гг. функционирование телефонных сетей, началось развитие теории массового обслуживания. Эта теория занимается вероятностно-статистическим моделированием систем, в которых происходит обслуживание потока заявок, с целью принятия оптимальных решений. Распределения Эрланга используют в тех же прикладных областях, в которых применяют экспоненциальные распределения. Это основано на следующем математическом факте: сумма k

независимых случайных величин, экспоненциально распределенных с одинаковыми параметрами л и *c*, имеет гамма-распределение с параметром формы *a* = *k*, параметром масштаба *b* = 1/л и параметром сдвига *kc*. При *c* = 0 получаем распределение Эрланга.

Если случайная величина *X* имеет гамма-распределение с параметром формы *a* таким, что *d* = *2a* - целое число, *b* = 1 и *c* = 0, то 2*X* имеет распределение хи-квадрат с *d* степенями свободы.

Случайная величина *X* с гвмма-распределением имеет следующие характеристики:

- математическое ожидание $M(X) = ab + c$,
- дисперсию $D(X) = y^2 = ab^2$,
- коэффициент вариации $v = \frac{b\sqrt{a}}{ab + c}$,
- асимметрию $M[(X - M(X))^3] = \frac{2}{\sqrt{a}}$,
- эксцесс $\frac{M[(X - M(X))^4]}{\sigma^4} - 3 = \frac{6}{a}$.

Нормальное распределение - предельный случай гамма-распределения. Точнее, пусть Z - случайная величина, имеющая стандартное гамма-распределение, заданное формулой (18). Тогда

$$\lim_{a \to \infty} P\left\{\frac{Z - a}{\sqrt{a}} < x\right\} = \Phi(x)$$

для любого действительного числа *x*, где *Ф(x)* - функция стандартного нормального распределения *N*(0,1).

В прикладных исследованиях используются и другие параметрические семейства распределений, из которых наиболее известны система кривых Пирсона, ряды Эджворта и Шарлье. Здесь они не рассматриваются.

Дискретные распределения, используемые в вероятностно-статистических методах. Наиболее часто используют три семейства дискретных распределений -

биномиальных, гипергеометрических и Пуассона, а также некоторые другие семейства - геометрических, отрицательных биномиальных, мультиномиальных, отрицательных гипергеометрических и т.д.

Подробнее о биномиальном распределении. Как уже говорилось, биномиальное распределение имеет место при независимых испытаниях, в каждом из которых с вероятностью p появляется событие A. Если общее число испытаний n задано, то число испытаний Y, в которых появилось событие A, имеет биномиальное распределение. Для биномиального распределения вероятность принятия случайной величиной Y значения y определяется формулой

$$P(Y = y \mid p, n) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,2,\ldots,n, \quad (19)$$

где

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} = C_n^y -$$

- число сочетаний из n элементов по y, известное из комбинаторики. Для всех y, кроме 0, 1, 2, …, n, имеем $P(Y=y)=0$. Биномиальное распределение при фиксированном объеме выборки n задается параметром p, т.е. биномиальные распределения образуют однопараметрическое семейство. Они применяются при анализе данных выборочных исследований [2], в частности, при изучении предпочтений потребителей, выборочном контроле качества продукции по планам одноступенчатого контроля, при испытаниях совокупностей индивидуумов в демографии, социологии, медицине, биологии и др.

Если Y_1 и Y_2 - независимые биномиальные случайные величины с одним и тем же параметром p_0, определенные по выборкам с объемами n_1 и n_2 соответственно, то $Y_1 + Y_2$ - биномиальная случайная величина, имеющая распределение (19) с p

= p_0 и $n = n_1 + n_2$. Это замечание расширяет область применимости биномиального распределения, позволяя объединять результаты нескольких групп испытаний, когда есть основания полагать, что всем этим группам соответствует один и тот же параметр.

Характеристики биномиального распределения вычислены ранее:

$$M(Y) = np,\ \ D(Y) = np(1-p).$$

В главе "События и вероятности" для биномиальной случайной величины доказан закон больших чисел:

$$\lim_{n\to\infty} P\left\{\left|\frac{Y}{n} - p\right| \geq \varepsilon\right\} = 0$$

для любого $\varepsilon > 0$. С помощью центральной предельной теоремы закон больших чисел можно уточнить, указав, насколько *Y/n* отличается от *p*.

Теорема Муавра-Лапласа. Для любых чисел a и *b*, *a*<*b*, имеем

$$\lim_{n\to\infty} P\left\{a \leq \frac{Y - np}{\sqrt{np(1-p)}} < b\right\} = \Phi(b) - \Phi(a),$$

где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

Для доказательства достаточно воспользоваться представлением *Y* в виде суммы независимых случайных величин, соответствующих исходам отдельных испытаний, формулами для *M(Y)* и *D(Y)* и центральной предельной теоремой.

Эта теорема для случая *p* = S доказана английским математиком А.Муавром (1667-1754) в 1730 г. В приведенной выше формулировке она была доказана в 1810 г. французским математиком Пьером Симоном Лапласом (1749 – 1827).

Гипергеометрическое распределение имеет место при выборочном контроле конечной совокупности объектов объема N по альтернативному признаку. Каждый контролируемый объект классифицируется либо как обладающий признаком *A*, либо как не

обладающий этим признаком. Гипергеометрическое распределение имеет случайная величина *Y*, равная числу объектов, обладающих признаком *A* в случайной выборке объема *n*, где *n*<*N*. Например, число *Y* дефектных единиц продукции в случайной выборке объема *n* из партии объема *N* имеет гипергеометрическое распределение, если *n*<*N*. Другой пример – лотерея. Пусть признак *A* билета – это признак «быть выигрышным». Пусть всего билетов *N,* а некоторое лицо приобрело *n* из них. Тогда число выигрышных билетов у этого лица имеет гипергеометрическое распределение.

Для гипергеометрического распределения вероятность принятия случайной величиной Y значения y имеет вид

$$P(Y = y \mid N, d, n) = \frac{\binom{n}{y}\binom{N-n}{D-y}}{\binom{N}{D}}, \quad (20)$$

где *D* – число объектов, обладающих признаком *A*, в рассматриваемой совокупности объема *N*. При этом *y* принимает значения от $\max\{0,\ n - (N - D)\}$ до $\min\{n,\ D\}$, при прочих *y* вероятность в формуле (20) равна 0. Таким образом, гипергеометрическое распределение определяется тремя параметрами – объемом генеральной совокупности *N,* числом объектов *D* в ней, обладающих рассматриваемым признаком *A*, и объемом выборки *n.*

Простой случайной выборкой объема *n* из совокупности объема *N* называется выборка, полученная в результате случайного отбора, при котором любой из $\binom{N}{n}$ наборов из *n* объектов имеет одну и ту же вероятность быть отобранным. Методы случайного отбора выборок респондентов (опрашиваемых) или единиц штучной продукции рассматриваются в инструктивно-методических и нормативно-технических документах. Один из

методов отбора таков: объекты отбирают один из другим, причем на каждом шаге каждый из оставшихся в совокупности объектов имеет одинаковые шансы быть отобранным. В литературе для рассматриваемого типа выборок используются также термины «случайная выборка», «случайная выборка без возвращения».

Поскольку объемы генеральной совокупности (партии) N и выборки n обычно известны, то подлежащим оцениванию параметром гипергеометрического распределения является D. В статистических методах управления качеством продукции D – обычно число дефектных единиц продукции в партии. Представляет интерес также характеристика распределения D/N – уровень дефектности.

Для гипергеометрического распределения

$$M(Y) = n\frac{D}{N}, \quad D(Y) = n\frac{D}{N}\left(1-\frac{D}{N}\right)\left((1-\frac{n-1}{N-1}\right).$$

Последний множитель в выражении для дисперсии близок к 1, если $N>10n$. Если при этом сделать замену $p = D/N$, то выражения для математического ожидания и дисперсии гипергеометрического распределения перейдут в выражения для математического ожидания и дисперсии биномиального распределения. Это не случайно. Можно показать, что

$$P(Y = y \mid N, d, n) = \frac{\binom{n}{y}\binom{N-n}{D-y}}{\binom{N}{D}} \approx P(Y = y \mid p, n) = \binom{n}{y} p^{y}(1-p)^{n-y}, \;\; y = 0,1,2,...,n,$$

при $N>10n$, где $p = D/N$. Точнее, справедливо предельное соотношение

$$\lim_{N\to\infty, \frac{D}{N}\to p} P(Y = y \mid N, d, n) = P(Y = y \mid p, n), \;\; y = 0,1,2,...,n,$$

и этим предельным соотношением можно пользоваться при $N>10n$.

Распределение Пуассона. Третье широко используемое дискретное распределение – распределение Пуассона. Случайная величина Y имеет распределение Пуассона, если

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0,1,2,\ldots,$$

где л – параметр распределения Пуассона, и $P(Y=y)=0$ для всех прочих y (при y=0 обозначено 0! =1). Для распределения Пуассона

$$M(Y) = \text{л}, \; D(Y) = \text{л}.$$

Это распределение названо в честь французского математика С.Д.Пуассона (1781-1840), впервые получившего его в 1837 г. Распределение Пуассона является предельным случаем биномиального распределения, когда вероятность p осуществления события мала, но число испытаний n велико, причем np = л. Точнее, справедливо предельное соотношение

$$\lim_{n\to\infty, np\to\lambda} P(Y = y \mid p, n) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0,1,2,\ldots$$

Поэтому распределение Пуассона (в старой терминологии «закон распределения») часто называют также «законом редких событий».

Распределение Пуассона возникает в теории потоков событий (см. выше). Доказано, что для простейшего потока с постоянной интенсивностью Л число событий (вызовов), происшедших за время t, имеет распределение Пуассона с параметром л = Лt. Следовательно, вероятность того, что за время t не произойдет ни одного события, равна $e^{-\text{Л}t}$, т.е. функция распределения длины промежутка между событиями является экспоненциальной.

Распределение Пуассона используется при анализе результатов выборочных маркетинговых обследований потребителей, расчете оперативных характеристик планов статистического приемочного контроля в случае малых значений приемочного уровня дефектности, для описания числа разладок статистически управляемого технологического процесса в единицу

времени, числа «требований на обслуживание», поступающих в единицу времени в систему массового обслуживания, статистических закономерностей несчастных случаев и редких заболеваний, и т.д.

Описание иных параметрических семейств дискретных распределений и возможности их практического использования рассматриваются в обширной (более миллиона названий статей и книг на десятках языков) литературе по вероятностно-статистическим методам.

5. Основные проблемы прикладной статистики - описание данных, оценивание и проверка гипотез

Выделяют три основные области статистических методов обработки результатов наблюдений – описание данных, оценивание (характеристик и параметров распределений, регрессионных зависимостей и др.) и проверка статистических гипотез. Рассмотрим основные понятия, применяемые в этих областях.

Основные понятия, используемые при описании данных. Описание данных – предварительный этап статистической обработки. Используемые при описании данных величины применяются при дальнейших этапах статистического анализа – оценивании и проверке гипотез, а также при решении иных задач, возникающих при применении вероятностно-статистических методов принятия решений, например, при статистическом контроле качества продукции и статистическом регулировании технологических процессов.

Статистические данные – это результаты наблюдений (измерений, испытаний, опытов, анализов). Функции результатов наблюдений, используемые, в частности, для оценки параметров распределений и (или) для проверки статистических гипотез,

называют «статистиками». (Для математиков надо добавить, что речь идет об измеримых функциях.) Если в вероятностной модели результаты наблюдений рассматриваются как случайные величины (или случайные элементы), то статистики, как функции случайных величин (элементов), сами являются случайными величинами (элементами). Статистики, являющиеся выборочными аналогами характеристик случайных величин (математического ожидания, медианы, дисперсии, моментов и др.) и используемые для оценивания этих характеристик, называют статистическими характеристиками.

Виды выборок. Основополагающее понятие в вероятностно-статистических методах принятия решений – выборка. Как уже говорилось, выборка – это

1) набор наблюдаемых значений или

2) множество объектов, отобранные из изучаемой совокупности.

Например, единицы продукции, отобранные из контролируемой партии или потока продукции для контроля и принятия решений. Наблюдаемые значения обозначим $x_1, x_2, \ldots, x_n$, где n – объем выборки, т.е. число наблюдаемых значений, составляющих выборку. О втором виде выборок уже шла речь при рассмотрении гипергеометрического распределения, когда под выборкой понимался набор единиц продукции, отобранных из партии. Там же обсуждалась вероятностная модель случайной выборки.

В вероятностной модели выборки первого вида наблюдаемые значения обычно рассматривают как реализацию независимых одинаково распределенных случайных величин $X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \omega \in \Omega$. При этом считают, что полученные при наблюдениях конкретные значения $x_1, x_2, \ldots, x_n$ соответствуют определенному элементарному событию $\omega = \omega_0$, т.е.

$$x_1 = X_1(\omega_0), x_2 = X_2(\omega_0), ..., x_n = X_n(\omega_0), \omega_0 \in \Omega .$$

При повторных наблюдениях будут получены иные наблюдаемые значения, соответствующие другому элементарному событию $\omega = \omega_1$. Цель обработки статистических данных состоит в том, чтобы по результатам наблюдений, соответствующим элементарному событию $\omega = \omega_0$, сделать выводы о вероятностной мере P и результатах наблюдений при различных возможных $\omega = \omega_1$.

Применяют и другие, более сложные вероятностные модели выборок. Например, цензурированные выборки соответствуют испытаниям, проводящимся в течение определенного промежутка времени. При этом для части изделий удается замерить время наработки на отказ, а для остальных лишь констатируется, что наработки на отказ для них больше времени испытания. Для выборок второго вида отбор объектов может проводиться в несколько этапов. Например, для входного контроля сигарет могут сначала отбираться коробки, в отобранных коробках – блоки, в выбранных блоках – пачки, а в пачках – сигареты. Четыре ступени отбора. Ясно, что выборка будет обладать иными свойствами, чем простая случайная выборка из совокупности сигарет.

Частоты. Из приведенного выше определения математической статистики следует, что описание статистических данных дается с помощью частот. Частота – это отношение числа X наблюдаемых единиц, которые принимают заданное значение или лежат в заданном интервале, к общему числу наблюдений n, т.е. частота – это X/n. (В более старой литературе иногда X/n называется относительной частотой, а под частотой имеется в виду X. В старой терминологии можно сказать, что относительная частота – это отношение частоты к общему числу наблюдений.)

Отметим, что обсуждаемое определение приспособлено к нуждам одномерной статистики. В случае многомерного статистического анализа, статистики случайных процессов и временных рядов, статистики объектов нечисловой природы нужны несколько иные определения понятия «статистические данные». Не считая нужным давать такие определения, отметим, что в подавляющем большинстве практических постановок исходные статистические данные – это выборка или несколько выборок. А выборка – это конечная совокупность соответствующих математических объектов (чисел, векторов, функций, объектов нечисловой природы).

Число X имеет биномиальное распределение, задаваемое вероятностью p того, что случайная величина, с помощью которой моделируются результаты наблюдений, принимает заданное значение или лежит в заданном интервале, и общим числом наблюдений n. Из закона больших чисел (теорема Бернулли) следует, что

$$\frac{X}{n} \to p$$

при $n \to \infty$ (сходимость по вероятности), т.е. частота сходится к вероятности. Теорема Муавра-Лапласа позволяет уточнить скорость сходимости в этом предельном соотношении.

Эмпирическая функция распределения. Чтобы от отдельных событий перейти к одновременному рассмотрению многих событий, используют накопленную частоту. Так называется отношение числа единиц, для которых результаты наблюдения меньше заданного значения, к общему числу наблюдений. (Это понятие используется, если результаты наблюдения – действительные числа, а не вектора, функции или объекты нечисловой природы.) Функция, которая выражает зависимость между значениями количественного признака и накопленной

частотой, называется эмпирической функцией распределения. Итак, эмпирической функцией распределения $F_n(x)$ называется доля элементов выборки, меньших x. Эмпирическая функция распределения содержит всю информацию о результатах наблюдений.

Чтобы записать выражение для эмпирической функции распределения в виде формулы, введем функцию $c(x, y)$ двух переменных:

$$c(x,y)=\begin{cases}0, & x \le y,\\ 1, & x > y.\end{cases}$$

Случайные величины, моделирующие результаты наблюдений, обозначим $X_1(\omega), X_2(\omega),..., X_n(\omega), \omega \in \Omega$. Тогда эмпирическая функция распределения $F_n(x)$ имеет вид

$$F_n(x) = F_n(x,\omega) = \frac{1}{n}\sum_{1\le i\le n} c(x, X_i(\omega)).$$

Из закона больших чисел следует, что для каждого действительного числа x эмпирическая функция распределения $F_n(x)$ сходится к функции распределения $F(x)$ результатов наблюдений, т.е.

$$F_n(x) \to F(x) \quad (1)$$

при $n \to \infty$. Советский математик В.И. Гливенко (1897-1940) доказал в 1933 г. более сильное утверждение: сходимость в (1) равномерна по x, т.е.

$$\sup_x | F_n(x) - F(x) | \to 0 \quad (2)$$

при $n \to \infty$ (сходимость по вероятности).

В (2) использовано обозначение sup (читается как «супремум»). Для функции $g(x)$ под $\sup_x g(x)$ понимают наименьшее из чисел a таких, что $g(x) \le a$ при всех x. Если функция $g(x)$ достигает максимума в точке x_0, то $\sup_x g(x) = g(x_0)$. В таком случае

вместо sup пишут max. Хорошо известно, что не все функции достигают максимума.

В том же 1933 г. А.Н.Колмогоров усилил результат В.И. Гливенко для непрерывных функций распределения *F(x)*. Рассмотрим случайную величину

$$D_n = \sqrt{n} \sup_x | F_n(x) - F_0(x) |$$

и ее функцию распределения

$$K_n(x) = P\{D_n \le x\}.$$

По теореме А.Н.Колмогорова

$$\lim_{n\to\infty} K_n(x) = K(x)$$

при каждом *x*, где *K(x)* – т.н. функция распределения Колмогорова.

Рассматриваемая работа А.Н. Колмогорова породила одно из основных направлений математической статистики – т.н. непараметрическую статистику. И в настоящее время непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат широко используются. Они были разработаны для проверки согласия с *полностью известным* теоретическим распределением, т.е. предназначены для проверки гипотезы $H_0 : F(x) \equiv F_0(x)$. Основная идея критериев Колмогорова, омега-квадрат и аналогичных им состоит в измерении расстояния между функцией эмпирического распределения и функцией теоретического распределения. Различаются эти критерии видом расстояний в пространстве функций распределения. Аналитические выражения для предельных распределений статистик, расчетные формулы, таблицы распределений и критических значений широко распространены [8], поэтому не будем их приводить.

Выборочные характеристики распределения. Кроме эмпирической функции распределения, для описания данных используют и другие статистические характеристики. В качестве выборочных средних величин постоянно используют выборочное

среднее арифметическое, т.е. сумму значений рассматриваемой величины, полученных по результатам испытания выборки, деленную на ее объем:

$$\bar{x} = \frac{1}{n} \sum_{1 \le i \le n} x_i ,$$

где n – объем выборки, x_i – результат измерения (испытания) i-ого элемента выборки.

Другой вид выборочного среднего – выборочная медиана. Она определяется через порядковые статистики.

Порядковые статистики – это члены вариационного ряда, который получается, если элементы выборки $x_1, x_2, \ldots, x_n$ расположить в порядке неубывания:

$$x(1) \le x(2) \le \ldots \le x(k) \le \ldots \le x(n).$$

Пример 1. Для выборки $x_1 = 1$, $x_2 = 7$, $x_3 = 4$, $x_4 = 2$, $x_5 = 8$, $x_6 = 0$, $x_7 = 5$, $x_8 = 7$ вариационный ряд имеет вид 0, 1, 2, 4, 5, 7, 7, 8, т.е. $x(1) = 0 = x_6$, $x(2) = 1 = x_1$, $x(3) = 2 = x_4$, $x(4) = 4 = x_3$, $x(5) = 5 = x_7$, $x(6) = x(7) = 7 = x_2 = x_8$, $x(8) = 8 = x_5$.

В вариационном ряду элемент $x(k)$ называется k-той порядковой статистикой. Порядковые статистики и функции от них широко используются в вероятностно-статистических методах принятия решений, в эконометрике и в других прикладных областях [2].

Выборочная медиана $\tilde{x}$ - результат наблюдения, занимающий центральное место в вариационном ряду, построенном по выборке с нечетным числом элементов, или полусумма двух результатов наблюдений, занимающих два центральных места в вариационном ряду, построенном по выборке с четным числом элементов. Таким образом, если объем выборки n – нечетное число, $n = 2k+1$, то медиана $\tilde{x} = x(k+1)$, если же n – четное число, $n = 2k$, то медиана $\tilde{x} = [x(k) + x(k+1)]/2$, где $x(k)$ и $x(k+1)$ – порядковые статистики.

В качестве выборочных показателей рассеивания результатов наблюдений чаще всего используют выборочную дисперсию, выборочное среднее квадратическое отклонение и размах выборки.

Согласно [8] выборочная дисперсия s^2 – это сумма квадратов отклонений выборочных результатов наблюдений от их среднего арифметического, деленная на объем выборки:

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

Выборочное среднее квадратическое отклонение s – неотрицательный квадратный корень из дисперсии, т.е. $s = +\sqrt{s^2}$.

В некоторых литературных источниках выборочной дисперсией называют другую величину:

$$s_0^2 = \frac{1}{n-1} \sum_{1 \le i \le n} (x_i - \bar{x})^2.$$

Она отличается от s^2 постоянным множителем:

$$s^2 = \left(1 - \frac{1}{n}\right) s_0^2.$$

Соответственно выборочным средним квадратическим отклонением в этих литературных источниках называют величину $s_0 = +\sqrt{s_0^2}$. Тогда, очевидно,

$$s = \sqrt{1 - \frac{1}{n}} s_0.$$

Различие в определениях приводит к различию в алгоритмах расчетов, правилах принятия решений и соответствующих таблицах. Поэтому при использовании тех или иных нормативно-технических и инструктивно-методических материалов, программных продуктов, таблиц необходимо обращать внимание на способ определения выборочных характеристик.

Выбор s_0^2, а не s^2, объясняется тем, что

$$M(s_0^2) = D(X) = \sigma^2,$$

где X – случайная величина, имеющая такое же распределение, как и результаты наблюдений. В терминах теории статистического оценивания это означает, что s_0^2 - несмещенная оценка дисперсии (см. ниже). В то же время статистика s^2 не является несмещенной оценкой дисперсии результатов наблюдений, поскольку

$$M(s^2) = \left(1 - \frac{1}{n}\right)\sigma^2.$$

Однако у s^2 есть другое свойство, оправдывающее использование этой статистики в качестве выборочного показателя рассеивания. Для известных результатов наблюдений $x_1, x_2, \ldots, x_n$ рассмотрим случайную величину Y с распределением вероятностей

$$P(Y = x_i) = \frac{1}{n}, \quad i = 1,2,\ldots,n,$$

и $P(Y = x) = 0$ для всех прочих x. Это распределение вероятностей называется эмпирическим. Тогда функция распределения Y – это эмпирическая функция распределения, построенная по результатам наблюдений $x_1, x_2, \ldots, x_n$. Вычислим математическое ожидание и дисперсию случайной величины Y:

$$M(Y) = \bar{x}, \quad D(Y) = s^2.$$

Второе из этих равенств и является основанием для использования s^2 в качестве выборочного показателя рассеивания.

Отметим, что математические ожидания выборочных средних квадратических отклонений $M(s)$ и $M(s_0)$, вообще говоря, не равняются теоретическому среднему квадратическому отклонению у. Например, если X имеет нормальное распределение, объем выборки $n = 3$, то

$$M(s) = 0{,}724\sigma, \quad M(s_0) = 0{,}887\sigma.$$

Кроме перечисленных выше статистических характеристик, в качестве выборочного показателя рассеивания используют размах R – разность между n-й и первой порядковыми статистиками в

выборке объема n, т.е. разность между наибольшим и наименьшим значениями в выборке: $R = x(n) - x(1)$.

В ряде вероятностно-статистических методов применяют и иные показатели рассеивания. В частности, в методах статистического регулирования процессов используют средний размах – среднее арифметическое размахов, полученных в определенном количестве выборок одинакового объема. Популярно и межквартильное расстояние, т.е. расстояние между выборочными квартилями $x([0,75n])$ и $x([0,25n])$ порядка 0,75 и 0,25 соответственно, где $[0,75n]$ – целая часть числа $0,75n$, а $[0,25n]$ – целая часть числа $0,25n$.

Основные понятия, используемые при оценивании. Оценивание – это определение приближенного значения неизвестной характеристики или параметра распределения (генеральной совокупности), иной оцениваемой составляющей математической модели реального (экономического, технического и др.) явления или процесса по результатам наблюдений. Иногда формулируют более коротко: оценивание – это определение приближенного значения неизвестного параметра генеральной совокупности по результатам наблюдений. При этом параметром генеральной совокупности может быть либо число, либо набор чисел (вектор), либо функция, либо множество или иной объект нечисловой природы. Например, по результатам наблюдений, распределенных согласно биномиальному закону, оценивают число – параметр p (вероятность успеха). По результатам наблюдений, имеющих гамма-распределение, оценивают набор из трех чисел – параметры формы a, масштаба b и сдвига c. Способ оценивания функции распределения дается теоремами В.И. Гливенко и А.Н. Колмогорова. Оценивают также плотности вероятности, функции, выражающие зависимости между переменными, включенными в вероятностные модели экономических, управленческих или

технологических процессов, и т.д. Целью оценивания может быть нахождение упорядочения инвестиционных проектов по экономической эффективности или технических изделий (объектов) по качеству, формулировка правил технической или медицинской диагностики и т.д. (Упорядочения в математической статистике называют также ранжировками. Это – один из видов объектов нечисловой природы.)

Оценивание проводят с помощью оценок – статистик, являющихся основой для оценивания неизвестного параметра распределения. В ряде литературных источников термин «оценка» встречается в качестве синонима термина «оценивание». Употреблять одно и то же слово для обозначения двух разных понятий нецелесообразно: оценивание – это действие, а оценка – статистика (функция от результатов наблюдений), используемая в процессе указанного действия или являющаяся его результатом.

Оценивание бывает двух видов – точечное оценивание и оценивание с помощью доверительной области.

Точечное оценивание - способ оценивания, заключающийся в том, что значение оценки принимается как неизвестное значение параметра распределения.

Пример 2. Пусть результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают в вероятностной модели как случайную выборку из нормального распределения $N(m,y)$. Т.е. считают, что результаты наблюдений моделируются как реализации n независимых одинаково распределенных случайных величин, имеющих функцию нормального распределения $N(m,y)$ с некоторыми математическим ожиданием m и средним квадратическим отклонением y, неизвестными статистику. Требуется оценить параметры m и y (или y^2) по результатам наблюдений. Оценки обозначим m^* и $(y^2)^*$ соответственно. Обычно в качестве оценки m^* математического ожидания m используют выборочное среднее арифметическое $\bar{x}$, а

в качестве оценки $(y^2)^*$ дисперсии y^2 используют выборочную дисперсию s^2, т.е.

$$m^* = \bar{x}\,,\ (y^2)^* = s^2.$$

Для оценивания математического ожидания m могут использоваться и другие статистики, например, выборочная медиана $\tilde{x}$, полусумма минимального и максимального членов вариационного ряда

$$m^{**} = [x(1)+x(n)]/2$$

и др. Для оценивания дисперсии y^2 также имеется ряд оценок, в частности, s_0^2 (см. выше) и оценка, основанная на размахе R, имеющая вид

$$(y^2)^{**} = [a(n)R]^2,$$

где коэффициенты $a(n)$ берут из специальных таблиц [8]. Эти коэффициенты подобраны так, чтобы для выборок из нормального распределения

$$M[a(n)R] = y.$$

Наличие нескольких методов оценивания одних и тех же параметров приводит к необходимости выбора между этими методами.

Состоятельность, несмещенность и эффективность оценок. Как сравнивать методы оценивания между собой? Сравнение проводят на основе таких показателей качества методов оценивания, как состоятельность, несмещенность, эффективность и др.

Рассмотрим оценку $и_n$ числового параметра и, определенную при n = 1, 2, … Оценка $и_n$ называется *состоятельной*, если она сходится по вероятности к значению оцениваемого параметра и при безграничном возрастании объема выборки. Выразим сказанное более подробно. Статистика $и_n$ является состоятельной оценкой

параметра и тогда и только тогда, когда для любого положительного числа e справедливо предельное соотношение

$$\lim_{n\to\infty} P\{|\theta_n - \theta| > \varepsilon\} = 0.$$

Пример 3. Из закона больших чисел следует, что и$_n$ = $\bar{x}$ является состоятельной оценкой и = *M(X)* (в приведенной выше теореме Чебышёва предполагалось существование дисперсии *D(X);* однако, как доказал А.Я. Хинчин [6], достаточно выполнения более слабого условия – существования математического ожидания *M(X)*).

Пример 4. Все указанные выше оценки параметров нормального распределения являются состоятельными.

Вообще, все (за редчайшими исключениями) оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются состоятельными.

Пример 5. Так, согласно теореме В.И. Гливенко, эмпирическая функция распределения $F_n(x)$ является состоятельной оценкой функции распределения результатов наблюдений *F(x)*.

При разработке новых методов оценивания следует в первую очередь проверять состоятельность предлагаемых методов.

Второе важное свойство оценок – *несмещенность*. Несмещенная оценка и$_n$ – это оценка параметра и, математическое ожидание которой равно значению оцениваемого параметра: *M*(и$_n$) = и.

Пример 6. Из приведенных выше результатов следует, что $\bar{x}$ и s_0^2 являются несмещенными оценками параметров *m* и y^2 нормального распределения. Поскольку M($\tilde{x}$) = M(*m***) = *m*, то выборочная медиана $\tilde{x}$ и полусумма крайних членов вариационного ряда *m*** - также несмещенные оценки математического ожидания *m* нормального распределения. Однако

$$M(s^2) \neq \sigma^2, \quad M[(\sigma^2)^{**}] \neq \sigma^2,$$

поэтому оценки s^2 и (y^2)** не являются состоятельными оценками дисперсии y^2 нормального распределения.

Оценки, для которых соотношение $M(и_n) = и$ неверно, называются смещенными. При этом разность между математическим ожиданием оценки $и_n$ и оцениваемым параметром и, т.е. $M(и_n) - и$, называется смещением оценки.

Пример 7. Для оценки s^2, как следует из сказанного выше, смещение равно

$$M(s^2) - y^2 = - y^2/n.$$

Смещение оценки s^2 стремится к 0 при $n \to \infty$.

Оценка, для которой смещение стремится к 0, когда объем выборки стремится к бесконечности, называется *асимптотически несмещенной*. В примере 7 показано, что оценка s^2 является асимптотически несмещенной.

Практически все оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются либо несмещенными, либо асимптотически несмещенными. Для несмещенных оценок показателем точности оценки служит дисперсия – чем дисперсия меньше, тем оценка лучше. Для смещенных оценок показателем точности служит математическое ожидание квадрата оценки $M(и_n - и)^2$. Как следует из основных свойств математического ожидания и дисперсии,

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] = D(\theta_n) + (M(\theta_n) - \theta)^2, \quad (3)$$

т.е. математическое ожидание квадрата ошибки складывается из дисперсии оценки и квадрата ее смещения.

Для подавляющего большинства оценок параметров, используемых в вероятностно-статистических методах принятия решений, дисперсия имеет порядок $1/n$, а смещение – не более чем $1/n$, где n – объем выборки. Для таких оценок при больших n второе

слагаемое в правой части (3) пренебрежимо мало по сравнению с первым, и для них справедливо приближенное равенство

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] \approx D(\theta_n) \approx \frac{c}{n}, \quad c = c(\theta_n, \theta), \quad (4)$$

где c – число, определяемое методом вычисления оценок и$_n$ и истинным значением оцениваемого параметра и.

С дисперсией оценки связано третье важное свойство метода оценивания – *эффективность*. Эффективная оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра.

Доказано [11], что $\bar{x}$ и s_0^2 являются эффективными оценками параметров m и y^2 нормального распределения. В то же время для выборочной медианы $\tilde{x}$ справедливо предельное соотношение

$$\lim_{n\to\infty} \frac{D(\bar{x})}{D(\tilde{x})} = \frac{2}{\pi} \approx 0{,}637.$$

Другими словами, эффективность выборочной медианы, т.е. отношение дисперсии эффективной оценки $\bar{x}$ параметра m к дисперсии несмещенной оценки $\tilde{x}$ этого параметра при больших n близка к 0,637. Именно из-за сравнительно низкой эффективности выборочной медианы в качестве оценки математического ожидания нормального распределения обычно используют выборочное среднее арифметическое.

Понятие эффективности вводится для несмещенных оценок, для которых M(и$_n$) = и для всех возможных значений параметра и. Если не требовать несмещенности, то можно указать оценки, при некоторых и имеющие меньшую дисперсию и средний квадрат ошибки, чем эффективные.

Пример 8. Рассмотрим «оценку» математического ожидания $m_1 \equiv 0$. Тогда $D(m_1) = 0$, т.е. всегда меньше дисперсии $D(\bar{x})$ эффективной оценки $\bar{x}$. Математическое ожидание среднего квадрата ошибки $d_n(m_1) = m^2$, т.е. при $|m| < \sigma/\sqrt{n}$ имеем $d_n(m_1) <$

$d_n(\bar{x})$. Ясно, однако, что статистику $m_1 \equiv 0$ бессмысленно рассматривать в качестве оценки математического ожидания m.

Пример 9. Более интересный пример рассмотрен американским математиком Дж. Ходжесом:

$$T_n = \begin{cases} \bar{x}, & |\bar{x}| > n^{-1/4}, \\ 0{,}5\bar{x}, & |\bar{x}| \le n^{-1/4}. \end{cases}$$

Ясно, что T_n – состоятельная, асимптотически несмещенная оценка математического ожидания m, при этом, как нетрудно вычислить,

$$\lim_{n\to\infty} n d_n(T_n) = \begin{cases} \sigma^2, & m \neq 0, \\ \dfrac{\sigma^2}{4}, & m = 0. \end{cases}$$

Последняя формула показывает, что при $m \neq 0$ оценка T_n не хуже $\bar{x}$ (при сравнении по среднему квадрату ошибки d_n), а при $m = 0$ – в четыре раза лучше.

Подавляющее большинство оценок $и_n$, используемых в вероятностно-статистических методах, являются асимптотически нормальными, т.е. для них справедливы предельные соотношения:

$$\lim_{n\to\infty} P\left\{ \frac{\theta_n - M(\theta_n)}{\sqrt{D(\theta_n)}} < x \right\} = \Phi(x)$$

для любого x, где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Это означает, что для больших объемов выборок (практически - несколько десятков или сотен наблюдений) распределения оценок полностью описываются их математическими ожиданиями и дисперсиями, а качество оценок – значениями средних квадратов ошибок $d_n(и_n)$.

Наилучшие асимптотически нормальные оценки, сокращенно НАН - оценки, - это оценки, для которых средний квадрат ошибки $d_n(и_n)$ принимает при больших объемах выборки наименьшее возможное значение, т.е. величина $c = c(и_n, и)$ в формуле (4) минимальна. Ряд видов оценок – так называемые

одношаговые оценки и оценки максимального правдоподобия – являются НАН - оценками, именно они обычно используются в вероятностно-статистических методах принятия решений.

Доверительное оценивание. Какова точность оценки параметра? В каких границах он может лежать? В научных публикациях и учебной литературе, в нормативно-технической и инструктивно-методической документации, в таблицах и программных продуктах наряду с алгоритмами расчетов точечных оценок даются правила нахождения доверительных границ. Они и указывают точность точечной оценки. При этом используются такие термины, как доверительная вероятность, доверительный интервал. Если речь идет об оценивании нескольких числовых параметров, или же функции, упорядочения и т.п., то говорят об оценивании с помощью доверительной области.

Доверительная область – это область в пространстве параметров, в которую с заданной вероятностью входит неизвестное значение оцениваемого параметра распределения. «Заданная вероятность» называется *доверительной вероятностью* и обычно обозначается г. Пусть И – пространство параметров. Рассмотрим статистику $И_1 = И_1(x_1, x_2, \ldots, x_n)$ – функцию от результатов наблюдений $x_1, x_2, \ldots, x_n$, значениями которой являются подмножества пространства параметров И. Так как результаты наблюдений – случайные величины, то $И_1$ – также случайная величина, значения которой – подмножества множества И, т.е. $И_1$ – случайное множество. Напомним, что множество – один из видов объектов нечисловой природы, случайные множества изучают в теории вероятностей и статистике объектов нечисловой природы.

В ряде литературных источников, к настоящему времени во многом устаревших, под случайными величинами понимают только те из них, которые в качестве значений принимают действительные числа. Согласно справочнику академика РАН Ю.В.Прохорова и

проф. Ю.А.Розанова [12] случайные величины могут принимать значения из любого множества. Так, случайные вектора, случайные функции, случайные множества, случайные ранжировки (упорядочения) – это отдельные виды случайных величин. Используется и иная терминология: термин «случайная величина» сохраняется только за числовыми функциями, определенными на пространстве элементарных событий, а в случае иных областей значений используется термин «случайный элемент». (Замечание для математиков: все рассматриваемые функции, определенные на пространстве элементарных событий, предполагаются измеримыми.)

Статистика $И_1$ называется *доверительной областью*, соответствующей доверительной вероятности г, если

$$P\{\theta \in \Theta_1(x_1, x_2, ..., x_n)\} = \gamma. \quad (5)$$

Ясно, что этому условию удовлетворяет, как правило, не одна, а много доверительных областей. Из них выбирают для практического применения какую-либо одну, исходя из дополнительных соображений, например, из соображений симметрии или минимизируя объем доверительной области, т.е. меру множества $И_1$.

При оценке одного числового параметра в качестве доверительных областей обычно применяют доверительные интервалы (в том числе лучи), а не иные типа подмножеств прямой. Более того, для многих двухпараметрических и трехпараметрических распределений (нормальных, логарифмически нормальных, Вейбулла-Гнеденко, гамма-распределений и др.) обычно используют точечные оценки и построенные на их основе доверительные границы для каждого из двух или трех параметров отдельно. Это делают для удобства пользования результатами

расчетов: доверительные интервалы легче применять, чем фигуры на плоскости или тела в трехмерном пространстве.

Как следует из сказанного выше, *доверительный интервал* – это интервал, который с заданной вероятностью накроет неизвестное значение оцениваемого параметра распределения. Границы доверительного интервала называют *доверительными границами*. Доверительная вероятность г – вероятность того, что доверительный интервал накроет действительное значение параметра, оцениваемого по выборочным данным. Оцениванием с помощью доверительного интервала называют способ оценки, при котором с заданной доверительной вероятностью устанавливают границы доверительного интервала.

Для числового параметра и рассматривают верхнюю доверительную границу $и_В$, нижнюю доверительную границу $и_Н$ и двусторонние доверительные границы – верхнюю $и_{1В}$ и нижнюю $и_{1Н}$. Все четыре доверительные границы – функции от результатов наблюдений $x_1, x_2, \ldots, x_n$ и доверительной вероятности г.

Верхняя доверительная граница $и_В$ – случайная величина $и_В = и_В(x_1, x_2, \ldots, x_n; г)$, для которой $Р(и \le и_В) = г$, где и – истинное значение оцениваемого параметра. Доверительный интервал в этом случае имеет вид $(-\infty; и_В]$.

Нижняя доверительная граница $и_Н$ – случайная величина $и_Н = и_Н(x_1, x_2, \ldots, x_n; г)$, для которой $Р(и \ge и_Н) = г$, где и – истинное значение оцениваемого параметра. Доверительный интервал в этом случае имеет вид $[и_Н; +\infty)$.

Двусторонние доверительные границы - верхняя $и_{1В}$ и нижняя $и_{1Н}$ - это случайные величины $и_{1В} = и_{1В}(x_1, x_2, \ldots, x_n; г)$ и $и_{1Н} = и_{1Н}(x_1, x_2, \ldots, x_n; г)$ такие, что $Р(и_{1Н} \le и \le и_{1В}) = г$, где и – истинное значение оцениваемого параметра. Доверительный интервал в этом случае имеет вид $[и_{1Н}; и_{1В}]$.

Вероятности, связанные с доверительными границами, можно записать в виде частных случаев формулы (5):

$$P\{\theta \in (-\infty;\theta_B]\} = \gamma,\ \ P\{\theta \in [\theta_H;+\infty)\} = \gamma,\ \ P\{\theta \in [\theta_H;\theta_B]\} = \gamma.$$

В нормативно-технической и инструктивно-методической документации, научной и учебной литературе используют два типа правил определения доверительных границ – построенных на основе точного распределения и построенных на основе асимптотического распределения некоторой точечной оценки $и_n$ параметра и. Рассмотрим примеры.

Пример 10. Пусть $x_1, x_2, \ldots, x_n$ – выборка из нормального закона $N(m,$ у), параметры m и у неизвестны. Укажем доверительные границы для m.

Известно [11], что случайная величина

$$Y = \sqrt{n}\,\frac{\bar{x} - m}{s_0}$$

имеет распределение Стьюдента с (т-1) степенью свободы, где $\bar{x}$ - выборочное среднее арифметическое и s_0 – выборочное среднее квадратическое отклонение. Пусть $t_\gamma(n-1)$ и $t_{1-\gamma}(n-1)$ - квантили указанного распределения порядка г и 1-г соответственно. Тогда

$$P\{Y \le t_г(n\text{-}1)\} = г,\ P\{Y \ge t_{1\text{-}г}(n\text{-}1)\} = г.$$

Следовательно,

$$P\left\{m \ge \bar{x} - t_\gamma(n-1)\frac{s_0}{\sqrt{n}}\right\} = \gamma,$$

т.е. в качестве нижней доверительной границы $и_Н$, соответствующей доверительной вероятности г, следует взять

$$\theta_H(x_1, x_2, \ldots, x_n;\gamma) = \bar{x} - t_\gamma(n-1)\frac{s_0}{\sqrt{n}}. \quad (6)$$

Аналогично получаем, что

$$P\left\{m \le \bar{x} - t_{1-\gamma}(n-1)\frac{s_0}{\sqrt{n}}\right\} = \gamma.$$

Поскольку распределение Стьюдента симметрично относительно 0, то $t_{1-\gamma}(n-1) = - t_{1-\gamma}(n-1)\, t_{\gamma}(n-1)$. Следовательно, в качестве верхней доверительной границы $и_В$ для m, соответствующей доверительной вероятности г, следует взять

$$\theta_B(x_1, x_2, ..., x_n; \gamma) = \bar{x} + t_{\gamma}(n-1)\frac{s_0}{\sqrt{n}}. \quad (7)$$

Как построить двусторонние доверительные границы? Положим

$$\theta_{1H} = \theta_H(x_1, x_2, ..., x_n; \gamma_1), \ \theta_{1B} = \theta_B(x_1, x_2, ..., x_n; \gamma_2),$$

где $и_{1Н}$ и $и_{1В}$ заданы формулами (6) и (7) соответственно. Поскольку неравенство $и_{1Н} \le m \le и_{1В}$ выполнено тогда и только тогда, когда

$$t_{\gamma_2}(n-1) \ge Y \ge t_{1-\gamma_1}(n-1),$$

то

$$P\{и_{1Н} \le m \le и_{1В}\} = г_1 + г_2 - 1,$$

(в предположении, что $г_1 > 0{,}5$; $г_2 > 0{,}5$). Следовательно, если $г = г_1 + г_2 - 1$, то $и_{1Н}$ и $и_{1В}$ – двусторонние доверительные границы для m, соответствующие доверительной вероятности г. Обычно полагают $г_1 = г_2$, т.е. в качестве двусторонних доверительных границ $и_{1Н}$ и $и_{1В}$, соответствующих доверительной вероятности г, используют односторонние доверительные границы $и_Н$ и $и_В$, соответствующие доверительной вероятности (1+г)/2.

Другой вид правил построения доверительных границ для параметра и основан на асимптотической нормальности некоторой точечной оценки $и_n$ этого параметра. В вероятностно-статистических методах принятия решений используют, как уже отмечалось, несмещенные или асимптотически несмещенные оценки $и_n$, для которых смещение либо равно 0, либо при больших объемах выборки пренебрежимо мало по сравнению со средним квадратическим отклонением оценки $и_n$. Для таких оценок при всех x

$$\lim_{n\to\infty} P\left\{\frac{\theta_n-\theta}{\sqrt{D(\theta_n)}}\le x\right\}=\Phi(x),$$

где *Ф(x)* – функция нормального распределения *N*(0;1). Пусть $u_г$ – квантиль порядка г распределения *N*(0;1). Тогда

$$\lim_{n\to\infty} P\left\{\frac{\theta_n-\theta}{\sqrt{D(\theta_n)}}\le u_\gamma\right\}=\gamma \quad (8)$$

Поскольку неравенство

$$\frac{\theta_n-\theta}{\sqrt{D(\theta_n)}}\le u_\gamma$$

равносильно неравенству

$$\theta_n-u_\gamma\sqrt{D(\theta_n)}\le\theta,$$

то в качестве $и_Н$ можно было бы взять левую часть последнего неравенства. Однако точное значение дисперсии $D(и_n)$ обычно неизвестно. Зато часто удается доказать, что дисперсия оценки имеет вид

$$D(\theta_n)=\frac{h(\theta)}{n}$$

(с точностью до пренебрежимо малых при росте *n* слагаемых), где *h*(и) – некоторая функция от неизвестного параметра и. Справедлива теорема о наследовании сходимости [7, §2.4], согласно которой при подстановке в *h*(и) оценки $и_n$ вместо и соотношение (8) остается справедливым, т.е.

$$\lim_{n\to\infty} P\left\{\theta_n-u_\gamma\frac{\sqrt{h(\theta_n)}}{\sqrt{n}}\le\theta\right\}=\gamma.$$

Следовательно, в качестве приближенной нижней доверительной границы следует взять

$$\theta_H=\theta_n-u_\gamma\frac{\sqrt{h(\theta_n)}}{\sqrt{n}},$$

а в качестве приближенной верхней доверительной границы -

$$\theta_B=\theta_n+u_\gamma\frac{\sqrt{h(\theta_n)}}{\sqrt{n}}.$$

С ростом объема выборки качество приближенных доверительных границ улучшается, т.к. вероятности событий {и $\geq$ $и_Н$} и {и $\leq и_В$} стремятся к г. Для построения двусторонних доверительных границ поступают аналогично правилу, указанному выше в примере 10 для интервального оценивания параметра *m* нормального распределения. А именно, используют односторонние доверительные границы, соответствующие доверительной вероятности (1+г)/2.

При обработке экономических, управленческих или технических статистических данных обычно используют значение доверительной вероятности г = 0,95. Применяют также значения г = 0,99 или г = 0,90. Иногда встречаются значения г = 0,80, г = 0,975, г = 0,98 и др.

Доверительное оценивание для дискретных распределений. Для дискретных распределений, таких, как биномиальное, гипергеометрическое или распределение Пуассона (а также распределения статистики Колмогорова

$$D_n = \sqrt{n} \sup_x | F_n(x) - F_0(x) |$$

и других непараметрических статистик), функции распределения имеют скачки. Поэтому для заданного заранее значения г, например, г= 0,95, нельзя указать доверительные границы, поскольку уравнения, с помощью которых вводятся доверительные границы, не имеют ни одного решения. Так, рассмотрим биномиальное распределение

$$P(Y = y \mid p, n) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,2,\ldots,n,$$

где *Y* – число осуществлений события, *n* – объем выборки. Для него нельзя указать статистику *K(Y, n)* такую, что

$$P\{p \leq K(Y, n)\} = г,$$

поскольку $K(Y, n)$ – функция от Y и может принимать не больше значений, чем принимает Y, т.е. $n + 1$, а для г имеется бесконечно много возможных значений – столько, сколько точек на отрезке. Сказанная означает, что верхней доверительной границы в случае биномиального распределения не существует.

Для дискретных распределений приходится изменить определения доверительных границ. Покажем изменения на примере биномиального распределения. Так, в качестве верхней доверительной границы $и_В$ используют наименьшее $K(Y, n)$ такое, что

$$P\{p \leq K(Y, n)\} \geq г.$$

Аналогичным образом поступают для других доверительных границ и других распределений. Необходимо иметь в виду, что при небольших n и p истинная доверительная вероятность $P\{p \leq K(Y, n)\}$ может существенно отличаться от номинальной г, как это подробно продемонстрировано в работе [13]. Поэтому наряду с величинами типа $K(Y, n)$ (т.е. доверительных границ) при разработке таблиц и компьютерных программ необходимо предусматривать возможность получения и величин типа $P\{p \leq K(Y, n)\}$ (т.е. достигаемых доверительных вероятностей).

Основные понятия, используемые при проверке гипотез. Статистическая гипотеза – любое предположение, касающееся неизвестного распределения случайных величин (элементов). Приведем формулировки нескольких статистических гипотез:

1. Результаты наблюдений имеют нормальное распределение с нулевым математическим ожиданием.

2. Результаты наблюдений имеют функцию распределения $N(0,1)$.

3. Результаты наблюдений имеют нормальное распределение.

4. Результаты наблюдений в двух независимых выборках имеют одно и то же нормальное распределение.

5. Результаты наблюдений в двух независимых выборках имеют одно и то же распределение.

Различают нулевую и альтернативную гипотезы. Нулевая гипотеза – гипотеза, подлежащая проверке. Альтернативная гипотеза – каждая допустимая гипотеза, отличная от нулевой. Нулевую гипотезу обозначают H_0, альтернативную – H_1 (от Hypothesis – «гипотеза» (англ.)).

Выбор тех или иных нулевых или альтернативных гипотез определяется стоящими перед менеджером, экономистом, инженером, исследователем прикладными задачами. Рассмотрим примеры.

Пример 11. Пусть нулевая гипотеза – гипотеза 2 из приведенного выше списка, а альтернативная – гипотеза 1. Сказанное означает, то реальная ситуация описывается вероятностной моделью, согласно которой результаты наблюдений рассматриваются как реализации независимых одинаково распределенных случайных величин с функцией распределения $N(0,\sigma)$, где параметр σ неизвестен статистику. В рамках этой модели нулевую гипотезу записывают так:

$$H_0: \sigma = 1,$$

а альтернативную так:

$$H_1: \sigma \neq 1.$$

Пример 12. Пусть нулевая гипотеза – по-прежнему гипотеза 2 из приведенного выше списка, а альтернативная – гипотеза 3 из того же списка. Тогда в вероятностной модели управленческой, экономической или производственной ситуации предполагается, что результаты наблюдений образуют выборку из нормального распределения $N(m, \sigma)$ при некоторых значениях m и σ. Гипотезы записываются так:

$$H_0: m = 0, \sigma = 1$$

(оба параметра принимают фиксированные значения);

H_1: $m \neq 0$ и/или $\sigma \neq 1$

(т.е. либо $m \neq 0$, либо $\sigma \neq 1$, либо и $m \neq 0$, и $\sigma \neq 1$).

Пример 13. Пусть H_0 – гипотеза 1 из приведенного выше списка, а H_1 – гипотеза 3 из того же списка. Тогда вероятностная модель – та же, что в примере 12,

H_0: $m = 0$, σ произвольно;

H_1: $m \neq 0$, σ произвольно.

Пример 14. Пусть H_0 – гипотеза 2 из приведенного выше списка, а согласно H_1 результаты наблюдений имеют функцию распределения $F(x)$, не совпадающую с функцией стандартного нормального распределения $\Phi(x)$. Тогда

H_0: $F(x) = \Phi(x)$ при всех x (записывается как $F(x) \equiv \Phi(x)$);

H_1: $F(x_0) \neq \Phi(x_0)$ при некотором x_0 (т.е. неверно, что $F(x) \equiv \Phi(x)$).

Примечание. Здесь $\equiv$ - знак тождественного совпадения функций (т.е. совпадения при всех возможных значениях аргумента x).

Пример 15. Пусть H_0 – гипотеза 3 из приведенного выше списка, а согласно H_1 результаты наблюдений имеют функцию распределения $F(x)$, не являющуюся нормальной. Тогда

$H_0 : F(x) \equiv \Phi\left(\frac{x-m}{\sigma}\right)$ при некоторых m, σ;

H_1: для любых m, σ найдется $x_0 = x_0(m, \sigma)$ такое, что $F(x_0) \neq \Phi\left(\frac{x_0 - m}{\sigma}\right)$.

Пример 16. Пусть H_0 – гипотеза 4 из приведенного выше списка, согласно вероятностной модели две выборки извлечены из совокупностей с функциями распределения $F(x)$ и $G(x)$, являющихся нормальными с параметрами m_1, σ_1 и m_2, σ_2 соответственно, а H_1 – отрицание H_0. Тогда

H_0: $m_1 = m_2$, $\sigma_1 = \sigma_2$, причем m_1и σ_1 произвольны;

H_1: $m_1 \neq m_2$ и/или $\sigma_1 \neq \sigma_2$.

Пример 17. Пусть в условиях примера 16 дополнительно известно, что $y_1 = y_2$. Тогда

H_0: $m_1 = m_2$, $y > 0$, причем m_1и y произвольны;

H_1: $m_1 \neq m_2$, $y > 0$.

Пример 18. Пусть H_0 – гипотеза 5 из приведенного выше списка, согласно вероятностной модели две выборки извлечены из совокупностей с функциями распределения *F(x)* и *G(x)* соответственно, а H_1 – отрицание H_0. Тогда

H_0: *F(x)* ≡ *G(x)*, где *F(x)* – произвольная функция распределения;

H_1: *F(x)* и *G(x)* - произвольные функции распределения, причем

F(x) ≠ *G(x)* при некоторых *x*.

Пример 19. Пусть в условиях примера 17 дополнительно предполагается, что функции распределения *F(x)* и *G(x)* отличаются только сдвигом, т.е. *G(x) = F(x - a)* при некотором *a*. Тогда

H_0: *F(x)* ≡ *G(x)*,

где *F(x)* – произвольная функция распределения;

H_1: *G(x) = F(x - a), a* ≠ 0,

где *F(x)* – произвольная функция распределения.

Пример 20. Пусть в условиях примера 14 дополнительно известно, что согласно вероятностной модели ситуации *F(x)* - функция нормального распределения с единичной дисперсией, т.е. имеет вид *N*(*m*, 1). Тогда

H_0: $m = 0$ (т.е. *F(x) = Φ(x)*

при всех *x*);(записывается как *F(x)* ≡ *Φ(x)*);

H_1: $m \neq 0$

(т.е. неверно, что *F(x)* ≡ *Φ(x)*).

Пример 21. При статистическом регулировании технологических, экономических, управленческих или иных процессов [2] рассматривают выборку, извлеченную из совокупности с нормальным распределением и известной дисперсией, и гипотезы

$$H_0: m = m_0,$$

$$H_1: m = m_1,$$

где значение параметра $m = m_0$ соответствует налаженному ходу процесса, а переход к $m = m_1$ свидетельствует о разладке.

Пример 22. При статистическом приемочном контроле [2] число дефектных единиц продукции в выборке подчиняется гипергеометрическому распределению, неизвестным параметром является $p = D/N$ – уровень дефектности, где N – объем партии продукции, D – общее число дефектных единиц продукции в партии. Используемые в нормативно-технической и коммерческой документации (стандартах, договорах на поставку и др.) планы контроля часто нацелены на проверку гипотезы

$$H_0: p \leq AQL$$

против альтернативной гипотезы

$$H_1: p \geq LQ,$$

где AQL – приемочный уровень дефектности, LQ – браковочный уровень дефектности (очевидно, что $AQL < LQ$).

Пример 23. В качестве показателей стабильности технологического, экономического, управленческого или иного процесса используют ряд характеристик распределений контролируемых показателей, в частности, коэффициент вариации $v = \text{у}/M(X)$. Требуется проверить нулевую гипотезу

$$H_0: v \leq v_0$$

при альтернативной гипотезе

$$H_1: v > v_0,$$

где v_0 – некоторое заранее заданное граничное значение.

Пример 24. Пусть вероятностная модель двух выборок – та же, что в примере 18, математические ожидания результатов наблюдений в первой и второй выборках обозначим $M(X)$ и $M(У)$ соответственно. В ряде ситуаций проверяют нулевую гипотезу

$$H_0: M(X) = M(У)$$

против альтернативной гипотезы

$$H_1: M(X) \neq M(У).$$

Пример 25. Выше отмечалось большое значение в математической статистике функций распределения, симметричных относительно 0, При проверке симметричности

H_0: $F(-x) = 1 - F(x)$ при всех x, в остальном F произвольна;

H_1: $F(-x_0) \neq 1 - F(x_0)$ при некотором x_0, в остальном F произвольна.

В вероятностно-статистических методах принятия решений используются и многие другие постановки задач проверки статистических гипотез. Некоторые из них рассматриваются ниже.

Конкретная задача проверки статистической гипотезы полностью описана, если заданы нулевая и альтернативная гипотезы. Выбор метода проверки статистической гипотезы, свойства и характеристики методов определяются как нулевой, так и альтернативной гипотезами. Для проверки одной и той же нулевой гипотезы при различных альтернативных гипотезах следует использовать, вообще говоря, различные методы. Так, в примерах 14 и 20 нулевая гипотеза одна и та же, а альтернативные – различны. Поэтому в условиях примера 14 следует применять методы, основанные на критериях согласия с параметрическим семейством (типа Колмогорова или типа омега-квадрат), а в условиях примера 20 – методы на основе критерия Стьюдента или критерия Крамера-Уэлча [2,11]. Если в условиях примера 14 использовать критерий Стьюдента, то он не будет решать поставленных задач. Если в условиях примера 20 использовать критерий согласия типа Колмогорова, то он, напротив, будет решать поставленные задачи, хотя, возможно, и хуже, чем специально приспособленный для этого случая критерий Стьюдента.

При обработке реальных данных большое значение имеет правильный выбор гипотез H_0 и H_1. Принимаемые предположения,

например, нормальность распределения, должны быть тщательно обоснованы, в частности, статистическими методами. Отметим, что в подавляющем большинстве конкретных прикладных постановок распределение результатов наблюдений отлично от нормального [2].

Часто возникает ситуация, когда вид нулевой гипотезы вытекает из постановки прикладной задачи, а вид альтернативной гипотезы не ясен. В таких случаях следует рассматривать альтернативную гипотезу наиболее общего вида и использовать методы, решающие поставленную задачу при всех возможных H_1. В частности при проверке гипотезы 2 (из приведенного выше списка) как нулевой следует в качестве альтернативной гипотезы использовать H_1 из примера 14, а не из примера 20, если нет специальных обоснований нормальности распределения результатов наблюдений при альтернативной гипотезе.

Параметрические и непараметрические гипотезы. Статистические гипотезы бывают параметрические и непараметрические. Предположение, которое касается неизвестного значения параметра распределения, входящего в некоторое параметрическое семейство распределений, называется параметрической гипотезой (напомним, что параметр может быть и многомерным). Предположение, при котором вид распределения неизвестен (т.е. не предполагается, что оно входит в некоторое параметрическое семейство распределений), называется непараметрической гипотезой. Таким образом, если распределение $F(x)$ результатов наблюдений в выборке согласно принятой вероятностной модели входит в некоторое параметрическое семейство $\{F(x;и), и \in И\}$, т.е. $F(x) = F(x;и_0)$ при некотором $и_0 \in И$, то рассматриваемая гипотеза – параметрическая, в противном случае – непараметрическая.

Если и H_0 и H_1 – параметрические гипотезы, то задача проверки статистической гипотезы – параметрическая. Если хотя бы одна из гипотез H_0 и H_1 – непараметрическая, то задача проверки статистической гипотезы – непараметрическая. Другими словами, если вероятностная модель ситуации – параметрическая, т.е. полностью описывается в терминах того или иного параметрического семейства распределений вероятностей, то и задача проверки статистической гипотезы – параметрическая. Если же вероятностная модель ситуации – непараметрическая, т.е. ее нельзя полностью описать в терминах какого-либо параметрического семейства распределений вероятностей, то и задача проверки статистической гипотезы – непараметрическая. В примерах 11-13, 16, 17, 20-22 даны постановки параметрических задач проверки гипотез, а в примерах 14, 15, 18, 19, 23-25 – непараметрических. Непараметрические задачи делятся на два класса: в одном из них речь идет о проверке утверждений, касающихся функций распределения (примеры 14, 15, 18, 19, 25), во втором – о проверке утверждений, касающихся характеристик распределений (примеры 23, 24).

Статистическая гипотеза называется простой, если она однозначно задает распределение результатов наблюдений, вошедших в выборку. В противном случае статистическая гипотеза называется сложной. Гипотеза 2 из приведенного выше списка, нулевые гипотезы в примерах 11, 12, 14, 20, нулевая и альтернативная гипотезы в примере 21 – простые, все остальные упомянутые выше гипотезы – сложные.

Статистические критерии. Однозначно определенный способ проверки статистических гипотез называется статистическим критерием. Статистический критерий строится с помощью статистики $U(x_1, x_2, \ldots, x_n)$ – функции от результатов наблюдений $x_1, x_2, \ldots, x_n$. В пространстве значений статистики U

выделяют критическую область Ш, т.е. область со следующим свойством: если значения применяемой статистики принадлежат данной области, то отклоняют (иногда говорят -отвергают) нулевую гипотезу, в противном случае – не отвергают (т.е. принимают).

Статистику U, используемую при построении определенного статистического критерия, называют статистикой этого критерия. Например, в задаче проверки статистической гипотезы, приведенной в примере 14, применяют критерий Колмогорова, основанный на статистике

$$D_n = \sqrt{n} \sup_x | F_n(x) - F_0(x) | .$$

При этом D_n называют статистикой критерия Колмогорова.

Частным случаем статистики U является векторзначная функция результатов наблюдений $U_0(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n)$, значения которой – набор результатов наблюдений. Если x_i – числа, то U_0 – набор n чисел, т.е. точка n–мерного пространства. Ясно, что статистика критерия U является функцией от U_0, т.е. $U = f(U_0)$. Поэтому можно считать, что Ш – область в том же n–мерном пространстве, нулевая гипотеза отвергается, если $(x_1, x_2, \ldots, x_n) \in$ Ш, и принимается в противном случае.

В вероятностно-статистических методах обработки данных и принятия решений статистические критерии, как правило, основаны на статистиках U, принимающих числовые значения, и критические области имеют вид

$$\text{Ш} = \{U(x_1, x_2, \ldots, x_n) > C\}, \quad (9)$$

где C – некоторые числа.

Статистические критерии делятся на параметрические и непараметрические. Параметрические критерии используются в параметрических задачах проверки статистических гипотез, а непараметрические – в непараметрических задачах.

Уровень значимости и мощность. При проверке статистической гипотезы возможны ошибки. Есть два рода ошибок. Ошибка первого рода заключается в том, что отвергают нулевую гипотезу, в то время как в действительности эта гипотеза верна. Ошибка второго рода состоит в том, что принимают нулевую гипотезу, в то время как в действительности эта гипотеза неверна.

Вероятность ошибки первого рода называется уровнем значимости и обозначается б. Таким образом, б = $P\{U \in Ш \mid H_0\}$, т.е. уровень значимости б – это вероятность события $\{U \in Ш\}$, вычисленная в предположении, что верна нулевая гипотеза H_0.

Уровень значимости однозначно определен, если H_0 – простая гипотеза. Если же H_0 – сложная гипотеза, то уровень значимости, вообще говоря, зависит от функции распределения результатов наблюдений, удовлетворяющей H_0. Статистику критерия U обычно строят так, чтобы вероятность события $\{U \in Ш\}$ не зависела от того, какое именно распределение (из удовлетворяющих нулевой гипотезе H_0) имеют результаты наблюдений. Для статистик критерия U общего вида под уровнем значимости понимают максимально возможную ошибку первого рода. Максимум (точнее, супремум) берется по всем возможным распределениям, удовлетворяющим нулевой гипотезе H_0, т.е. б = sup $P\{U \in Ш \mid H_0\}$.

Если критическая область имеет вид, указанный в формуле (9), то

$$P\{U > C \mid H_0\} = б. \quad (10)$$

Если C задано, то из последнего соотношения определяют б. Часто поступают по иному - задавая б (обычно б = 0,05, иногда б = 0,01 или б = 0,1, другие значения б используются гораздо реже), определяют C из уравнения (10), обозначая его $C_б$, и используют критическую область Ш = $\{U > C_б\}$ с заданным уровнем значимости б.

Вероятность ошибки второго рода есть $P\{U \notin Ш \mid H_1\}$. Обычно используют не эту вероятность, а ее дополнение до 1, т.е. $P\{U \in Ш \mid H_1\} = 1 - P\{U \notin Ш \mid H_1\}$. Эта величина носит название мощности критерия. Итак, мощность критерия – это вероятность того, что нулевая гипотеза будет отвергнута, когда альтернативная гипотеза верна.

Понятия уровня значимости и мощности критерия объединяются в понятии функции мощности критерия – функции, определяющей вероятность того, что нулевая гипотеза будет отвергнута. Функция мощности зависит от критической области Ш и действительного распределения результатов наблюдений. В параметрической задаче проверки гипотез распределение результатов наблюдений задается параметром и. В этом случае функция мощности обозначается M(Ш,и) и зависит от критической области Ш и действительного значения исследуемого параметра и. Если

$$H_0: и = и_0,$$

$$H_1: и = и_1,$$

то

$$M(Ш,и_0) = б,$$

$$M(Ш,и_1) = 1 – в,$$

где б – вероятность ошибки первого рода, в - вероятность ошибки второго рода. В статистическом приемочном контроле б – риск изготовителя, в – риск потребителя. При статистическом регулировании технологического процесса б – риск излишней наладки, в – риск незамеченной разладки.

Функция мощности M(Ш,и) в случае одномерного параметра и обычно достигает минимума, равного б, при $и = и_0$, монотонно возрастает при удалении от $и_0$ и приближается к 1 при $| и - и_0 | \to \infty$.

В ряде вероятностно-статистических методов принятия решений используется оперативная характеристика L(Ш,и) -

вероятность принятия нулевой гипотезы в зависимости от критической области Ш и действительного значения исследуемого параметра и. Ясно, что

$$L(\text{Ш},\text{и}) = 1 - M(\text{Ш},\text{и}).$$

Состоятельность и несмещенность критериев. Основной характеристикой статистического критерия является функция мощности. Для многих задач проверки статистических гипотез разработан не один статистический критерий, а целый ряд. Чтобы выбрать из них определенный критерий для использования в конкретной практической ситуации, проводят сравнение критериев по различным показателям качества [2, приложение 3], прежде всего с помощью их функций мощности. В качестве примера рассмотрим лишь два показателя качества критерия проверки статистической гипотезы – состоятельность и несмещенность.

Пусть объем выборки n растет, а U_n и Ш_n – статистики критерия и критические области соответственно. Критерий называется состоятельным, если

$$\lim_{n\to\infty} P\{U_n \in \Psi_n \mid H_1\} = 1,$$

т.е. вероятность отвергнуть нулевую гипотезу стремится к 1, если верна альтернативная гипотеза.

Статистический критерий называется несмещенным, если для любого и_0, удовлетворяющего H_0, и любого и_1 , удовлетворяющего H_1, справедливо неравенство

$$\text{Р}\{U \in \text{Ш} \mid \text{и}_0\} < \text{Р}\{U \in \text{Ш} \mid \text{и}_1\},$$

т.е. при справедливости H_0 вероятность отвергнуть H_0 меньше, чем при справедливости H_1.

При наличии нескольких статистических критериев в одной и той же задаче проверки статистических гипотез следует использовать состоятельные и несмещенные критерии.

6. Некоторые типовые задачи прикладной статистики и методы их решения

Статистические данные и прикладная статистика. Под прикладной статистикой обычно понимают часть математической статистики, посвященную методам обработки реальных статистических данных, а также соответствующее математическое и программное обеспечение. Таким образом, чисто математические задачи не включают в прикладную статистику. В последние десятилетия термин «математическая статистика» все чаще применяют для обозначения чисто математической дисциплины, которая изучает свойства математических объектов и структур, введенных в классической статистике ранее середины XX века. При таком понимании прикладная статистика – самостоятельная научно-практическая дисциплина, не имеющая пересечения с математической статистикой. Прикладную статистику и статистические методы в целом можно отнести к кибернетике или к прикладной математике.

Под статистическими данными понимают числовые или нечисловые значения контролируемых параметров (признаков) исследуемых объектов, которые получены в результате наблюдений (измерений, анализов, испытаний, опытов и т.д.) определенного числа признаков, у каждой единицы, вошедшей в исследование. Способы получения статистических данных и объемы выборок устанавливают, исходя из постановок конкретной прикладной задачи на основе методов математической теории планирования эксперимента.

Результат наблюдения x_i исследуемого признака X (или совокупности исследуемых признаков X) у i – ой единицы выборки отражает количественные и/или качественные свойства обследованной единицы с номером i (здесь $i = 1, 2, \ldots, n$, где n –

объем выборки). Деление прикладной статистики на направления соответственно виду обрабатываемых результатов наблюдений (т.е. на статистику случайных величин, многомерный статистический анализ, статистику временных рядов и статистику объектов нечисловой природы) обсуждалось выше.

Результаты наблюдений $x_1, x_2, \ldots, x_n$, где x_i – результат наблюдения i – ой единицы выборки, или результаты наблюдений для нескольких выборок, обрабатывают с помощью методов прикладной статистики, соответствующих поставленной задаче. Используют, как правило, аналитические методы, т.е. методы, основанные на численных расчетах (объекты нечисловой природы при этом описывают с помощью чисел). В отдельных случаях допустимо применение графических методов (визуального анализа).

Количество разработанных к настоящему времени методов обработки данных весьма велико. Они описаны в сотнях тысяч книг и статей, а также в стандартах и других нормативно-технических и инструктивно-методических документах.

Многие методы прикладной статистики требуют проведения трудоемких расчетов, поэтому для их реализации необходимо использовать компьютеры. Программы расчетов на ЭВМ должны соответствовать современному научному уровню. Однако для единичных расчетов при отсутствии соответствующего программного обеспечения успешно используют микрокалькуляторы.

Статистический анализ точности и стабильности технологических процессов и качества продукции. Статистические методы используют, в частности, для анализа точности и стабильности технологических процессов и качества продукции. Цель - подготовка решений, обеспечивающих эффективное функционирование технологических единиц и

повышение качества и конкурентоспособности выпускаемой продукции. Статистические методы следует применять во всех случаях, когда по результатам ограниченного числа наблюдений требуется установить причины улучшения или ухудшения точности и стабильности технологического оборудования. Под точностью технологического процесса понимают свойство технологического процесса, обусловливающее близость действительных и номинальных значений параметров производимой продукции. Под стабильностью технологического процесса понимают свойство технологического процесса, обусловливающее постоянство распределений вероятностей для его параметров в течение некоторого интервала времени без вмешательства извне.

Целями применения статистических методов анализа точности и стабильности технологических процессов и качества продукции на стадиях разработки, производства и эксплуатации (потребления) продукции являются, в частности:

• определение фактических показателей точности и стабильности технологического процесса, оборудования или качества продукции;

• установление соответствия качества продукции требованиям нормативно-технической документации;

• проверка соблюдения технологической дисциплины;

• изучение случайных и систематических факторов, способных привести к появлению дефектов;

• выявление резервов производства и технологии;

• обоснование технических норм и допусков на продукцию;

• оценка результатов испытаний опытных образцов при обосновании требований к продукции и нормативов на нее;

• обоснование выбора технологического оборудования и средств измерений и испытаний;

• сравнение различных образцов продукции;

• обоснование замены сплошного контроля статистическим;

• выявление возможности внедрения статистических методов управления качеством продукции, и т.д.

Для достижения перечисленных выше целей применяют различные методы описания данных, оценивания и проверки гипотез. Приведем примеры постановок задач.

Задачи одномерной статистики (статистики случайных величин). Сравнение математических ожиданий проводят в тех случаях, когда необходимо установить соответствие показателей качества изготовленной продукции и эталонного образца. Это – задача проверки гипотезы:

$$H_0: M(\mathrm{X}) = m_0,$$

где m_0 – значение соответствующее эталонному образцу; X – случайная величина, моделирующая результаты наблюдений. В зависимости от формулировки вероятностной модели ситуации и альтернативной гипотезы сравнение математических ожиданий проводят либо параметрическими, либо непараметрическими методами.

Сравнение дисперсий проводят тогда, когда требуется установить отличие рассеивания показателя качества от номинального. Для этого проверяют гипотезу:

$$H_0 : D(X) = \sigma_0^2.$$

Ряд иных постановок задач одномерной статистики приведен ниже. Не меньшее значение, чем задачи проверки гипотез, имеют задачи оценивания параметров. Они, как и задачи проверки гипотез, в зависимости от используемой вероятностной модели ситуации делятся на параметрические и непараметрические.

В параметрических задачах оценивания принимают вероятностную модель, согласно которой результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают как реализации n независимых

случайных величин с функцией распределения $F(x;и)$. Здесь и – неизвестный параметр, лежащий в пространстве параметров И заданном используемой вероятностной моделью. Задача оценивания состоит в определении точечной оценок и доверительных границ (либо доверительной области) для параметра и.

Параметр и – либо число, либо вектор фиксированной конечной размерности. Так, для нормального распределения и = (m, $у^2$) – двумерный вектор, для биномиального и = p – число, для гамма-распределения и = (a, b, c) – трехмерный вектор, и т.д.

В современной математической статистике разработан ряд общих методов определения оценок и доверительных границ – метод моментов, метод максимального правдоподобия, метод одношаговых оценок, метод устойчивых (робастных) оценок, метод несмещенных оценок и др. Кратко рассмотрим первые три из них. Теоретические основы различных методов оценивания и полученные с их помощью конкретные правила определения оценок и доверительных границ для тех или иных параметрических семейств распределений рассмотрены в специальной литературе, включены в нормативно-техническую и инструктивно-методическую документацию.

Метод моментов основан на использовании выражений для моментов рассматриваемых случайных величин через параметры их функций распределения. Оценки метода моментов получают, подставляя выборочные моменты вместо теоретических в функции, выражающие параметры через моменты.

В методе максимального правдоподобия, разработанном в основном Р.А.Фишером, в качестве оценки параметра и берут значение и*, для которого максимальна так называемая функция правдоподобия

$$f(x_1, и)\, f(x_2, и)\, \ldots\, f(x_n, и),$$

где $x_1, x_2, \ldots, x_n$ - результаты наблюдений; $f(x, и)$ – их плотность распределения, зависящая от параметра и, который необходимо оценить.

Оценки максимального правдоподобия, как правило, эффективны (или асимптотически эффективны) и имеют меньшую дисперсию, чем оценки метода моментов. В отдельных случаях формулы для них выписываются явно (нормальное распределение, экспоненциальное распределение без сдвига). Однако чаще для их нахождения необходимо численно решать систему трансцендентных уравнений (распределения Вейбулла-Гнеденко, гамма). В подобных случаях целесообразно использовать не оценки максимального правдоподобия, а другие виды оценок, прежде всего одношаговые оценки. В литературе их иногда не вполне точно называют «приближенные оценки максимального правдоподобия». При достаточно больших объемах выборок они имеют столь же хорошие свойства, как и оценки максимального правдоподобия. Поэтому их следует рассматривать не как «приближенные», а как оценки, полученные по *другому* методу, не менее обоснованному и эффективному, чем метод максимального правдоподобия. Одношаговые оценки вычисляют по явным формулам ([14]).

В непараметрических задачах оценивания принимают вероятностную модель, в которой результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают как реализации n независимых случайных величин с функцией распределения $F(x)$ общего вида. От $F(x)$ требуют лишь выполнения некоторых условий типа непрерывности, существования математического ожидания и дисперсии и т.п. Подобные условия не являются столь жесткими, как условие принадлежности к определенному параметрическому семейству.

Непараметрическое оценивание математического ожидания. В непараметрической постановке оценивают либо характеристики случайной величины (математическое ожидание,

дисперсию, коэффициент вариации), либо ее функцию распределения, плотность и т.п. Так, в силу закона больших чисел выборочное среднее арифметическое $\bar{x}$ является состоятельной оценкой математического ожидания $M(X)$ (при любой функции распределения $F(x)$ результатов наблюдений, для которой математическое ожидание существует). С помощью центральной предельной теоремы определяют асимптотические доверительные границы

$$(M(X))_{\text{Н}} = \bar{x} - u\left(\frac{1+\gamma}{2}\right)\frac{s}{\sqrt{n}},\ (M(\text{X}))_{\text{В}} = \bar{x} + u\left(\frac{1+\gamma}{2}\right)\frac{s}{\sqrt{n}}.$$

где г – доверительная вероятность, $u\left(\frac{1+\gamma}{2}\right)$ - квантиль порядка $\frac{1+\gamma}{2}$ стандартного нормального распределения N(0;1) с нулевым математическим ожиданием и единичной дисперсией, $\bar{x}$ - выборочное среднее арифметическое, s – выборочное среднее квадратическое отклонение. Термин «асимптотические доверительные границы» означает, что вероятности

$$P\{(M(X))_H < M(X)\},\ P\{(M(X))_B > M(X)\},$$

$$P\{(M(X))_H < M(X) < (M(X))_B\}$$

стремятся к $\frac{1+\gamma}{2}$, $\frac{1+\gamma}{2}$ и г соответственно при $n \to \infty$, но, вообще говоря, не равны этим значениям при конечных n. Практически асимптотические доверительные границы дают достаточную точность при n порядка 10.

Непараметрическое оценивание функции распределения. Второй пример непараметрического оценивания – оценивание функции распределения. По теореме Гливенко эмпирическая функция распределения $F_n(x)$ является состоятельной оценкой функции распределения $F(x)$. Если $F(x)$ – непрерывная функция, то на основе теоремы Колмогорова доверительные границы для функции распределения $F(x)$ задают в виде

$$(F(x))_{\text{Н}} = \max\left\{0, F_n(x) - \frac{k(\gamma,n)}{\sqrt{n}}\right\},\ (F(x))_{\text{В}} = \min\left\{1, F_n(x) + \frac{k(\gamma,n)}{\sqrt{n}}\right\},$$

где *k(г,n)* – квантиль порядка г распределения статистики Колмогорова при объеме выборки *n* (напомним, что распределение этой статистики не зависит от $F(x)$).

Правила определения оценок и доверительных границ в параметрическом случае строятся на основе параметрического семейства распределений $F(x;\text{и})$. При обработке реальных данных возникает вопрос – соответствуют ли эти данные принятой вероятностной модели? Т.е. статистической гипотезе о том, что результаты наблюдений имеют функцию распределения из семейства $\{F(x;\text{и}),\ \text{и} \in \text{И}\}$ при некотором и = и_0? Такие гипотезы называют гипотезами согласия, а критерии их проверки – критериями согласия.

Если истинное значение параметра и = и_0 известно, функция распределения $F(x;\text{и}_0)$ непрерывна, то для проверки гипотезы согласия часто применяют критерий Колмогорова, основанный на статистике

$$D_n = \sqrt{n}\sup_x |F_n(x) - F(x,\theta_0|,$$

где $F_n(x)$ – эмпирическая функция распределения.

Если истинное значение параметра и_0 неизвестно, например, при проверке гипотезы о нормальности распределения результатов наблюдения (т.е. при проверке принадлежности этого распределения к семейству нормальных распределений), то иногда используют статистику

$$D_n(\theta^*) = \sqrt{n}\sup_x |F_n(x) - F(x,\theta^*)|.$$

Она отличается от статистики Колмогорова D_n тем, что вместо истинного значения параметра и_0 подставлена его оценка и*.

Распределение статистики D_n(и*) сильно отличается от распределения статистики D_n. В качестве примера рассмотрим

проверку нормальности, когда и = (m, у^2), а и* = ($\bar{x}$, s^2). Для этого случая квантили распределений статистик D_n и D_n(и*) приведены в табл.1 (см., например, [15]). Таким образом, квантили отличаются примерно в 1,5 раза.

Таблица 1.

Квантили статистик D_n и D_n(и*) при проверке нормальности

p	0,85	0,90	0,95	0,975	0,99
Квантили порядка p для D_n	1,138	1,224	1,358	1,480	1,626
Квантили порядка p для D_n(и*)	0,775	0,819	0,895	0,955	1,035

Проблема исключения промахов. При первичной обработке статистических данных важной задачей является исключение результатов наблюдений, полученных в результате грубых погрешностей и промахов. Например, при просмотре данных о весе (в килограммах) новорожденных детей наряду с числами 3,500, 2,750, 4,200 может встретиться число 35,00. Ясно, что это промах, и получено ошибочное число при ошибочной записи – запятая сдвинута на один знак, в результате результат наблюдения ошибочно увеличен в 10 раз.

Статистические методы исключения резко выделяющихся результатов наблюдений основаны на предположении, что подобные результаты наблюдений имеют распределения, резко отличающиеся от изучаемых, а потому их следует исключить из выборки.

Простейшая вероятностная модель такова. При нулевой гипотезе результаты наблюдений рассматриваются как реализации независимых одинаково распределенных случайных величин X_1, X_2 , , X_n с функцией распределения $F(x)$. При альтернативной гипотезе X_1, X_2 , , X_{n-1} – такие же, как и при нулевой гипотезе, а X_n соответствует грубой погрешности и имеет функцию распределения

$G(x) = F(x - c)$, где c велико. Тогда с вероятностью, близкой к 1 (точнее, стремящейся к 1 при росте объема выборки),

$$X_n = \max \{ X_1, X_2 , \quad , X_n\} = X_{max} ,$$

т.е. при описании данных в качестве возможной грубой ошибки следует рассматривать X_{max} . Критическая область имеет вид

$$Ш = \{x: x \geq d\}.$$

Критическое значение $d = d(б,n)$ выбирают в зависимости от уровня значимости б и объема выборки n из условия

$$P\{X_{max} \geq d \mid H_0\} = б . \quad (1)$$

Условие (1) эквивалентно при больших n и малых б следующему:

$$F(d) = \sqrt[n]{1-\alpha} \approx 1 - \frac{\alpha}{n}. \quad (2)$$

Если функция распределения результатов наблюдений $F(x)$ известна, то критическое значение d находят из соотношения (2). Если $F(x)$ известна с точностью до параметров, например, известно, что $F(x)$ – нормальная функция распределения, то также разработаны правила проверки рассматриваемой гипотезы [8].

Однако часто вид функции распределения результатов наблюдений известен не абсолютно точно и не с точностью до параметров, а лишь с некоторой погрешностью. Тогда соотношение (2) становится практически бесполезным, поскольку малая погрешность в определении $F(x)$, как можно показать, приводит к большой погрешности при определении критического значения d из условия (2), а при фиксированном d уровень значимости критерия может существенно отличаться от номинального [2].

Поэтому в ситуации, когда о $F(x)$ нет полной информации, однако известны математическое ожидание $M(X)$ и дисперсия $у^2 = D(X)$ результатов наблюдений $X_1, X_2 , \quad , X_n$, можно использовать непараметрические правила отбраковки, основанные на неравенстве

Чебышёва. С помощью этого неравенства найдем критическое значение $d = d(б,n)$ такое, что

$$P\left\{\max_{1\le i\le n} | X_i - M(X) | \ge d\right\} \le \alpha.$$

Так как

$$P\left\{\max_{1\le i\le n} | X_i - M(X) | < d\right\} = \left[P\{| X - M(X) | < d\}\right]^n,$$

то соотношение (3) будет выполнено, если

$$P\{| X - M(X) | \ge d\} \le 1 - \sqrt[n]{1-\alpha} \approx \frac{\alpha}{n}. \quad (4)$$

По неравенству Чебышёва

$$P\{| X - M(X) | \ge d\} \le \frac{\sigma^2}{d^2}, \quad (5)$$

поэтому для того, чтобы (4) было выполнено, достаточно приравнять правые части формул (4) и (5), т.е. определить d из условия

$$\frac{\sigma^2}{d^2} = \frac{\alpha}{n}, \quad d = \frac{\sigma\sqrt{n}}{\sqrt{\alpha}}. \quad (6)$$

Правило отбраковки, основанное на критическом значении d, вычисленном по формуле (6), использует минимальную информацию о функции распределения F(x) и поэтому исключает лишь результаты наблюдений, весьма далеко отстоящие от основной массы. Другими словами, значение d_1, заданное соотношением (1), обычно много меньше, чем значение d_2, заданное соотношением (6).

Многомерный статистический анализ. Перейдем к многомерному статистическому анализу. Его применяют при решении следующих задач:

• исследование зависимости между признаками;

• классификация объектов или признаков, заданных векторами;

• снижение размерности пространства признаков.

При этом результат наблюдений – вектор значений фиксированного числа количественных и иногда качественных признаков, измеренных у объекта. Напомним, что количественный признак – признак наблюдаемой единицы, который можно непосредственно выразить числом и единицей измерения. Количественный признак противопоставляется качественному - признаку наблюдаемой единицы, определяемому отнесением к одной из двух или более условных категорий (если имеется ровно две категории, то признак называется альтернативным). Статистический анализ качественных признаков – часть статистики объектов нечисловой природы. Количественные признаки делятся на признаки, измеренные в шкалах интервалов, отношений, разностей, абсолютной. А качественные – на признаки, измеренные в шкале наименований и порядковой шкале. Методы обработки данных должны быть согласованы со шкалами, в которых измерены рассматриваемые признаки [2, 3, 7].

Корреляция и регрессия. Целями исследования зависимости между признаками являются доказательство наличия связи между признаками и изучение этой связи. Для доказательства наличия связи между двумя случайными величинами X и Y применяют корреляционный анализ. Если совместное распределение X и Y является нормальным, то статистические выводы основывают на выборочном коэффициенте линейной корреляции, в остальных случаях используют коэффициенты ранговой корреляции Кендалла и Спирмена, а для качественных признаков – критерий хи-квадрат.

Регрессионный анализ применяют для изучения функциональной зависимости количественного признака Y от количественных признаков $x(1)$, $x(2)$, … , $x(k)$. Эту зависимость называют регрессионной или, кратко, регрессией. Простейшая вероятностная модель регрессионного анализа (в случае $k = 1$)

использует в качестве исходной информации набор пар результатов наблюдений (x_i, y_i), $i = 1, 2, \dots, n$, и имеет вид

$$y_i = ax_i + b + e_i,\ i = 1, 2, \dots, n,$$

где e_i – ошибки наблюдений. Иногда предполагают, что e_i – независимые случайные величины с одним и тем же нормальным распределением $N(0, \sigma^2)$. Поскольку распределение ошибок наблюдения обычно отлично от нормального, то целесообразно рассматривать регрессионную модель в непараметрической постановке [2], т.е. при произвольном распределении e_i.

Основная задача регрессионного анализа состоит в оценке неизвестных параметров a и b, задающих линейную зависимость y от x. Для решения этой задачи применяют разработанный еще К.Гауссом в 1794 г. метод наименьших квадратов, т.е. находят оценки неизвестных параметров моделиa и b из условия минимизации суммы квадратов

$$\sum_{1 \le i \le n} (y_i - ax_i - b)^2$$

по переменным a и b.

Теория регрессионного анализа описана и расчетные формулы даны в специальной литературе [2, 16, 17]. В этой теории разработаны методы точечного и интервального оценивания параметров, задающих функциональную зависимость, а также непараметрические методы оценивания этой зависимости, методы проверки различных гипотез, связанных с регрессионными зависимостями. Выбор планов эксперимента, т.е. точек x_i, в которых будут проводиться эксперименты по наблюдению y_i – предмет теории планирования эксперимента [18].

Дисперсионный анализ применяют для изучения влияния качественных признаков на количественную переменную. Например, пусть имеются k выборок результатов измерений количественного показателя качества единиц продукции,

выпущенных на k станках, т.е. набор чисел $(x_1(j), x_2(j), \dots, x_n(j))$, где j – номер станка, $j = 1, 2, \dots, k$, а n – объем выборки. В распространенной постановке дисперсионного анализа предполагают, что результаты измерений независимы и в каждой выборке имеют нормальное распределение $N(m(j), y^2)$ с одной и той же дисперсией. Хорошо разработаны и непараметрические постановки [19].

Проверка однородности качества продукции, т.е. отсутствия влияния номера станка на качество продукции, сводится к проверке гипотезы

$$H_0: m(1) = m(2) = \dots = m(k).$$

В дисперсионном анализе разработаны методы проверки подобных гипотез. Теория дисперсионного анализа и расчетные формулы рассмотрены в специальной литературе [20].

Гипотезу H_0 проверяют против альтернативной гипотезы H_1, согласно которой хотя бы одно из указанных равенств не выполнено. Проверка этой гипотезы основана на следующем «разложении дисперсий», указанном Р.А.Фишером:

$$(kn)s^2 = n\sum_{j=1}^{k} s^2(j) + (kn)s_1^2, \quad (7)$$

где s^2 – выборочная дисперсия в объединенной выборке, т.е.

$$s^2 = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i(j) - \bar{x})^2, \quad \bar{x} = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k} x_i(j).$$

Далее, $s^2(j)$ – выборочная дисперсия в j-ой группе,

$$s^2(j) = \frac{1}{n}\sum_{i=1}^{n}(x_i(j) - \bar{x}(j))^2, \quad \bar{x}(j) = \frac{1}{n}\sum_{i=1}^{n} x_i(j), \quad j = 1,2,...,k.$$

Таким образом, первое слагаемое в правой части формулы (7) отражает внутригрупповую дисперсию. Наконец, s_1^2 - межгрупповая дисперсия,

$$s_1^2 = \frac{1}{k}\sum_{j=1}^{k}(\bar{x}(j) - \bar{x})^2.$$

Область прикладной статистики, связанную с разложениями дисперсии типа формулы (7), называют дисперсионным анализом. В качестве примера задачи дисперсионного анализа рассмотрим проверку приведенной выше гипотезы H_0 в предположении, что результаты измерений независимы и в каждой выборке имеют нормальное распределение $N(m(j), y^2)$ с одной и той же дисперсией. При справедливости H_0 первое слагаемое в правой части формулы (7), деленное на y^2, имеет распределение хи-квадрат с $k(n\text{-}1)$ степенями свободы, а второе слагаемое, деленное на y^2, также имеет распределение хи-квадрат, но с $(k\text{-}1)$ степенями свободы, причем первое и второе слагаемые независимы как случайные величины. Поэтому случайная величина

$$F = \frac{k(n-1)}{k-1}\frac{(kn)s_1^2}{n\sum_{j=1}^{k} s^2(j)} = \frac{k^2(n-1)s_1^2}{(k-1)\sum_{j=1}^{k} s^2(j)}$$

имеет распределение Фишера с $(k\text{-}1)$ степенями свободы числителя и $k(n\text{-}1)$ степенями свободы знаменателя. Гипотеза H_0 принимается, если $F \le F_{1\text{-}б}$, и отвергается в противном случае, где $F_{1\text{-}б}$ – квантиль порядка 1-б распределения Фишера с указанными числами степеней свободы. Такой выбор критической области определяется тем, что при H_1 величина F безгранично увеличивается при росте объема выборок n. Значения $F_{1\text{-}б}$ берут из соответствующих таблиц [8].

Разработаны непараметрические методы решения классических задач дисперсионного анализа [19], в частности, проверки гипотезы H_0.

Методы классификации. Следующий тип задач многомерного статистического анализа – задачи классификации. Они согласно [2, 20] делятся на три принципиально различных вида – дискриминантный анализ, кластер-анализ, задачи группировки.

Задача дискриминантного анализа состоит в нахождении правила отнесения наблюдаемого объекта к одному из ранее

описанных классов. При этом объекты описывают в математической модели с помощью векторов, координаты которых – результаты наблюдения ряда признаков у каждого объекта. Классы описывают либо непосредственно в математических терминах, либо с помощью обучающих выборок. Обучающая выборка – это выборка, для каждого элемента которой указано, к какому классу он относится.

Рассмотрим пример применения дискриминантного анализа для принятия решений в технической диагностике. Пусть по результатам измерения ряда параметров продукции необходимо установить наличие или отсутствие дефектов. В этом случае для элементов обучающей выборки указаны дефекты, обнаруженные в ходе дополнительного исследования, например, проведенного после определенного периода эксплуатации. Дискриминантный анализ позволяет сократить объем контроля, а также предсказать будущее поведение продукции. Дискриминантный анализ сходен с регрессионным – первый позволяет предсказывать значение качественного признака, а второй – количественного. В статистике объектов нечисловой природы разработана математическая схема, частными случаями которой являются регрессионный и дискриминантный анализы [21].

Кластерный анализ применяют, когда по статистическим данным необходимо разделить элементы выборки на группы. Причем два элемента группы из одной и той же группы должны быть «близкими» по совокупности значений измеренных у них признаков, а два элемента из разных групп должны быть «далекими» в том же смысле. В отличие от дискриминантного анализа в кластер-анализе классы не заданы, а формируются в процессе обработки статистических данных. Например, кластер-анализ может быть применен для разбиения совокупности марок стали (или марок холодильников) на группы сходных между собой.

Другой вид кластер-анализа – разбиение признаков на группы близких между собой. Показателем близости признаков может служить выборочный коэффициент корреляции. Цель кластер-анализа признаков может состоять в уменьшении числа контролируемых параметров, что позволяет существенно сократить затраты на контроль. Для этого из группы тесно связанных между собой признаков (у которых коэффициент корреляции близок к 1 – своему максимальному значению) измеряют значение одного, а значения остальных рассчитывают с помощью регрессионного анализа.

Задачи группировки решают тогда, когда классы заранее не заданы и не обязаны быть «далекими» друг от друга. Примером является группировка студентов по учебным группам. В технике решением задачи группировки часто является параметрический ряд – возможные типоразмеры группируются согласно элементам параметрического ряда. В литературе, нормативно-технических и инструктивно-методических документах по прикладной статистике также иногда используется группировка результатов наблюдений (например, при построении гистограмм).

Задачи классификации решают не только в многомерном статистическом анализе, но и тогда, когда результатами наблюдений являются числа, функции или объекты нечисловой природы. Так, многие алгоритмы кластер-анализа используют только расстояния между объектами. Поэтому их можно применять и для классификации объектов нечисловой природы, лишь бы были заданы расстояния между ними. Простейшая задача классификации такова: даны две независимые выборки, требуется определить, представляют они два класса или один. В одномерной статистике эта задача сводится к проверке гипотезы однородности [2].

Снижение размерности. Третий раздел многомерного статистического анализа – задачи снижения размерности с целью

сжатия информации. Цель их решения состоит в определении набора производных показателей, полученных преобразованием исходных признаков, такого, что число производных показателей значительно меньше числа исходных признаков, но они содержат возможно большую часть информации, имеющейся в исходных статистических данных. Задачи снижения размерности решают с помощью методов многомерного шкалирования, главных компонент, факторного анализа и др. Например, в простейшей модели многомерного шкалирования исходные данные – попарные расстояния $\rho_{ij}, i, j = 1,2,...,k, i \neq j,$ между k объектами, а цель расчетов состоит в представлении объектов точками на плоскости. Это дает возможность в буквальном смысле слова увидеть, как объекты соотносятся между собой. Для достижения этой цели необходимо каждому объекту поставить в соответствие точку на плоскости так, чтобы попарные расстояния s_{ij} между точками, соответствующими объектам с номерами i и j, возможно точнее воспроизводили расстояния c_{ij} между этими объектами. Согласно основной идее метода наименьших квадратов находят точки на плоскости так, чтобы величина

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(s_{ij} - \rho_{ij})^2$$

достигала своего наименьшего значения. Есть и многие другие постановки задач снижения размерности и визуализации данных.

Статистика случайных процессов и временных рядов. Методы статистики случайных процессов и временных рядов применяют для постановки и решения, в частности, следующих задач:

• предсказание будущего развития случайного процесса или временного ряда;

• управление случайным процессом (временным рядом) с целью достижения поставленных целей, например, заданных значений контролируемых параметров;

• построение вероятностной модели реального процесса, обычно длящегося во времени, и изучение свойств этой модели.

Пример 1. При внедрении статистического регулирования технологического процесса необходимо проверить, что в налаженном состоянии математическое ожидание контролируемого параметра не меняется со временем. Если подобное изменение будет обнаружено, то необходимо установить подналадочное устройство.

Пример 2. Следящие системы, например, входящие в состав автоматизированной системы управления технологическим процессом, должны выделять полезный сигнал на фоне шумов. Это – задача оценивания (полезного сигнала), в то время как в примере 1 речь шла о задаче проверки гипотезы.

Методы статистики случайных процессов и временных рядов описаны в литературе [2,20].

Статистика объектов нечисловой природы. Методы статистики объектов нечисловой природы (статистики нечисловых данных, или нечисловой статистики) применяют всегда, когда результаты наблюдений являются объектами нечисловой природы. Например, сообщениями о годности или дефектности единиц продукции. Информацией о сортности единиц продукции. Разбиениями единиц продукции на группы соответственно значения контролируемых параметров. Упорядочениями единиц продукции по качеству или инвестиционных проектов по предпочтительности. Фотографиями поверхности изделия, пораженной коррозией, и т.д. Итак, объекты нечисловой природы – это измерения по качественному признаку, множества, бинарные отношения (разбиения, упорядочения и др.) и многие другие

математические объекты [2]. Они используются в различных вероятностно-статистических методах принятия решений. В частности, в задачах управления качеством продукции, а также, например, в медицине и социологии, как для описания результатов приборных измерений, так и для анализа экспертных оценок.

Для описания данных, являющихся объектами нечисловой природы, применяют, в частности, таблицы сопряженности, а в качестве средних величин – решения оптимизационных задач [2]. В качестве выборочных средних для измерений в порядковой шкале используют медиану и моду, а в шкале наименований – только моду. О методах классификации нечисловых данных говорилось выше.

Для решения параметрических задач оценивания используют оптимизационный подход, метод одношаговых оценок, метод максимального правдоподобия, метод устойчивых оценок. Для решения непараметрических задач оценивания наряду с оптимизационными подходами к оцениванию характеристик используют непараметрические оценки распределения случайного элемента, плотности распределения, функции, выражающей зависимость [2].

В качестве примера методов проверки статистических гипотез для объектов нечисловой природы рассмотрим критерий «хи-квадрат» (обозначают $ч^2$), разработанный К.Пирсоном для проверки гипотезы однородности (другими словами, совпадения) распределений, соответствующих двум независимым выборкам.

Рассматриваются две выборки объемов n_1 и n_2, состоящие из результатов наблюдений качественного признака, имеющего k градаций. Пусть m_{1j} и m_{2j} – количества элементов первой и второй выборок соответственно, для которых наблюдается j–я градация, а p_{1j} и p_{2j} – вероятности того, что эта градация будет принята, для элементов первой и второй выборок, $j = 1, 2, \ldots, k$.

Для проверки гипотезы однородности распределений, соответствующих двум независимым выборкам,

$$H_0: p_{1j} = p_{2j}, j = 1, 2, \ldots, k,$$

применяют критерий ч^2 (хи-квадрат) со статистикой

$$X^2 = n_1 n_2 \sum_{j=1}^{k} \frac{1}{m_{1j} + m_{2j}} \left(\frac{m_{1j}}{n_1} - \frac{m_{2j}}{n_2} \right)^2 .$$

Установлено [9, 11], что статистика X^2 при больших объемах выборок n_1 и n_2 имеет асимптотическое распределение хи-квадрат с $(k - 1)$ степенью свободы.

Таблица 1

Распределения плавок стали по процентному содержанию серы

Содержание серы,	Число плавок	
в %	Завод А	Завод Б
0,00 ч 0,02	82	63
0,02 ч 0,04	535	429
0,04 ч 0,06	1173	995
0,06 ч 0,08	1714	1307

Пример 3. В табл.1 приведены данные о содержании серы в углеродистой стали, выплавляемой двумя металлургическими заводами. Проверим, можно ли считать распределения примеси серы в плавках стали этих двух заводов одинаковыми.

Расчет по данным табл.1 дает $X^2 = 3{,}39$. Квантиль порядка 0,95 распределения хи-квадрат с $k - 1 = 3$ степенями свободы равен $\chi^2_{0,95}(3) = 7{,}8,$ а потому гипотезу о совпадении функций распределения содержания серы в плавках двух заводов нельзя отклонить, т.е. ее следует принять (на уровне значимости б = 0,05).

Выше дано лишь краткое описание содержания прикладной статистики на современном этапе. Подробное изложение конкретных методов содержится в специальной литературе.

Цитированная литература

1. Вероятность и математическая статистика: Энциклопедия / Гл. ред. акад. РАН Ю.В.Прохоров. – М.: Большая Российская энциклопедия, 1999. – 910с.

2. Орлов А.И. Эконометрика. Учебник. 2-е изд. – М.: Экзамен, 2003. - 576 с.

3. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики / Орлов А.И., Фомин В.Н. и др. - М.: ВНИИСтандартизации, 1987. - 62 с.

4. Колмогоров А.Н. Основные понятия теории вероятностей. – М.-Л.: ОНТИ, 1936. - 80 с.

5. Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987. - 304 с.

6. Гнеденко Б.В. Курс теории вероятностей: Учебник. 7-е изд., исправл. - М.: Эдиториал УРСС, 2001. - 320 с.

7. Орлов А.И. Устойчивость в социально-экономических моделях. – М.: Наука, 1979. - 296 с.

8. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965 (1-е изд.), 1968 (2-е изд.), 1983 (3-е изд.).

9. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. – М.: Наука, 1969. - 512 с.

10. Колмогоров А.Н. О логарифмически нормальном законе распределения размеров частиц при дроблении / Доклады АН СССР. 1941. Т.31. С.99-101.

11. Крамер Г. Математические методы статистики. – М.: Мир, 1975. - 648 с.

12. Прохоров Ю.В., Розанов Ю.А. Теория вероятностей. (Основные понятия. Предельные теоремы. Случайные процессы.) – М.: Наука, 1973. - 496 с.

13. Камень Ю.Э., Камень Я.Э., Орлов А.И. Реальные и номинальные уровни значимости в задачах проверки статистических гипотез. - Журнал «Заводская лаборатория». 1986. Т.52. No.12. С.55-57.

14. Орлов А.И. О нецелесообразности использования итеративных процедур нахождения оценок максимального правдоподобия. – Журнал «Заводская лаборатория», 1986, т.52, No.5, с.67-69.

15. Орлов А.И. Распространенная ошибка при использовании критериев Колмогорова и омега-квадрат. – Журнал «Заводская лаборатория».1985. Т.51. No.1. С.60-62.

16. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. – 900 с.

17. Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980. - 456 с.

18. Математическая теория планирования эксперимента / Под ред. С.М.Ермакова. - М.: Наука, 1983. – 392 с.

19. Холлендер М., Вульф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. - 518 с.

20. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. – 736 с.

21. Орлов А.И. Некоторые неклассические постановки в регрессионном анализе и теории классификации. - В сб.: Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. - М.: Наука, 1987. С.27-40.

Контрольные вопросы и задачи

1. Расскажите о понятиях случайного события и его вероятности.

2. Почему закон больших чисел и центральная предельная теорема занимают центральное место в вероятностно-статистических методах?

3. Чем многомерный статистический анализ отличается от статистики объектов нечисловой природы?

4. Имеются три одинаковые с виду ящика. В первом a белых шаров и b черных; во втором c белых и d черных; в третьем только белые шары. Некто подходит наугад к одному из ящиков и вынимает из нее один шар. Найдите вероятность того, что этот шар белый.

5. Пассажир может воспользоваться трамваями двух маршрутов, следующих с интервалами T_1 и T_2 соответственно. Пассажир может прийти на остановку в некоторый произвольный момент времени. Какой может быть вероятность того, что пассажир, пришедший на остановку, будет ждать не дольше t, где $0<t<\min(T_1,T_2)$?

6. Два стрелка, независимо один от другого, делают по два выстрела (каждый по своей мишени). Вероятность попадания в мишень при одном выстреле для первого стрелка p_1, для второго p_2.Выигравшим соревнование считается тот стрелок, в мишени которого будет больше пробоин. Найти вероятность того, что выиграет первый стрелок.

7. Полная колода карт(52 листа) делится наугад на две равные пачки по 26 листов. Найти вероятности следующих событий:

A - в каждой из пачек окажется по два туза;

B - в одной из пачек не будет ни одного туза, а в другой все четыре;

C - в одной из пачек будет один туз, а в другой три.

8. Случайная величина X принимает значения 0 и 1, а случайная величина Y - значения (-1), 0 и 1. Вероятности $P(X=i,\ Y=j)$ задаются таблицей:

$P(X=i,\ Y=j)$	Y = - 1	Y = 0	Y = 1

X = 0	1/16	1/4	1/16
X = 1	1/16	1/4	5/16

Найдите распределение случайной величины $Z = XY$, ее математическое ожидание и дисперсию.

9. В условиях задачи 8 найдите распределение случайной величины $W = X/(Y+3)$, ее математическое ожидание и дисперсию.

10. Даны независимые случайные величины X и Y такие, что $M(X) = 1$, $D(X) = 3$, $M(Y) = -1$, $D(Y) = 2$. Найдите $M(aX + bY)$ и $D(aX + bY)$, где $a= 3$, $b= -2$.

Темы докладов, рефератов, исследовательских работ

1. Описание данных с помощью гистограмм и непараметрических оценок плотности.
2. Сравнительный анализ методов оценивания параметров и характеристик.
3. Преимущества одношаговых оценок по сравнению с оценками метода максимального правдоподобия.
4. Методы проверки однородности для независимых и связанных выборок.
5. Непараметрический регрессионный анализ.
6. Структура статистики нечисловых данных.
7. Аксиоматическое введение метрик и их использование в статистике объектов нечисловой природы.
8. Законы больших чисел в пространствах произвольной природы.
9. Непараметрические оценки плотности в пространствах произвольной природы, в том числе в дискретных пространствах.
10. Основные идеи статистики интервальных данных.
11. Оптимизационные постановки в вероятностно-статистических задачах принятия решений.

Приложение

Некоторые постановки задач прикладной статистики

Чтобы дать представление о богатом содержании теории рассматриваемых методов, приведем краткий перечень основных типов постановок задач прикладной статистики, широко используемых в практической деятельности и в научных исследованиях. Задачи рассмотрим в соответствии с описанной выше классификацией областей прикладной статистики.

1. Одномерная статистика.

1.1. Описание материала

1.1.1. Расчет выборочных характеристик распределения.

1.1.2. Построение гистограмм и полигонов часто.

1.1.3. Приближение эмпирических распределений с помощью распределений из системы Пирсона и других систем…

1.2. Оценивание.

1.2.1. Параметрическое оценивание.

1.2.1.1. Правила определения оценок и доверительных границ для параметров устойчивого распределения.

1.2.1.2. Правила определения оценок и доверительных границ для параметров логистического распределения.

1.2.1.3. Правила определения оценок и доверительных границ для параметров экспоненциального распределения и смеси экспоненциальных распределений… (и так далее для различных семейств распределений).

1.2.2. Непараметрическое оценивание.

1.2.2.1. Непараметрическое точечное и доверительное оценивание основных характеристик распределения – математического ожидания, дисперсии, среднего квадратического отклонения, коэффициента вариации, квантилей, прежде всего медианы.

1.2.2.2. Непараметрические оценки плотности и функции распределения.

1.2.2.3. Непараметрическое оценивание параметра сдвига…

1.3. Проверка гипотез.

1.3.1. Параметрические задачи проверки гипотез.

1.3.1.1. Проверка равенства математических ожиданий для двух нормальных совокупностей.

1.3.1.2. Проверка равенства дисперсий для двух нормальных совокупностей.

1.3.1.3. Проверка равенства коэффициентов вариации для двух нормальных совокупностей.

1.3.1.4. Проверка равенства математических ожиданий и дисперсий для двух нормальных совокупностей.

1.3.1.5. Проверка равенства математического ожидания нормального распределения определенному значению.

1.3.1.6. Проверка равенства дисперсии нормального распределения определенному значению…

1.3.1.7. Проверка равенства параметров двух экспоненциальных совокупностей… (и так далее – проверка утверждений о параметрах для различных семейств распределений).

1.3.2. Непараметрические задачи проверки гипотез.

1.3.2.1. Непараметрическая проверка равенства математических ожиданий для двух совокупностей.

1.3.2.2. Непараметрическая проверка равенства дисперсий для двух совокупностей.

1.3.2.3. Непараметрическая проверка равенства коэффициентов вариации для двух совокупностей.

1.3.2.4. Непараметрическая проверка равенства математических ожиданий и дисперсий для двух совокупностей.

1.3.2.5. Непараметрическая проверка равенства математического ожидания определенному значению.

1.3.2.6. Непараметрическая проверка равенства дисперсии определенному значению…

1.3.2.7. Проверка гипотезы согласия с равномерным распределением по критерию Колмогорова.

1.3.2.8. Проверка гипотезы согласия с равномерным распределением по критерию омега-квадрат (Крамера – Мизеса - Смирнова).

1.3.2.9. Проверка гипотезы согласия с равномерным распределением по критерию Смирнова.

1.3.2.10. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа Колмогорова при известной дисперсии.

1.3.2.11. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа Колмогорова при известном математическом ожидании.

1.3.2.12. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа Колмогорова (оба параметра неизвестны).

1.3.2.13. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа омега-квадрат при известной дисперсии.

1.3.2.14. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа омега-квадрат при известном математическом ожидании.

1.3.2.15. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа омега-квадрат (оба параметра неизвестны).

1.3.2.16. Проверка гипотезы согласия с экспоненциальным семейством распределений по критерию типа омега-квадрат… (и так далее для различных семейств распределений, тех или иных предположениях о параметрах, всевозможных критериев).

1.3.2.17. Проверка гипотезы однородности двух выборок методом Смирнова.

1.3.2.18. Проверка гипотезы однородности двух выборок методом омега-квадрат.

1.3.2.19. Проверка гипотезы однородности двух выборок с помощью критерия Вилкоксона.

1.3.2.20. Проверка гипотезы однородности двух выборок по критерию Ван-дер-Вардена.

1.3.2.21. Проверка гипотезы симметрии функции распределения относительно 0 методом Смирнова.

1.3.2.22. Проверка гипотезы симметрии функции распределения относительно 0 с помощью критерия типа омега-квадрат (Орлова).

1.3.2.23. Проверка гипотезы независимости элементов выборки.

1.3.2.24. Проверка гипотезы одинаковой распределенности элементов выборки…(и т.д.).

2. Многомерный статистический анализ.

2.1. Описание материала.

2.1.1. Расчет выборочных характеристик (вектора средних, ковариационной и корреляционной матриц и др.).

2.1.2. Таблицы сопряженности.

2.1.3. Детерминированные методы приближения функциональной зависимости.

2.1.3.1. Метод наименьших квадратов.

2.1.3.2. Метод наименьших модулей

2.1.3.3. Сплайны и др.

2.1.4. Методы снижения размерности.

2.1.4.1. Алгоритмы факторного анализа.

2.1.4.2. Алгоритмы метода главных компонент

2.1.4.3. Алгоритмы многомерного метрического шкалирования.

2.1.4.4. Алгоритмы многомерного неметрического шкалирования.

2.1.4.5. Методы оптимального проецирования и др.

2.1.5. Методы классификации.

2.1.5.1. Методы кластер-анализа – иерархические процедуры.

2.1.5.2. Методы кластер-анализа – оптимизационный подход.

2.1.5.3. Методы кластер-анализа – итерационные процедуры…

2.1.5.4. Методы группировки…

2.2. Оценивание.

2.2.1. Параметрическое оценивание.

2.2.1.1. Оценивание параметров многомерного нормального распределения.

2.2.1.2. Оценивание параметров в нормальной модели линейной регрессии.

2.2.1.3. Методы расщепления смесей.

2.2.1.4. Оценивание компонент дисперсии в дисперсионном анализе (в нормальной модели).

2.2.1.5. Оценивание размерности и структуры модели в регрессионном анализе (в нормальной модели).

2.2.1.6. Оценивание в дискриминантном анализе (в нормальной модели).

2.2.1.7. Оценивание в методах снижения размерности (в нормальной модели).

2.2.1.8. Нелинейная регрессия.

2.2.1.9. Методы планирования эксперимента.

2.2.2. Непараметрическое оценивание.

2.2.2.1. Непараметрические оценки многомерной плотности.

2.2.2.2. Непараметрическая регрессия (с погрешностями наблюдений произвольного вида).

2.2.2.3. Непараметрическая регрессия (на основе непараметрических оценок многомерной плотности).

2.2.2.4. Монотонная регрессия.

2.2.2.5. Непараметрический дискриминантный анализ.

2.2.2.6. Непараметрический дисперсионный анализ…

2.3. Проверка гипотез.

2.3.1. Параметрические задачи проверки гипотез.

2.3.1.1. Корреляционный анализ (нормальная модель).

2.3.1.2. Проверка гипотез об отличии коэффициентов при предикторах от 0 в линейной регрессии при справедливости нормальной модели.

2.3.1.3. Проверка гипотезы о равенстве математических ожиданий нормальных совокупностей (дисперсионный анализ).

2.3.1.4. Проверка гипотезы о совпадении двух линий регрессии (нормальная модель)…(и т.д.)

2.3.2. Непараметрические задачи проверки гипотез.

2.3.2.1. Непараметрический корреляционный анализ.

2.3.2.2. Проверка гипотез об отличии коэффициентов при предикторах от 0 в линейной регрессии (непараметрическая постановка).

2.3.2.3. Проверка гипотез в непараметрическом дисперсионном анализе.

2.3.2.4. Проверка гипотезы о совпадении двух линий регрессии (непараметрическая постановка)…(и т.д.)

Здесь остановимся, поскольку продолжение предполагало бы знакомство со многими достаточно сложными методами, о которых нет упоминаний в этой книге. Приведенный выше перечень ряда основных типов постановок задач, используемых в прикладной статистике, дает первоначальное представление об объеме арсенала разработанных к настоящему времени интеллектуальных инструментов в рассматриваемой области.

Об авторе

Орлов Александр Иванович, 1949 г.р., профессор (1995 г. – по кафедре математической экономики), доктор технических наук (1992 г. - по применению математических методов), кандидат физико-математических наук (1976 г. – по теории вероятностей и математической статистике). Профессор кафедры "Экономика и организация производства" факультета "Инженерный бизнес и менеджмент" Московского государственного технического университета им. Н.Э. Баумана, руководитель секции "Эконометрика", директор Института высоких статистических технологий и эконометрики. Профессор кафедры "Экология и право" Московского государственного института электроники и математики (технического университета), научный руководитель Лаборатории эконометрических исследований. Профессор кафедры «Анализ стохастических процессов в экономике» Российской экономической академии им. Г.В. Плеханова. Визит-профессор (в 2002-2003 г.г.) Академии народного хозяйства при Правительстве Российской Федерации, Международного университета (в Москве), Всероссийского государственного института кинематографии (ВГИК), Московского государственного университета прикладной биотехнологии, Международного юридического института при Министерстве юстиции Российской Федерации. Член редколлегий научных журналов «Заводская лаборатория», «Контроллинг», «Социология: методология, методы, математические модели». Главный редактор электронного еженедельника «Эконометрика». Академик Российской Академии статистических методов, член-корреспондент МО "СовАсК" (Международной организации «Советская Ассоциация Качества»). Вице-президент Всесоюзной Статистической Ассоциации, президент Российской ассоциации статистических методов.

В 1966 г. окончил физматшколу № 2 г. Москвы, в 1971 г. – механико-математический факультет МГУ им. М.В.Ломоносова. В 1971-1978 гг. работал в Центральном экономико-математическом институте АН СССР, в 1978-1981 г. – «кремлевской больнице», в 1981-1989 г. – во ВНИИ стандартизации Госстандарта СССР. Создал и руководил Всесоюзным центром статистических методов и информатики (1989-1992). С 1993 г. – на преподавательской работе, профессор ряда московских вузов.

Основные направления научной и педагогической деятельности: прикладная статистика и другие статистические методы, эконометрика, экономико-математические методы, теория принятия решений, экспертные оценки, менеджмент, экономика предприятия, макроэкономика, экология, социология. Автор более 500 публикаций в России и за рубежом, в том числе более 20 книг.

Основные книги проф. А.И.Орлова

1. Устойчивость в социально-экономических моделях. - М.: Наука,1979. - 296 с.
2. Задачи оптимизации и нечеткие переменные. - М.: Знание, 1980. - 64 с.
3. Анализ нечисловой информации (препринт) (совместно с Ю.Н. Тюриным, Б.Г. Литваком, Г.А. Сатаровым, Д.С. Шмерлингом). - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981. - 80 с.
4. Внеклассная работа по математике в 6-8 классах (совместно с В.А. Гусевым, А.Л. Розенталем). - М.: Просвещение, 1977. - 288 с. - Второе издание, исправленное и дополненное (М.: Просвещение, 1984). Переводы на казахский, литовский, молдавский языки.
5. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-

распределения (в соавторстве). - М.: Изд-во стандартов, 1984. - 53 с. - Переиздание: М.: Изд-во стандартов, 1985, - 50 с.

6. Анализ нечисловой информации в социологических исследованиях (в соавторстве). - М.: Наука, 1985. - 220 с.

7. Пакет программ анализа данных "ППАНД". Учебное пособие (совместно с И.Л. Легостаевой, О.М. Черномордиком и др.). - М.: Сотрудничающий центр Всемирной организации здравоохранения по профессиональной гигиене, 1990. - 93 с.

8. О теоретических основах внеклассной работы по математике и опыте Вечерней математической школы при Московском математическом обществе. - М.: Всесоюзный центр статистических методов и информатики, 1991. - 48 с.

9. Математическое моделирование процессов налогообложения (подходы к проблеме) (совместно с В. Г. Кольцовым, Н.Ю. Ивановой и др.). - М.: Изд-во ЦЭО Министерства общего и профессионального образования РФ, 1997. – 232 с.

10. Экология. Учебное пособие (совместно с С.А. Боголюбовым и др.). - М.: Знание, 1999. - 288 с.

11. Менеджмент. Учебное пособие (совместно с С.А. Боголюбовым, Ж.В.Прокофьевой и др.). - М.: Знание, 2000. - 288 с.

12. Управление качеством окружающей среды. Учебник. Т.1 (совместно с С.А.Боголюбовым и др.). - М.: МГИЭМ(ту), 2000. – 283 с.

13. Системы экологического управления: Учебник (совместно с С.А.Боголюбовым и др.). - М.: Европейский центр по качеству, 2002. – 224 с.

14. Эконометрика. Учебник для вузов. – М.: Экзамен, 2002 (1-е изд.), 2003 (2-е изд.). – 576 с.

15. Управление промышленной и экологической безопасностью: Учебное пособие (совместно с В.Н. Федосеевым, В.Г. Ларионовым,

А.Ф. Козьяковым). - М.: УРАО, 2002 (1-е изд.), 2003 (2-е изд.). – 220 с.

16. Менеджмент в техносфере. Учебное пособие (совместно с В.Н. Федосеевым). – М.: Академия, 2003. – 384 с.

Уважаемые читатели!

Предлагаемая книга входит в новую серию «Статистические методы» издательства «МЗ-Пресс». В этой серии будут выпускаться научные монографии по различным теоретическим и прикладным направлениям статистических методов, учебники и учебные пособия, написанные ведущими исследователями. Основная цель серии – выпуск научных монографий, являющихся одновременно учебниками и позволяющих студентам и специалистам выйти на передовой фронт современных исследований.

Книги серии посвящены прикладной статистике и другим статистическим методам обработки и анализа данных, а также применению статистических методов в технических, социально-экономических, медицинских, исторических и иных исследованиях. Они окажутся полезными для инженеров, экономистов, менеджеров, социологов, врачей, всех научных работников и специалистов, чья профессиональная деятельность связаны с обработкой и анализом данных.

Редакционный совет серии создан Правлением Российской ассоциации статистических методов (учреждена в 1990 г.). По оценке Правления, выпуск серии «Статистические методы» позволит заметно повысить научный уровень и практическую значимость отечественных научных исследований, прикладных разработок и преподавания в области статистических методов.

Надеемся, что новая серия привлечет внимание и будет полезна как студентов и преподавателям, так и профессиональным исследователям. Желаем всем потенциальным читателям найти что-то полезное для себя.