МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)

О.Н. САПРЫКИН

СТАТИСТИЧЕСКИЙ АНАЛИЗ РИСКОВ В СИСТЕМАХ КОМПЛЕКСНОЙ БЕЗОПАСНОСТИ

Рекомендовано редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева» в качестве учебного пособия для обучающихся по основной образовательной программе высшего образования по направлению подготовки 23.03.01 Технологии транспортных процессов

САМАРА
Издательство Самарского университета
2020

УДК 656.05(075)+004.056(075)
ББК 65.37я7+32.973-018.2я7
С197

Рецензенты: д-р техн. наук, проф. А.Н. Коптев,
канд. экон. наук А.В. Зиновьев

Сапрыкин, Олег Николаевич

С197 Статистический анализ рисков в системах комплексной безопасности: учебное пособие / *О.Н. Сапрыкин*. – Самара: Издательство Самарского университета, 2020. – 72 с.: ил.

ISBN 978-5-7883-1565-2

Изложены основные методики анализа рисков при организации проактивных средств защиты информации ИТ-инфраструктуры предприятия, модели экспертных систем и методы автоматического вывода правил из накопленных данных. Рассмотрены вопросы оценки рисков с помощью статистического анализа данных с применением метода регрессионного анализа.

Содержание учебного пособия соответствует тематике лекций для бакалавров по дисциплине «Основы информационной безопасности на воздушном транспорте», читаемых автором в Самарском университете.

Предназначено для студентов направления подготовки 23.03.01 Технологии транспортных процессов.

Подготовлено на кафедре организации и управления перевозками на транспорте.

УДК 656.05(075)+004.056(075)
ББК 65.37я7+32.973-018.2я7

ISBN 978-5-7883-1565-2

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Функционирование любого предприятия в современном мире сложно представить без телефонов, компьютеров и Интернета. Не составляют исключения и транспортные предприятия, информационная инфраструктура которых содержит, помимо систем общего назначения, специализированные системы, обеспечивающие контроль и управление процессом перевозки. Глубокая интеграция информационных технологий в бизнес-процессы предприятий ведет к уязвимости последних к любым сбоям и нарушениям в информационной инфраструктуре. Особенно стоит отметить сохранность конфиденциальной информации, нарушение которой может нанести серьезный ущерб работе предприятия, его репутации и безопасности физических и юридических лиц, взаимодействующих с ним. Предупредить несанкционированный доступ позволяют системы безопасности нового поколения. Они анализируют сетевой трафик с целью выявления и блокировки пакетов, несущих угрозу.

Разработка подобных систем информационной защиты требует подготовки высококвалифицированных специалистов в области статистического анализа рисков в системах комплексной безопасности. Целью данного учебного пособия является методическая поддержка подготовки указанных специалистов.

Содержание пособия соответствует тематике лекций для бакалавров направления подготовки «Технологии транспортных процессов» по дисциплине «Информационные системы и технологии в комплексной безопасности», читаемых автором в Самарском университете. Вопросы, затронутые в учебном пособии, могут быть интересны также магистрантам и аспирантам.

Глубокую благодарность автор выражает доктору технических наук, профессору, профессору кафедры эксплуатации авиационной техники Самарского университета А.Н. Коптеву и генеральному директору ООО «Средневолжская логистическая компания», кандидату экономических наук А.В. Зиновьеву за полезные замечания, сделанные при рецензировании рукописи.

ПЕРЕЧЕНЬ ОСНОВНЫХ СОКРАЩЕНИЙ

ДИ – Доверительный интервал
МНК – Метод наименьших квадратов
ОО – Объект оценки
СКО – Среднеквадратическое отклонение
СФБ – Стойкость функции безопасности
ЦТП – Центральная предельная теорема
CI – Confidence Interval
RSS – Reduced Sum of Squares
SSE – Sum of Squared Errors

ВВЕДЕНИЕ

Для современных информационных технологий характерен экспоненциальный рост объема передаваемой и хранимой информации. С точки зрения обеспечения безопасности это имеет как отрицательные, так и положительные стороны. С одной стороны, бесперебойная работа системы, оперирующей большими потоками данных, является сложной задачей. С другой стороны, данные могут быть использованы для вывода закономерностей в цепочке сообщений и выявить подозрительный трафик.

Под системой защиты информации понимается организованная совокупность специальных органов, средств, методов и мероприятий, обеспечивающих защиту информации от внутренних и внешних угроз. Модель защиты информации можно разложить на следующие компоненты: объекты угроз, угрозы, источники угроз, цели угроз со стороны злоумышленников, источники информации, способы неправомерного овладения конфиденциальной информацией, направления защиты информации, способы защиты информации, средства защиты информации.

Объектом угроз информационной безопасности выступают сведения о составе, состоянии и деятельности объекта защиты (персонала, материальных и финансовых ценностей, информационных ресурсов). Угрозы информации выражаются в нарушении ее целостности, конфиденциальности, полноты и доступности. Источниками угроз выступают конкуренты, преступники, коррупционеры, административно-управленческие органы. Источники угроз преследуют при этом следующие цели: ознакомление с охраняемыми сведениями, их модификация в корыстных целях и уничтожение для нанесения прямого материального ущерба.

Неправомерное овладение конфиденциальной информацией возможно за счет ее разглашения источниками сведений, за счет утечки информации через технические средства и за счет несанкционированного доступа к охраняемым сведениям. Источниками конфиденциальной информации являются люди, документы, публикации, технические носители информации, технические средства обеспечения производственной и трудовой деятельности, продукция и отходы производства. Основными направлениями защиты информации являются правовая, организационная и инженерно-техническая защиты информации как выразители комплексного подхода к обеспечению информационной безопасности. Средствами защиты информации являются физические средства, аппаратные средства, программные средства и криптографические методы. Последние могут быть реализованы как аппаратно, программно, так и смешанно – программно-аппаратными средствами. В качестве способов защиты выступают всевозможные меры, пути, способы и действия, обеспечивающие упреждение противоправных действий, их предотвращение, пресечение и противодействие несанкционированному доступу.

Под угрозами конфиденциальной информации принято понимать потенциальные или реально возможные действия по отношению к информационным ресурсам, приводящие к неправомерному овладению охраняемыми сведениями. Такими действиями являются:

- ознакомление с конфиденциальной информацией различными путями и способами без нарушения ее целостности;
- модификация информации в криминальных целях как частичное или значительное изменение состава и содержания сведений;
- разрушение (уничтожение) информации как акт вандализма с целью прямого нанесения материального ущерба.

В конечном итоге противоправные действия с информацией приводят к нарушению ее конфиденциальности, полноты, достоверности и доступности, что, в свою очередь, приводит к наруше-

нию как режима управления, так и его качества в условиях ложной или неполной информации.

Своевременно организованные контрмеры могут предотвратить негативные последствия угроз. Современные системы оценки рисков безопасности основаны на моделировании угроз и глубоком анализе сетевого трафика. В первой главе данного учебного пособия рассматриваются методики, основанные на моделировании. Во второй главе рассматриваются методы статистического анализа для идентификации угроз. Третья глава знакомит читателя с деталями регрессионного анализа и его применения в системах комплексной безопасности. Четвертая глава посвящена вопросам временного анализа рисков с помощью моделей выживания.

1 МЕТОДИКИ АНАЛИЗА РИСКОВ БЕЗОПАСНОСТИ

1.1 Функция безопасности

Анализ рисков безопасности предприятия главным образом опирается на оценку вероятности успеха атаки на информационную систему при попытке ее реализации. Этот показатель зависит от того, насколько эффективно реализуются функции безопасности объекта.

Стойкость функции безопасности (СФБ) – это характеристика функции безопасности объекта оценки (ОО), выражающая минимальные усилия, предположительно необходимые для нарушения ее ожидаемого безопасного поведения при прямой атаке на лежащие в ее основе механизмы безопасности. Процедуры анализа стойкости функции безопасности (СФБ) объекта оценки (ОО) и анализа уязвимостей используют понятие «потенциал нападения» [1].

Потенциал нападения – это прогнозируемый потенциал для успешного (в случае реализации) нападения, выраженный в показателях компетентности, ресурсов и мотивации.

СФБ может быть базовой, средней и высокой:

- Базовая стойкость – функция обеспечивает адекватную защиту от случайного нарушения безопасности ОО нарушителем с низким потенциалом нападения.
- Средняя стойкость – функция обеспечивает защиту от целенаправленного нарушения безопасности ОО нарушителем с умеренным потенциалом нападения.
- Высокая стойкость – уровень стойкости функции безопасности ОО, на котором она обеспечивает защиту от тщательно спланированного и организованного нарушения безопасности ОО нарушителем с высоким потенциалом нападения.

Потенциал нападения зависит от компетенции ресурсов и мотивации нарушителя. Мотивация – фактор потенциала нападения, который может использоваться, чтобы описать разные аспекты, связанные с нарушителем и активами, которые его интересуют. Мотивация может:

- косвенно выражать вероятность нападения;
- быть связана с ценностью актива (хотя ценность актива может быть субъективна);
- быть связана с компетентностью и ресурсами нарушителя.

Вычисление потенциала нападения может производиться следующим образом. Предполагается, что для реализации угрозы нарушитель сначала должен выявить соответствующую уязвимость. Поэтому, при анализе потенциала нападения учитываются следующие факторы:

1) при идентификации уязвимости:
   - время, затрачиваемое на идентификацию уязвимости (x_1) («за минуты», «за часы», «за дни», «за месяцы»);
   - уровень специальной подготовки (x_2) («эксперт», «специалист», «неспециалист»);
   - знание проекта и функционирования ОО (x_3) («отсутствие информации об ОО», «общедоступная информация об ОО», «закрытая информация об ОО»);
   - доступ к ОО (x_4) (требуемое время на доступ к ОО, как в случае x_1);
   - аппаратные средства, программное обеспечение или другое оборудование (x_5) («стандартное оборудование», «специализированное оборудование», «уникальное оборудование»).
2) При использовании:
   - время, затраченное на использование уязвимости (y_1);
   - уровень специальной подготовки (y_2);
   - знание проекта функционирования ОО (y_3);
   - доступ к ОО (y_4);
   - аппаратные средства, программное обеспечение или другое оборудование, необходимое для использования уязвимости (y_5).

Далее 10 факторам x_1-x_5 и y_1-y_5 назначаются веса. Анализ факторов проводится либо методами статистического анализа (см. глава 2), либо методами машинного обучения [2]. Полученные на выходе характеристики используются для оценки уязвимости (потенциала нападения и СФБ ОО).

Анализ рисков безопасности является базовым элементом модели комплексной безопасности предприятия. Рассмотрим некоторые из подобных моделей.

1.2 Модель безопасности Lifecycle Security

Модель Lifecycle Security разработана компанией Axent, впоследствии приобретенной Symantec. Lifecycle Security – это обобщенная схема построения комплексной защиты компьютерной сети предприятия. Выполнение, описываемого в ней набора процедур, позволяет системно решать задачи, связанные с защитой информации, и дает возможность оценить эффект от затраченных средств и ресурсов. Идеология Lifecycle Security может быть противопоставлена тактике «точечных решений», заключающейся в том, что все усилия сосредотачиваются на внедрении отдельных частных решений (например, межсетевых экранов или систем аутентификации пользователей по смарт-картам). Без предварительного анализа и планирования подобная тактика может привести к появлению в компьютерной системе набора разрозненных продуктов, которые не стыкуются друг с другом и не позволяют решить проблемы предприятия в сфере информационной безопасности.

Lifecycle Security включает в себя 7 основных компонентов (рис 1).

Политики безопасности, стандарты, процедуры и метрики. Этот компонент определяет рамки, в которых осуществляются мероприятия по обеспечению безопасности информации, и задает критерии оценки полученных результатов. Под стандартами здесь понимаются не только государственные и международные стандарты в сфере информационной безопасности, но и корпоративные

Рис. 1. Компоненты модели Lifecycle Security

стандарты, которые в ряде случаев могут оказать очень существенное влияние на создаваемую систему защиты информации. Метрики позволяют оценить состояние системы до и после проведения работ по защите информации. Метрика определяет, в чем и как измеряем защищенность системы, и позволяет соотнести сделанные затраты и полученный эффект.

Анализ рисков. Этот этап является отправной точкой для установления и поддержания эффективного управления системой защиты. Проведение анализа рисков позволяет подробно описать состав и структуру информационной системы, расположить имеющиеся ресурсы по приоритетам, основываясь на степени их важности для нормальной работы предприятия, оценить угрозы и идентифицировать уязвимости системы.

Стратегический план построения системы защиты. Результаты анализа рисков используются как основа для разработки стратегического плана построения системы защиты. Наличие подобного плана помогает распределить по приоритетам бюджеты и ресурсы, и в последующем осуществить выбор продуктов и разработать стратегию их внедрения.

Выбор и внедрение решений. Хорошо структурированные критерии выбора решений в сфере защиты информации и наличие программы внедрения уменьшает вероятность приобретения продуктов, становящихся «мертвым грузом», мешающим развитию информационной системы предприятия. Кроме непосредственно выбора решений, также должно учитываться качество предоставляемых поставщиками сервисных и обучающих услуг. Кроме того, необходимо четко определить роль внедряемого решения в выполнении разработанных планов и достижении поставленных целей в сфере безопасности.

Обучение персонала. Знания в области компьютерной безопасности и технические тренинги необходимы для построения и обслуживания безопасной вычислительной среды. Усилия, затраченные на обучение персонала, значительно повышают шансы на успех мероприятий по защите сети.

Мониторинг защиты помогает обнаруживать аномалии или вторжения в ваши компьютеры и сети и является средством контроля над системой защиты, чтобы гарантировать эффективность программ защиты информации.

Разработка методов реагирования в случае инцидентов и восстановление. Без наличия заранее разработанных и «отрепетированных» процедур реагирования на инциденты в сфере безопасности невозможно гарантировать, что в случае обнаружения атаки ей будут противопоставлены эффективные меры защиты, и работоспособность системы будет быстро восстановлена.

Все компоненты программы взаимосвязаны и предполагается, что процесс совершенствования системы защиты идет непрерывно.

По мнению разработчиков модели Lifecycle Security, он должен проводиться в следующих случаях:

- до и после обновления или существенных изменений в структуре системы;
- до и после перехода на новые технологии;
- до и после подключения к новым сетям (например, подключения локальной сети филиала к сети головного офиса);

- до и после подключения к глобальным сетям (в первую очередь, Интернет);
- до и после изменений в порядке ведения бизнеса (например, при открытии электронного магазина);
- периодически, для проверки эффективности системы защиты.

Ключевые моменты этапа анализа рисков:

1. Подробное документирование компьютерной системы предприятия. При этом особое внимание необходимо уделять критически важным приложениям.
2. Определение степени зависимости организации от нормального функционирования фрагментов компьютерной сети, конкретных узлов, от безопасности хранимых и обрабатываемых данных.
3. Определение уязвимых мест компьютерной системы.
4. Определение угроз, которые могут быть реализованы в отношении выявленных уязвимых мест.
5. Определение и оценка всех рисков, связанных с эксплуатацией компьютерной системы.

1.3 Методика управления рисками, предлагаемая Microsoft

Управление рисками в Microsoft рассматривается как одна из составляющих общей программы управления, предназначенной для руководства компаний и позволяющей контролировать ведение бизнеса и принимать обоснованные решения.

Процесс управления рисками безопасности, предлагаемый Майкрософт, включает следующие четыре этапа (рис. 2):

Оценка рисков:

- Планирование сбора данных. Обсуждение основных условий успешной реализации и подготовка рекомендаций.
- Сбор данных о рисках. Описание процесса сбора и анализа данных.
- Установка приоритетов рисков. Подробное описание шагов по качественной и количественной оценке рисков.

Поддержка принятия решений:

- Определение функциональных требований. Определение функциональных требований для снижения рисков.

- Выбор возможных решений для контроля. Описание подхода к выбору решений по нейтрализации риска.
- Экспертиза решения. Проверка предложенных элементов контроля на соответствие функциональным требованиям.
- Оценка снижения риска. Оценка снижения подверженности воздействию или вероятности рисков.
- Оценка стоимости решения. Оценка прямых и косвенных затрат, связанных с решениями по нейтрализации риска.
- Выбор стратегии нейтрализации риска. Определение наиболее экономически эффективного решения по нейтрализации риска путем анализа выгод и затрат.

Реализация контроля. Развертывание и использование решений для контроля, снижающих риск для организации:

- Поиск целостного подхода. Включение персонала, процессов и технологий в решение по нейтрализации риска.
- Организация по принципу многоуровневой защиты. Упорядочение решений по нейтрализации риска в рамках предприятия.

Оценка эффективности программы. Анализ эффективности процесса управления рисками и проверка того, обеспечивают ли элементы контроля надлежащий уровень безопасности:

- Разработка системы показателей рисков. Оценка уровня и изменения риска.
- Оценка эффективности программы. Оценка программы управления рисками для выявления возможностей усовершенствования.

Термины управление рисками и оценка рисков не являются взаимозаменяемыми. Под управлением рисками понимаются общие мероприятия по снижению риска в рамках организации до приемлемого уровня. Управление рисками представляет собой непрерывный процесс, но производимые оценки чаще всего делаются для годичного интервала. Под оценкой рисков понимается процесс выявления и распределения приоритетов рисков для бизнеса, являющийся составной частью управления рисками.

Рис. 2. Процесс управления рисками безопасности, предлагаемый корпорацией Майкрософт

При описании риска делается указание на то, какое влияние он оказывает на бизнес и насколько вероятно данное событие. Компоненты, описывающие риск изображены на рис. 3.

Рис. 3 Компоненты «полной формулировки» риска

На начальном этапе проведения оценки рискам присваиваются значения в соответствии со шкалой: «высокий», «средний» и «низкий». После этого, для выявленных наиболее существенных рисков проводится количественная оценка.

Перед внедрением в организации процесса управления рисками безопасности, предлагаемого корпорацией Майкрософт, необходимо проверить уровень зрелости организации с точки зрения управления рисками безопасности. Организациям, в которых отсутствуют формальные политики или процессы, относящиеся к управлению рисками безопасности, будет очень трудно сразу внедрить все аспекты рассматриваемого процесса. Если окажется, что уровень зрелости является достаточно низким, рассматриваемый процесс можно внедрять последовательными этапами на протяжении нескольких месяцев (например, путем эксплуатации пилотного проекта в отдельном подразделении на протяжении нескольких полных циклов данного процесса). Продемонстрировав эффективность процесса управления рисками безопасности, предлагаемого корпорацией Майкрософт, на примере этого пилотного проекта, группа управления рисками безопасности может перейти к внедрению данного процесса в других подразделениях, постепенно охватывая всю организацию.

Уровни зрелости управления рисками безопасности.

Отсутствует. Политика или процесс не документированы. Ранее организация не знала о деловых рисках, связанных с управлением рисками, и не рассматривала данный вопрос

Узкоспециализированный. Некоторые члены организации признают значимость управления рисками, однако операции по управлению рисками являются узкоспециализированными. Политики и процессы в организации не документированы, процессы не являются полностью повторяемыми. В результате проекты по управлению рисками являются хаотичными и некоординируемыми, а получаемые результаты не измеряются и не подвергаются аудиту.

Повторяемый. Организации известно об управлении рисками. Процесс управления рисками является повторяемым, но развит

слабо. Процесс документирован не полностью, однако соответствующие операции выполняются регулярно, и организация стремится внедрить всеобъемлющий процесс управления рисками с привлечением высшего руководства. В организации не проводится формальное обучение и информирование по управлению рисками; ответственность за выполнение соответствующих мероприятий возложена на отдельных сотрудников.

Наличие определенного процесса. Организация приняла формальное решение об интенсивном внедрении управления рисками для управления программой защиты информации. В организации разработан базовый процесс с четко определенными целями и задокументированными процессами достижения и оценки результатов. Проводится обучение всего персонала основам управления рисками. Организация активно внедряет задокументированные процессы управления рисками.

Управляемый. На всех уровнях организации имеется глубокое понимание управления рисками. В организации существуют процедура управления рисками и четко определенный процесс, широко распространена информация об управлении рисками, доступно подробное обучение, существуют начальные формы измерений показателей эффективности. Программе управления рисками выделен достаточный объем ресурсов, результаты управления рисками оказывают положительное влияние на работу многих подразделений организации, а группа управления рисками безопасности может постоянно совершенствовать свои процессы и средства. В организации используются некоторые технологические средства, помогающие в управлении рисками, однако большая часть (если не подавляющее большинство) процедур оценки рисков, определения элементов контроля и анализа выгод и затрат выполняется вручную.

Оптимизированный. Организация выделила на управление рисками безопасности значительные ресурсы, а сотрудники пытаются прогнозировать, какие проблемы могут встретиться в течение следующих месяцев и лет и каким образом их нужно будет решать.

Процесс управления рисками глубоко изучен и в значительной степени автоматизирован путем применения различных средств (разработанных в организации или приобретенных у сторонних разработчиков). При возникновении проблем в системе безопасности выявляется основная причина возникшей проблемы, и предпринимаются необходимые действия для снижения риска ее повторного возникновения. Сотрудники организации могут проходить обучение, обеспечивающее различные уровни подготовки.

2 СТАТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

Прогнозирование играет большую роль в математике и других областях науки. Некоторые явления можно спрогнозировать с очень высокой точностью. Например:

- законы физики – предсказывание времени падения тела под действием силы тяжести;
- законы химии – предсказывание свойства смеси двух химических компонентов.

Однако результаты некоторых процессов невозможно предсказать с большой точностью. Подбросьте монету и попытайтесь предсказать, какой стороной она упадет на стол. Исход этого явления нельзя спрогнозировать точно, но это вовсе не значит, что нам ничего не известно о характеристиках данного явления. Если подбрасывать монету многократно, то она приблизительно одинаковое количество раз упадет одной и другой стороной вверх. Подбрасывание монеты является классическим примером случайного явления, в котором нельзя точно определить отдельные исходы, но суммарный исход имеет определенную закономерность [3].

Случаи, подобные подбрасыванию монеты – это вычисление, с какой вероятностью произойдет авария и при каких условиях, окупится ли предприятие, также не подчиняются законам физики, но могут быть описаны посредством математической статистики.

2.1 Общие понятия

Математическая статистика – наука, занимающаяся методами обработки экспериментальных данных, полученных в результате наблюдений над случайными явлениями. При этом существуют следующие задачи:

- описание явлений – упорядочить статистический материал, представить в удобном для экспериментатора виде (таблица, график, диаграмма);
- анализ и прогноз – приближенная оценка интересующих числовых событий (средняя, дисперсия) и погрешности этих величин;
- выработка оптимальных решений – в результате возникает задача проверки правдоподобности гипотез, решением которой является принятие или неприятие выдвинутой гипотезы [4].

Статистические методы помогают измерить, описать, проанализировать и смоделировать подобную изменчивость даже при наличии ограниченного объема данных. Статистический анализ данных может помочь при формировании лучшего понимания природы, сроков и причин изменчивости, а в дальнейшем при решении и даже предупреждении проблем, связанных с такого рода изменчивостью.

Таким образом, статистические методы позволяют наилучшим образом использовать имеющиеся в распоряжении данные при принятии решений и улучшить качество продукции и процессов на стадиях проектирования, разработки, производства, поставки и технического обслуживания.

Объектом исследования для статистического анализа являются статистические данные, полученные в результате наблюдений или экспериментов. Статистические данные – это совокупность объектов (наблюдений, случаев) и признаков (переменных), их характеризующих. Например, объекты исследования – страны мира и признаки, географические и экономические показатели их характеризующие: континент; высота местности над уровнем моря; расходы общества на здравоохранение, образование, армию; средняя продолжительность жизни; доля безработицы, безграмотных; индекс качества жизни и т.д. [5]. Если же рассматривать в качестве объекта исследования автомобильные аварии, то показателями, их характеризующими, будут являться число жертв, возраст автомобиля, пол и возраст водителя, стаж вождения, тип дорожного покрытия и т.д.

Совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений случайной величины, или совокупность результатов всех мыслимых наблюдений, проводимых в неизменных условиях над одной из случайных величин, связанных с данным видом объектов называют генеральной совокупностью.

Часто генеральная совокупность содержит конечное число объектов. Однако если это число достаточно велико, то иногда в целях упрощения вычислений допускают, что генеральная совокупность состоит из бесчисленного множества объектов. Такое допущение оправдывается тем, что увеличение объема генеральной совокупности (достаточно большого объема) практически не сказывается на результатах обработки данных выборки [6].

Вместо того, чтобы изучать генеральную совокупность объектов, изучают выборку, а затем результаты, полученные по выборке, распространяют на всю совокупность. Выборочные исследования занимают меньше времени, они дешевле, проще и практичнее, чем полное исследование.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности. Например, если из 1000 деталей отобрано для обследования 100 деталей, то объем генеральной совокупности $N = 1000$, а объем выборки $n = 100$.

Число объектов генеральной совокупности N значительно превосходит объем выборки n [7].

Важным свойством выборки является репрезентативность. Репрезентативность – свойство выборки воспроизводить характеристики генеральной совокупности. Таким образом, выборка должна быть копией генеральной совокупности относительно характеристик, существующих для цели исследования [8].

2.2 Виды выборок

Существует два вида выборок: детерминированные и вероятностные [9].

2.2.1 Детерминированная выборка

Выборка состоит из элементов, включенных в нее без учета вероятности их появления. Поскольку детерминированные выборки содержат элементы без учета вероятности их появления, причем в некоторых случаях респонденты участвуют в опросах по собственной инициативе, к ним нельзя применить теорию, разработанную для вероятностных выборок. Типичным примером детерминированных выборок являются нерепрезентативные выборки. Объекты включаются в такие выборки на основе соображений простоты, дешевизны или удобства отбора. Например, многие компании проводят опросы, предоставляя посетителям их Web-страниц возможность заполнить анкету. Такие анкеты позволяют собрать большое количество информации за короткий промежуток времени, однако выборки состоят из ответов пользователей интернета, которые принимают участие в опросе по собственной инициативе.

Нерепрезентативные выборки обладают некоторыми преимуществами, в частности, их можно легко и быстро создавать, не расходуя больших средств. С другой стороны, у них есть два важных недостатка – низкая точность, являющаяся следствием тенденциозности, и ограниченность результатов. Преимущества детерминированных выборок не компенсируют их недостатки. Следовательно, детерминированные выборки следует применять лишь для грубых и недорогих оценок, предназначенных для удовлетворения любопытства, либо в качестве учебного или пилотного проекта, который подлежит дальнейшему уточнению.

2.2.2 Вероятностная выборка

Выборка состоит из элементов, вероятность появления которых известна заранее. Вероятностные выборки следует применять всегда, когда это возможно, поскольку лишь они позволяют сделать корректные статистические выводы о генеральной совокупности. На практике получить истинно вероятностную выборку очень

трудно или просто невозможно. Однако для создания вероятностной выборки необходимо следовать правилам и учитывать любую возможную тенденциозность. Существует четыре вида вероятностных выборок: простая случайная, систематическая, стратифицированная и кластер. Каждой из этих выборок соответствует свой метод выбора, который характеризуется собственной стоимостью, точностью и сложностью.

2.2.2.1 Простая случайная выборка

Вероятность выбора элементов простой случайной выборки из основы совпадает с вероятностью выбора любого другого элемента. Кроме того, вероятность извлечения из основной совокупности любых выборок фиксированного объема является постоянной для данного объема. Простой случайный выбор представляет собой элементарную процедуру, на основе которой создаются более сложные методы выбора.

2.2.2.2 Стратифицированная выборка

При формировании стратифицированной выборки N элементов генеральной совокупности или основы разделяются на отдельные подмножества, или страты, обладающие общими свойствами. Затем к каждому подмножеству применяется простой случайный выбор, и его результаты объединяются в одно целое. Этот метод выбора более эффективен, чем методы простого или систематического выбора, поскольку он обеспечивает большую репрезентативность выборки. Точность оценки параметров генеральной совокупности гарантируется однородностью элементов, принадлежащих одному подмножеству.

2.2.2.3 Кластерная выборка

Для образования кластерной выборки основа, состоящая из N элементов, разбивается на несколько кластеров так, чтобы каждый

кластер отражал свойства всей генеральной совокупности. Затем осуществляется простой случайный выбор кластеров, в которых изучаются все элементы. Кластеры естественным образом получаются при статистическом анализе округов, избирательных участков, городов, районов или семей.

Метод кластерного выбора может оказаться менее дорогостоящим, чем метод простого случайного выбора, особенно если генеральная совокупность распределена по широкому географическому региону. Однако метод кластерного анализа в целом менее эффективен, чем методы простого случайного и систематического выбора, и для получения более точной оценки свойств генеральной совокупности приходится значительно увеличивать объем выборки.

2.2.3 Типы данных

В математической статистике существуют следующие типы данных:

- количественные (объём, цена, количество лошадиных сил);
- номинативные (отражают отношение объекта к некоторому классу: пол водителя, категория водительского удостоверения);
- ранговые (с определенным объектом связано некоторое целочисленное число).

Например:

- длина – количественная, непрерывная величина;
- количество перекрестков со светофорами – количественная, дискретная;
- ночь/день – номинативная;
- ранг маршрута – ранговая;
- количество аварий – количественная, дискретная.

2.3 Основные характеристики математической статистики

2.3.1 Меры центральной тенденции

Различные совокупности данных предполагают разные определения «центрального положения» [10].

Наиболее просто получаемой мерой центральной тенденции является мода**.** Мода (чаще всего обозначается как M) – это такое значение в множестве наблюдений, которое встречается наиболее часто.

В совокупности значений (2, 6, 6, 8, 9, 9, 9, 10) модой является 9, потому что данное значение встречается чаще любого другого. Обратите внимание, что мода представляет собой наиболее частое значение (в данном примере 9), а не частоту появления этого значения (в примере равную 3).

Медиана (обозначается *Md*) – это значение, которое делит упорядоченное множество данных пополам, так что одна половина значений оказывается больше медианы, а другая – меньше.

Если данные содержат нечетное число различных значений, например 11, 13, 18, 19, 20, то медиана есть центральное значение для случая, когда они упорядочены, т. е. *Md* = 18.

Если данные содержат четное число различных значений, например 4, 9, 13, 14, то медиана есть точка, лежащая посредине между двумя центральными значениями, когда они упорядочены: *Md* = (9 + 13)/2 = 11.

Теперь определим третью меру – выборочное среднее (называемое иногда «средним» или «арифметическим средним»). Среднее совокупности *n* значений обозначается через $\bar{X}$ и определяется как:

$$\bar{X} = \frac{(X1 + X2 + \cdots + Xn)}{n}$$

или

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_{\text{i.}}$$

Каждая из приведенных выше мер центральной тенденции обладает характеристиками, которые делают ее ценной в определенных условиях.

Мода вычисляется наиболее просто, ее можно определить на глаз. Кроме того, для очень больших групп данных это достаточно стабильная мера центра распределения. Во многих распределениях значительного числа измерений, используемых в педагогике и психологии, мода близка к двум другим мерам – медиане и среднему.

На величину среднего влияют значения всех результатов, особенно те результаты, которые можно назвать «выбросами», т.е. данные, находящиеся далеко от центра группы оценок. Медиана и мода не требуют для определения всех значений. Посмотрим, что произойдет, например, со средним, медианой и модой, когда удвоится максимальное значение в следующем множестве результатов наблюдений, приведенных в табл. 1:

Таблица 1. **Результаты наблюдений**

	Среднее	**Медиана**	**Мода**
1 множество: 1, 3, 3, 5, 6, 7, 8	4,7	5	3
2 множество: 1, 3, 3, 5, 6, 7, 16	5,9	5	3

2.3.2 Меры изменчивости

К мерам изменчивости переменной относятся следующие характеристики: размах, дисперсия, среднее квадратическое (стандартное) отклонение.

Размах измеряет на числовой шкале расстояние, в пределах которого изменяются оценки. Размах представляет собой меру рассеяния, разброса, неоднородности или изменчивости. Эта величина возрастает с ростом рассеяния и уменьшением однородности. Размах является довольно грубой, но достаточно распространенной мерой изменчивости.

Например, размах значений 0, 2, 3, 5, 8 равен 8 – 0 = 8. Значения: –0,2; 0,4; 0,8; 1,6 имеют размах, равный 1,6 – (–0,2) = 1,8.

Более точной мерой изменчивости является дисперсия (обозначается D, или σ^2_x или s^2_x), которая определяется по формуле:

$$D = \frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n-1}.$$

Ценность дисперсии заключается в том, что, являясь мерой варьирования числовых значений признака вокруг его среднего значения, она измеряет внутреннюю изменчивость значений признака, зависящую от разностей между наблюдениями. Преимущество дисперсии перед другими показателями вариации состоит также и в том, что она разлагается на составные компоненты, позволяя тем самым оценивать влияние различных факторов на величину учитываемого признака.

Мерой изменчивости, тесно связанной с дисперсией, является стандартное отклонение. Среднее квадратическое или стандартное отклонение, обозначаемое Sd (или σ_x), определяется как положительное значение квадратного корня из дисперсии:

$$Sd = \sqrt{D} = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n-1}}.$$

Свойства дисперсии и среднего квадратического отклонения:

$D_{X+C} = D_X$;

$Sd_{X+C} = Sd_X$;

$D_{X \cdot C} = D_X \cdot C^2$;

$Sd_{X \cdot C} = Sd_X \cdot C$.

Стандартное отклонение часто является полезной мерой вариации, так как для многих распределений мы приблизительно знаем, какой процент данных лежит внутри одного, двух, трех и более стандартных отклонений среднего. Например, мы можем знать, что 70 % значений лежит между $\bar{X} - Sd$ и $\bar{X} + Sd$.

Часто для описания совокупности используются понятия квантиль, квартиль и интерквартильный размах.

Квантиль в математической статистике – значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Квантили распределения делят упорядоченное множество значений на равные части.

Квартили (от лат. quarta – четверть) нормального распределения делят множество значений на 4 равные части, то есть во множестве существуют три такие точки, которые делят множество на 4 части:

- 0,25-квантиль называется первым (или нижним) квартилем;
- 0,5-квантиль называется медианой или вторым квартилем;
- 0,75-квантиль называется третьим (или верхним) квартилем.

Интерквартильным размахом (англ. Interquartile range – IQR) называется разность между третьим и первым квартилями, т.е. $X_{0,75} - X_{0,25}$. Вместе медиана и интерквартильный размах могут быть использованы вместо математического ожидания и дисперсии в случае распределений с большими выбросами, либо при невозможности вычисления последних.

Ящик с усами, диаграмма размаха (англ. box plot) – график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей [11].

Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы (рис. 4). Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально (рис. 5). Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы. Построение ящика с усами представлено на рис. 6.

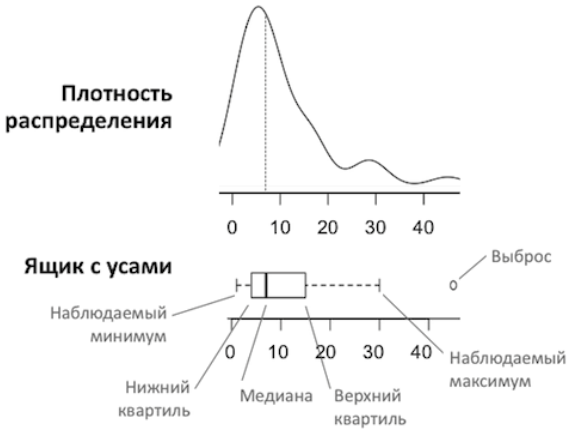

Рис. 4. Ящик с усами

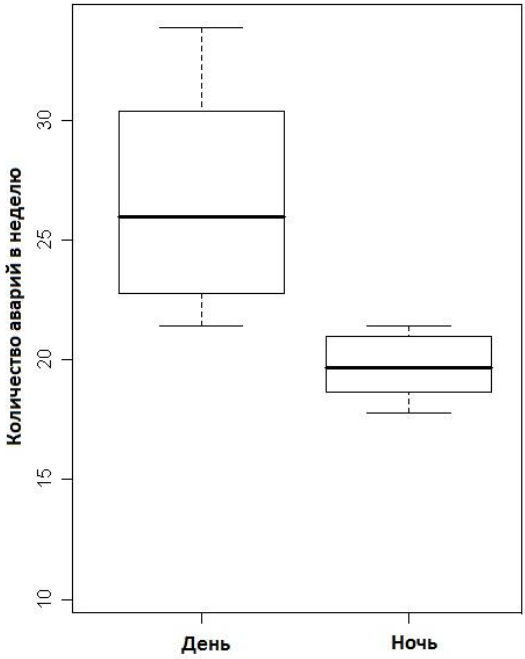

Рис. 5. Пример изображения ящика с усами в программе RStudio

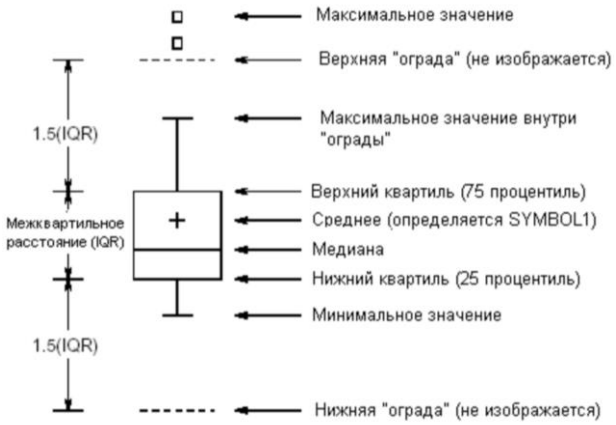

Рис. 6. Построение ящика с усами

2.4 Нормальное распределение

Нормальное распределение, также называемое распределением Гаусса – распределение вероятностей, которое играет важнейшую роль во многих областях знаний, особенно в физике. Физическая величина подчиняется нормальному распределению, когда она подвержена влиянию огромного числа случайных помех. Ясно, что такая ситуация крайне распространена, поэтому можно сказать, что из всех распределений, в природе чаще всего встречается именно нормальное распределение – отсюда и произошло одно из его названий [12].

Нормальное распределение зависит от двух параметров – смещения и масштаба, т.е., является, с математической точки зрения, не одним распределением, а целым их семейством. Значения параметров соответствуют значениям среднего (математического ожидания) и разброса (стандартного отклонения) (рис. 7).

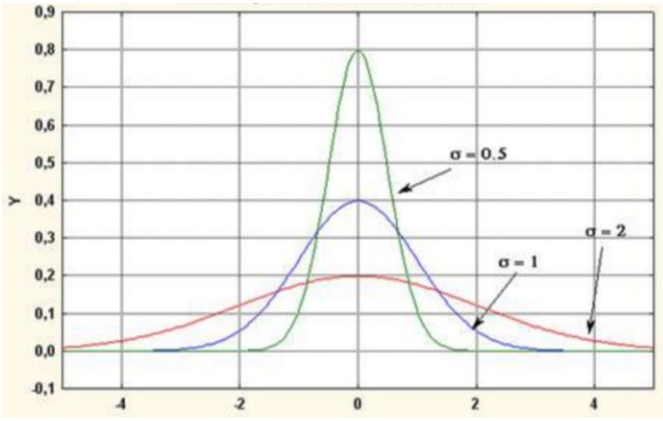

Рис. 7. Графики плотностей вероятностей нормальных распределений с разными стандартными отклонениями

Стандартным нормальным распределением называется нормальное распределение с математическим ожиданием 0 и стандартным отклонением 1.

Свойства нормально распределенной случайной величины X:

- существует только одна мода, то есть распределения является унимодальным;
- плотность симметрична относительно μ, т.е. отклонения x вправо и влево относительно центра μ равновероятны;
- как видно из рис. 4, формально плотность распределения существует для значений X в пределах от (-∞) до (+∞), однако с удалением значений от μ влево или вправо плотность быстро падает.

При удалении от μ на расстояние σ влево или вправо плотность убывает до величины 0,6065 от максимального значения в точке μ. При удалении от μ на 2σ плотность убывает до величины 0,1353, а при удалении на 3σ до величины 0,01111 от максимального значения. С достаточной, для практики, точностью можно считать, что плотность равна нулю при отклонении от μ на расстояние более 4-5 значений σ.

Доля распределения значений в интервале, как функция от ширины интервала, изображена на рис. 8. Ширина интервала задана в определенном количестве значений σ, при этом μ находится в центре интервала.

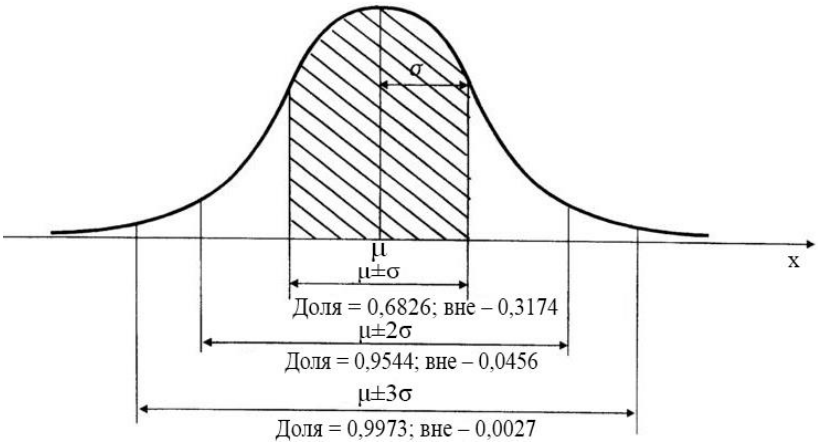

Рис. 8. Доля распределения в интервале,
как функция от ширины интервала

2.4.1 Правило «3 сигма»

Видно, что результат измерения с вероятностью 68% попадет в интервал, т.е. примерно каждое третье измерение даст результат за пределами этого интервала. За пределами интервала $\mu \mp 3\sigma$ окажется один результат из двадцати, а для интервала – только один из трехсот. Значит, интервал ±3σ вокруг среднего значения является почти достоверным, так как подавляющее большинство отдельных результатов многократного измерения случайной величины окажется сосредоточенным именно в нем.

При обработке результатов эксперимента часто используется «правило 3σ», или правило «трех стандартов», которое основано на указанном свойстве нормального распределения. С учетом проведенного выше анализа, можно установить наличие промаха в результате отдельного измерения, а значит, отбросить его, если результат измерения более чем на 3σ отличается от измеренного среднего значения случайной величины.

2.4.2 Z-преобразования

Все кривые плотностей нормальных законов распределения геометрически подобны. Изменяя значение μ, т.е. осуществляя «сдвиг» плотности по оси X, и изменяя σ, т.е. ширину колоколообразной кривой, всегда можно плотность одного нормального распределения привести к плотности другого нормального распределения (рис. 9) [13].

Это важнейшее свойство позволяет любое нормальное распределение преобразовать к стандартному нормальному распределению и наоборот.

Таким образом, любое множество n данных со средним $\bar{X}$ и стандартным отклонением Sd можно преобразовать в другое множество со средним 0 и стандартным отклонением 1 таким образом, что преобразованные значения будут непосредственно выражаться в отклонениях исходных значений от среднего, измеренных в единицах стандартного отклонения. Новые значения называют значениями z:

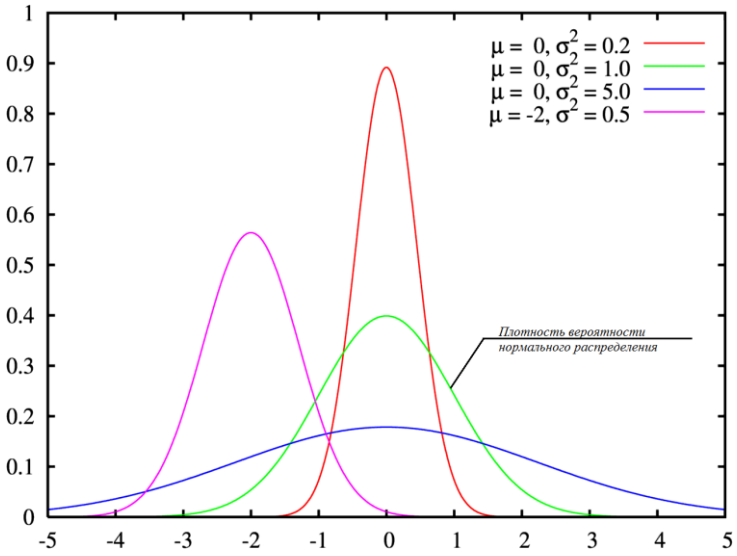

Рис. 9. Преобразование любого нормального распределения к стандартному нормальному распределению

$$Zi = \frac{Xi - \bar{X}}{\sigma},$$

где Zi – элемент преобразованной выборки;

Xi – элемент исходной выборки;

$\bar{X}$ – среднее значение исходной выборки;

σ – СКО исходной выборки

Значение Z – не только удобное средство информации о положении некоторого значения, связанного со средним и измеренного в единицах стандартного отклонения, но и шаг вперед к преобразованию множества X в произвольную шкалу с удобными характеристиками среднего и стандартного отклонения. Сами оценки Z могут не подходить для некоторых целей. Отрицательные оценки, например, могут оказаться неудобными, а множество Z будет, конечно, содержать дроби. Преобразование самих Z позволяет устранить эти несущественные трудности.

2.4.3 Центральная предельная теорема

Простейший вариант центральной предельной теоремы (ЦПТ) теории вероятностей таков:

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

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

где $\Phi(x)$ – функция стандартного нормального распределения [14].

Другими словами, центральная предельная теорема гласит: если случайная величина есть результат совместного действия очень многих факторов, причем:

- ни один из факторов по «силе своего действия» не превосходит многократно остальные факторы;
- факторы действуют независимо друг от друга и не подчиняются какой-то общей тенденции;
- количество факторов достаточно велико,

то независимо от того, какие воздействия производят отдельные факторы, результатом их совместного действия будет нормальное распределение.

Заметим в качестве примера, что если взять несколько, например, 10 случайных величин, имеющих равномерное распределение в интервале [0,1] и вычислить среднее арифметическое от этих равномерно распределенных величин, то это среднее будет иметь почти нормальное распределение, и притом, с очень высокой точностью. То есть центральная предельная теорема здесь срабатывает уже при количестве факторов (случайных чисел), равном всего 10, но все они входят в результат «с равной силой». Такой метод и применяется в ЭВМ для моделирования нормально распределенной случайной величины.

На практике с достаточно хорошей точностью нормальное распределение получается при усреднении всего 4-х равномерно-распределенных случайных величин. Это значит, что средние арифметические значения от выборок объема n=4 или более можно считать нормально распределенными величинами, даже если исходные выборочные данные достаточно далеки от нормального распределения.

Стандартная ошибка среднего Se в математической статистике – величина, характеризующая стандартное отклонение выборочного среднего, рассчитанное по выборке размера n из генеральной совокупности [15].

Стандартная ошибка среднего вычисляется по формуле:

$$Se = \frac{\sigma}{\sqrt{n}},$$

где σ – величина стандартного отклонения генеральной совокупности;

n – объем выборки.

Стандартная ошибка среднего показывает, насколько в среднем отклонение значений в выборке отличается от отклонений в генеральной совокупности.

Рассмотрим пример. Предположим, что существует некоторая генеральная совокупность значений, подчиненных нормальному закону распределения (рис. 10).

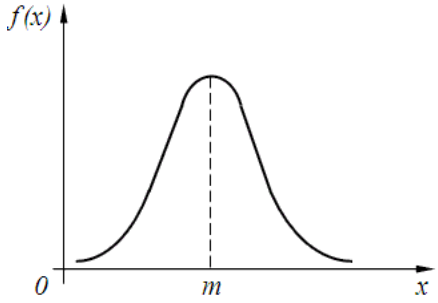

Рис. 10. Вид исходного распределения случайной величины

Последовательно делаем несколько выборок из генеральной совокупности, строим графики (рис. 11).

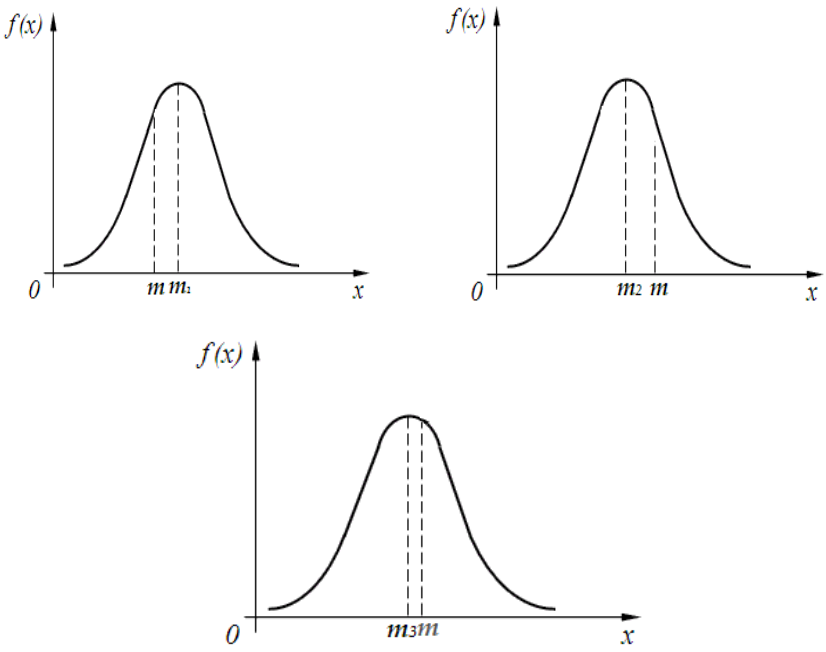

Рис. 11. Графики распределения плотностей вероятностей различных выборок из одной совокупности

Как видно из графиков, согласно центральной предельной теореме, математические ожидания выборок распределяются вокруг математического ожидания генеральной совокупности по нормальному закону. При этом получаем, что распределение случайной величины m также подчиняется нормальному закону.

В таком случае стандартное отклонение σ для генеральной совокупности может быть вычислено как стандартная ошибка среднего:

$$Se = \frac{\sigma 1}{\sqrt{n}},$$

где σ_1 – СКО 1-ой выборки.

2.5 Доверительный интервал для среднего

Для большинства случаев стандартная ошибка как такова не приемлема. Гораздо полезнее объединить эту меру точности с интервальной оценкой для параметра популяции. Это можно сделать, используя знания о теоретическом распределении вероятности выборочной статистики (параметра) для того, чтобы вычислить доверительный интервал (CI – Confidence Interval, ДИ – Доверительный интервал) для параметра.

Доверительный интервал расширяет оценки в обе стороны некоторой величиной, кратной стандартной ошибке (данного параметра); два значения (доверительные границы), определяющие интервал, обычно отделяют запятой и заключают в скобки.

2.5.1 Использование нормального распределения

Выборочное среднее $\bar{X}$ имеет нормальное распределение, если объем выборки большой, поэтому можно применить знания о нормальном распределении при рассмотрении выборочного среднего.

В частности, 95% распределения выборочных средних находится в пределах 1,96 стандартных отклонений (Sd) среднего совокупности [16].

Когда у нас есть только одна выборка, мы называем это стандартной ошибкой среднего (Se) и вычисляем 95% доверительного интервала для среднего следующим образом:

$$(\bar{X} - 1{,}96Se;\ \bar{X} + 1{,}96Se).$$

Если повторить этот эксперимент несколько раз, то интервал будет содержать истинное среднее популяции в 95% случаев. Обычно это доверительный интервал как, например, интервал значений, в пределах которого с доверительной вероятностью 95% находится истинное среднее популяции (генеральное среднее).

Рассмотрим пример.

Пусть дана выборка, состоящая из 64 элементов. Необходимо рассчитать доверительный интервал для данной выборки. Известно, что $M_1 = \bar{X} = 100, \quad \sigma = Sd = 8$.

Решение:

Стандартная ошибка среднего:

$$Se = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{64}} = 1.$$

Доверительный интервал:

$$(\bar{X} - 1{,}96Se;\ \bar{X} + 1{,}96Se);$$

$$(100 - 1{,}96;\ 100 + 1{,}96);$$

$$(98{,}04;\ 101{,}96).$$

2.5.2 Использование t-распределения

Нормальное распределение используется, если известно значение дисперсии в совокупности. Кроме того, когда объем выборки небольшой, выборочное среднее отвечает нормальному распределению, если данные, лежащие в основе популяции, распределены нормально.

Если данные, содержащиеся в совокупности, распределены ненормально и/или неизвестна генеральная дисперсия (дисперсия в совокупности), выборочное среднее подчиняется t-распределению Стьюдента [17].

Вычисляем 95% доверительный интервал для генерального среднего в совокупности следующим образом:

$$(\bar{X} - (t_{0,05} \times Se);\ (\bar{X} + (t_{0,05} \times Se),$$

где $t_{0,05}$ – процентная точка (процентиль);

t – распределения Стьюдента с (n-1) степенями свободы, которая даёт двухстороннюю вероятность 0,05.

Данное распределение обеспечивает более широкий интервал, чем при использовании нормального распределения, поскольку учитывает дополнительную неопределенность, которую вводят, оценивая стандартное отклонение совокупности и/или из-за небольшого объёма выборки.

Когда объём выборки большой (порядка 100 и более), разница между двумя распределениями (t-Стьюдента и нормальным) незначительна. Тем не менее, всегда используют t-распределение при вычислении доверительных интервалов, даже если объем выборки большой.

Обычно указывают 95% доверительный интервал. Можно вычислить другие доверительные интервалы, например 99% доверительный интервал для среднего.

Вместо произведения стандартной ошибки и табличного значения t-распределения, которое соответствует двусторонней вероятности 0,05, умножают стандартную ошибку на значение, которое соответствует двусторонней вероятности 0,01. Это более широкий доверительный интервал, чем в случае 95%, поскольку он отражает увеличенное доверие к тому, что интервал действительно включает среднее совокупности.

2.6 Проверка статистических гипотез

Статистическая гипотеза представляет собой некоторое предположение о законе распределения случайной величины или о параметрах этого закона, формулируемое на основе выборки [18].

Примерами статистических гипотез являются предположения: генеральная совокупность распределена по экспоненциальному закону; математические ожидания двух экспоненциально распределенных выборок равны друг другу. В первой из них высказано предположение о виде закона распределения, а во второй – о параметрах двух распределений. Гипотезы, в основе которых нет никаких допущений о конкретном виде закона распределения, называют непараметрическими, в противном случае – параметрическими.

Гипотезу, утверждающую, что различие между сравниваемыми характеристиками отсутствует, а наблюдаемые отклонения объясняются лишь случайными колебаниями в выборках, на основании которых производится сравнение, называют нулевой (основной) гипотезой и обозначают H_0. Наряду с основной гипотезой рассматривают и альтернативную (конкурирующую, противоречащую) ей гипотезу H_1. И если нулевая гипотеза будет отвергнута, то будет иметь место альтернативная гипотеза.

Различают простые и сложные гипотезы. Гипотезу называют простой, если она однозначно характеризует параметр распределения случайной величины. Например, если l является параметром экспоненциального распределения, то гипотеза H_0 о равенстве $l = 10$ – простая гипотеза. Сложной называют гипотезу, которая состоит из конечного или бесконечного множества простых гипотез. Сложная гипотеза H_0 о неравенстве $l > 10$ состоит из бесконечного множества простых гипотез H_0 о равенстве $l = b_i$, где b_i – любое число, большее 10. Гипотеза H_0 о том, что математическое ожидание нормального распределения равно двум при неизвестной дисперсии, тоже является сложной. Сложной гипотезой будет предположение о распределении случайной величины X по нормальному закону, если не фиксируются конкретные значения математического ожидания и дисперсии.

Проверка гипотезы основывается на вычислении некоторой случайной величины – критерия, точное или приближенное распределение которого известно. Обозначим эту величину через z, ее значение является функцией от элементов выборки $z = z(x_1, x_2, \ldots, x_n)$. Процедура проверки гипотезы предписывает каждому значению критерия одно из двух решений – принять или отвергнуть гипотезу. Тем самым все выборочное пространство и соответственно множество значений критерия делятся на два непересекающихся подмножества S_0 и S_1. Если значение критерия z попадает в область S_0, то гипотеза принимается, а если в область S_1 – гипотеза отклоняется. Множество S_0 называется областью принятия гипотезы или областью допустимых значений, а множество S_1 – областью отклонения

гипотезы или критической областью. Выбор одной области однозначно определяет и другую область.

Принятие или отклонение гипотезы H_0 по случайной выборке соответствует истине с некоторой вероятностью и, соответственно, возможны два рода ошибок.

Ошибка первого рода – возникает с вероятностью a тогда, когда отвергается верная гипотеза H_0 и принимается конкурирующая гипотеза H_1.

Ошибка второго рода – возникает с вероятностью b в том случае, когда принимается неверная гипотеза H_0, в то время как справедлива конкурирующая гипотеза H_1.

3. РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ – это статистический метод исследования зависимости случайной величины y от переменных (аргументов) x_j (j = 1, 2,..., k), рассматриваемых в регрессионном анализе как неслучайные величины независимо от истинного закона распределения x_j.(1) [19].

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Последовательность этапов регрессионного анализа:

- формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений;
- определение зависимых и независимых (объясняющих) переменных;
- сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель;
- формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная);
- определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии);
- оценка точности регрессионного анализа;
- интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов;
- предсказание неизвестных значений зависимой переменной.

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вы-

числяются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та, где оно меньше нуля, к другому классу.

Задачи регрессионного анализа:

1) установление формы зависимости (линейная, нелинейная, положительная, отрицательная):

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

- положительная линейная регрессия (выражается в равномерном росте функции);
- положительная равноускоренно возрастающая регрессия;
- положительная равнозамедленно возрастающая регрессия;
- отрицательная линейная регрессия (выражается в равномерном падении функции);
- отрицательная равноускоренно убывающая регрессия;
- отрицательная равнозамедленно убывающая регрессия [20].

Однако описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии;

2) определение функции регрессии:

Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов. Функция регрессии определяется в виде математического уравнения того или иного типа;

3) определение влияния на функцию регрессии отдельных факторов;

4) решение задач экстраполяции и интерполяции (определение значений функций в неисследованных участках, например, при решении задач прогнозирования).

Решение этой задачи сводится к решению задачи одного из типов:

- оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений, при этом решается задача интерполяции;
- оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

Рассмотрим некоторые предположения, на которые опирается регрессионный анализ:

1. *Предположение линейности*, т.е. предполагается, что связь между рассматриваемыми переменными является линейной. Так, в рассматриваемом примере мы построили диаграмму рассеивания и смогли увидеть явную линейную связь. Если же на диаграмме рассеивания переменных мы видим явное отсутствие линейной связи, т.е. присутствует нелинейная связь, следует использовать нелинейные методы анализа.

2. *Предположение о нормальности остатков*. Оно допускает, что распределение разницы предсказанных и наблюдаемых значений является нормальным. Для визуального определения характера распределения можно воспользоваться гистограммами остатков.

При использовании регрессионного анализа следует учитывать его основное ограничение**.** Оно состоит в том, что регрессионный анализ позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей.

Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений.

Остаток – это отклонение отдельной точки (наблюдения) от линии регрессии (предсказанного значения).

Величина R-квадрат, называемая также мерой определенности, характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала [0;1].

В большинстве случаев значение R-квадрат находится между этими значениями, называемыми экстремальными, т.е. между нулем и единицей.

Если значение R-квадрата близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение R-квадрата, близкое к нулю, означает плохое качество построенной модели.

Множественный R – коэффициент множественной корреляции R – выражает степень зависимости независимых переменных (X) и зависимой переменной (Y). Множественный R равен квадратному корню из коэффициента детерминации, эта величина принимает значения в интервале от нуля до единицы. В простом линейном регрессионном анализе множественный R равен коэффициенту корреляции Пирсона [21].

Виды регрессий:

1. Регрессия относительно числа переменных:
   - простая регрессия – регрессия между двумя переменными;
   - множественная регрессия – регрессия между зависимой переменной Y и несколькими объясняющими переменными X.
2. Регрессия относительно формы зависимости:
   - линейная регрессия, выражаемая линейной функцией;
   - нелинейная регрессия, выражаемая нелинейной функцией.
3. В зависимости от характера регрессии различаются следующие ее виды:
   - положительная регрессия: она имеет место, если с увеличением (уменьшением) объясняющей переменной значения зависимой переменной также соответственно увеличиваются (уменьшаются);

- отрицательная регрессия: в этом случае с увеличением или уменьшением объясняющей переменной зависимая переменная уменьшается или увеличивается.

4. Относительно типа соединения явлений различаются:

- непосредственная регрессия: в этом случае зависимая и объясняющая переменные связаны непосредственно друг с другом;
- косвенная регрессия: в этом случае объясняющая переменная действует на зависимую через ряд других переменных;
- ложная регрессия: она возникает при формальном подходе к исследуемым явлениям без уяснения того, какие причины обусловливают данную связь.

Рассмотрим четыре вида регрессионного анализа: линейную регрессию и нелинейную регрессию, одномерную и множественную (многомерную) регрессию [22].

3.1 Линейная простая регрессия

Линейная регрессия предназначена для получения прогноза непрерывных числовых переменных [23].

Достоинства линейной регрессии:

1. Скорость и простота получения модели.

2. Интерпретируемость модели. Линейная модель является прозрачной и понятной для аналитика. По полученным коэффициентам регрессии можно судить о том, как тот или иной фактор влияет на результат, сделать на этой основе дополнительные полезные выводы.

3. Широкая применимость. Большое количество реальных процессов в экономике и бизнесе можно с достаточной точностью описать линейными моделями.

4. Изученность данного подхода. Для линейной регрессии известны типичные проблемы (например, мультиколлинеарность) и их решения, разработаны и реализованы тесты оценки статической значимости получаемых моделей.

Математическое уравнение, которое оценивает линию простой (парной) линейной регрессии:

$$Y = a + b * X, \qquad (3.1)$$

где $X -$ называется независимой переменной или предиктором;

Y – зависимая переменная или переменная отклика. Это значение, которое ожидается для Y (в среднем), если мы знаем величину X, т.е. это «предсказанное» значение Y;

a – свободный член (пересечение) линии оценки; это значение Y, когда X =0 (рис. 12);

b – угловой коэффициент или градиент оценённой линии; она представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем x на одну единицу;

a и b называют коэффициентами регрессии оценённой линии, хотя этот термин часто используют только для b.

В большинстве случаев (если не всегда) наблюдается определенный разброс наблюдений относительно регрессионной прямой.

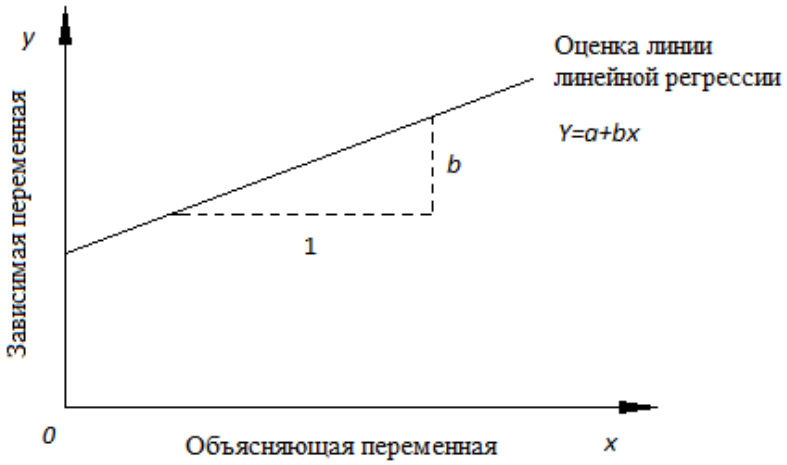

Рис. 12. Линия линейной регрессии, показывающая пересечение a и углового коэффициента b

3.1.1 Метод наименьших квадратов

При выполнении регрессионного анализа необходимо использовать выборку наблюдений, где a и b – выборочные оценки истинных (генеральных) параметров α и β, которые определяют линию линейной регрессии в популяции (генеральной совокупности).

Наиболее простым методом определения коэффициентов a и b является метод наименьших квадратов (МНК).

Подгонка оценивается, рассматривая остатки (вертикальное расстояние каждой точки от линии, например, остаток = наблюдаемому y – предсказанный y, (рис. 13).

Линию лучшей подгонки выбирают так, чтобы сумма квадратов остатков была минимальной.

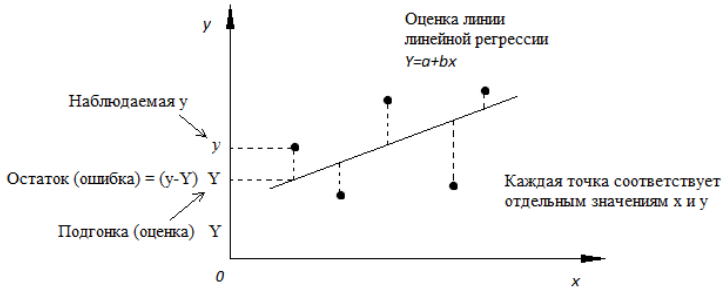

Рис. 13. Линия линейной регрессии с изображенными остатками

Для каждой наблюдаемой величины x остаток равен разнице y и соответствующего предсказанного Y. Каждый остаток может быть положительным или отрицательным.

Можно использовать остатки для проверки следующих предположений, лежащих в основе линейной регрессии:

1. Между x и y существует линейное соотношение: для любых пар $(x; y)$ данные должны аппроксимировать прямую линию. Если нанести на двумерный график остатки, то наблюдается случайное рассеяние точек, а не какая-либо систематическая картина.

2. Остатки нормально распределены с нулевым средним значением.

3. Остатки имеют одну и ту же вариабельность (постоянную дисперсию) для всех предсказанных величин *y*. Если нанести остатки против предсказанных величин *Y* от *y* должно наблюдаться случайное рассеяние точек. Если график рассеяния остатков увеличивается или уменьшается с увеличением *Y*, то это допущение не выполняется.

Если допущения линейности, нормальности и/или постоянной дисперсии сомнительны, можно преобразовать *x* или *y* и рассчитать новую линию регрессии, для которой эти допущения удовлетворяются (например, использовать логарифмическое преобразование или др.).

Аномальные значения (выбросы) и точки влияния. «Влиятельное» наблюдение, если оно опущено, изменяет одну или больше оценок параметров модели (т.е. угловой коэффициент или свободный член). Выброс (наблюдение, которое противоречит большинству значений в наборе данных) может быть «влиятельным» наблюдением и может хорошо обнаруживаться визуально, при осмотре двумерной диаграммы рассеяния или графика остатков.

И для выбросов, и для «влиятельных» наблюдений (точек) используют модели, как с их включением, так и без них, обращают внимание на изменение оценки (коэффициентов регрессии). При проведении анализа не стоит отбрасывать выбросы или точки влияния автоматически, поскольку простое игнорирование может повлиять на полученные результаты. Всегда изучайте причины появления этих выбросов и анализируйте их.

Гипотеза линейной регрессии. При построении линейной регрессии проверяется нулевая гипотеза о том, что генеральный угловой коэффициент линии регрессии β равен нулю.

Если угловой коэффициент линии равен нулю, между *x* и *y* нет линейного соотношения: изменение *x* не влияет на *y*.

Для тестирования нулевой гипотезы о том, что истинный угловой коэффициент β равен нулю, можно воспользоваться следующим алгоритмом.

Вычислить статистику критерия, равную отношению $\frac{b}{SE(b)}$, которая подчиняется t-распределению с (n-2) степенями свободы, где

SE(b) – стандартная ошибка коэффициента (3.3), оценка дисперсии остатков (3.4)

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}; \quad (3.2)$$

$$SE(b) = \frac{S_{res}}{\sum(x-\overline{x})^2}; \quad (3,3)$$

$$S_{res}^2 = \frac{\sum(y-Y)^2}{(n-2)}. \quad (3.4)$$

Обычно если достигнутый уровень значимости P < 0.05, нулевая гипотеза отклоняется.

Можно рассчитать 95% доверительный интервал для генерального углового коэффициента β:

$$b \pm t_{0.05} SE(b),$$

где $t_{0.05}$ – процентная точка t-распределения со степенями свободы (n-2), что дает вероятность двустороннего критерия 0.05.

Это тот интервал, который содержит генеральный угловой коэффициент с вероятностью 95%.

Для больших выборок, скажем n ≥100, мы можем аппроксимировать значением 1,96 (т.е. статистика критерия будет стремиться к нормальному распределению) [24].

3.1.2 Метод градиентного спуска

Метод наименьших квадратов используется только в тех случаях, когда количество наблюдений не велико. В противном случае используются эвристические методы. Примером этих методов может служить метод градиентного спуска, который рассмотрен ниже.

Пусть RSS – средняя квадратичная ошибка:

$$RSS = \sum e^2 \rightarrow min. \quad (3.5)$$

Основная идея этого метода состоит в том, чтобы двигаться к минимуму в направлении наиболее быстрого убывания функции,

которое определяется антиградиентом. Эта идея реализуется следующим образом.

Выберем каким-либо способом начальную точку, вычислим в ней градиент рассматриваемой функции и сделаем небольшой шаг в обратном, антиградиентном направлении. В результате мы придем в точку, в которой значение функции будет меньше первоначального. В новой точке повторим процедуру: снова вычислим градиент функции и сделаем шаг в обратном направлении. Продолжая этот процесс, мы будем двигаться в сторону убывания функции. Специальный выбор направления движения на каждом шаге позволяет надеяться на то, что в данном случае приближение к наименьшему значению функции будет более быстрым, чем в методе покоординатного спуска [25].

Метод градиентного спуска требует вычисления градиента ∇f целевой функции на каждом шаге:

$$\nabla f = \frac{\partial f}{\partial \beta} * e_1 + \frac{\partial f}{\partial \beta} * e_2 + \cdots + \frac{\partial f}{\partial \beta} * e_n. \quad (3.6)$$

На каждом шаге рассчитывается коэффициент β (3.7):

$$\beta_i = \beta_{i-1} - \alpha * \nabla f, \quad (3.7)$$

где $\alpha -$ скорость обучению алгоритма.

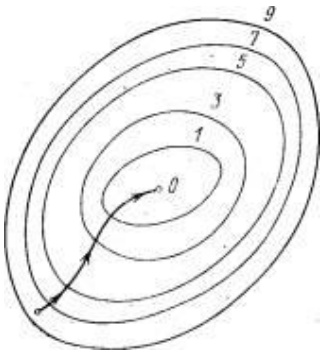

Рис. 14. Поиск наименьшего значения функций методом градиентного спуска

3.2 Одномерная линейная регрессия

Одномерная (простая) линейная регрессия – линейная регрессия с одной независимой скалярной переменной (объясняющей переменной). Под одномерной линейной регрессией также понимают и сопряженный модели метод наименьших квадратов, оценивающий параметры регрессии. Данную модель называют простой, так как это одна из самых простых моделей регрессии. На двумерной плоскости функция регрессии является прямой. Модель характеризуется двумя параметрами: угловым коэффициентом и свободным членом прямой [26].

Дана выборка $x_m = \{(x_1, y_1), \dots, (x_m, y_m), x_i, y_i \in R\}$.

Модель описывается уравнением:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \tag{3.8}$$

где y_i – зависимая переменная (отклик);

x_i – известная константа (значение объясняющей переменной, измеренной в i-том эксперименте);

β_0, β_1 — параметры модели (свободный член и угловой коэффициент);

ε_i – случайная ошибка со средним $M(\varepsilon_i) = 0$ и дисперсией σ $\sigma^2(\varepsilon_i) = \sigma^2$, i=1…m.

Согласно методу наименьших квадратов, искомый вектор параметров $\beta = (\beta_0, \beta_1)^T$ есть решение нормального уравнения (3.9)

$$\beta = (A^T A)^{-1} A^T y_i, \tag{3.9}$$

где y – вектор, состоящий из значений зависимой переменной, $y = (y_1, \dots, y_m)$.

Столбцы матрицы A есть подстановки значений свободной переменной $x_i^0 \to a_{i1}$ и $x_i^1 \to a_{i2}$, i=1…m. Матрица имеет вид (3.10):

$$A = \begin{pmatrix} 1 & x_1 \\ & x_2 \\ \dots & \dots \\ 1 & x_m \end{pmatrix}. \tag{3.10}$$

Зависимая переменная восстанавливается по полученным весам и заданным значениям свободной переменной (3.11):

$$y_i^* = \beta_0 + \beta_1 x_i. \quad (3.11)$$

Иначе

$$y^* = A\beta. \quad (3.12)$$

Для оценки качества модели используется критерий суммы квадратов регрессионных остатков, SSE – Sum of Squared Errors (3.13).

$$SSE = \sum_{i=1}^{m} (y_{i-} y_i^*)^2 = (y - y^*)^T (y - y^*). \quad (3.13)$$

3.3 Нелинейная регрессия

Нелинейная регрессия – частный случай регрессионного анализа, в котором рассматриваемая регрессионная модель есть функция, зависящая от параметров и от одной или нескольких свободных переменных. Зависимость от параметров предполагается нелинейной (например, степенная).

Различают два класса нелинейных регрессий:

- регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам;
- регрессии, нелинейные по оцениваемым параметрам.

Примером нелинейной регрессии по включаемым в неё объясняющим переменным могут служить следующие функции:

- полиномы разных степеней – $y = a + b * x + c * x^2 + \varepsilon$, $y = a + b * x + c * x^2 + d * x^3 + \varepsilon$;
- равносторонняя гипербола – $y = a + \frac{b}{x} + \varepsilon$.

К нелинейным регрессиям по оцениваемым параметрам относятся функции:

- степенная – $y = a * x^b * \varepsilon$;
- показательная – $y = a * b^x * \varepsilon$;
- экспоненциальная – $y = e^{a+bx} * \varepsilon$.

Параметры нелинейной регрессии по включенным переменным оцениваются, как и в линейной регрессии, методом наименьших квадратов, поскольку эти функции линейны по параметрам [27].

3.4 Множественная (многомерная) регрессия

Многомерная линейная регрессия – это линейная регрессия в n-мерном пространстве (объекты и признаки являются n-мерными векторами).

Модель множественной регрессии представляет собой уравнение, отражающее корреляционную связь между объясняемой переменной и несколькими объясняющими переменными. В общем виде уравнение множественной регрессии может быть записано, как:

$$y = f(x_1, x_2, \dots, x_n, \varepsilon),$$

где y – объясняемая переменная;

$x_1, x_2, \dots, x_n$ – объясняющие переменные (факторы модели);

ε – остатки модели;

f – некоторая функция.

В качестве функций множественной регрессии чаще всего выбирают наиболее простые – линейную, степенную и показательную функции. Данные функции также могут быть использованы и при формировании смешанных моделей, включающих в себя несколько видов зависимостей. При проведении регрессионного и корреляционного анализа предполагается, что наблюдения были получены по относительно однородной совокупности единиц.

Рассмотрим линейную множественную регрессию в матричном виде:

1. $$y = Xa + \varepsilon, \tag{3.14}$$

где $$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}; \tag{3.15}$$

$$\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}; \quad (3.16)$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}. \quad (3.17)$$

2. $y = Xb + e$ – собственно модель, (3.18)

где b – оценка параметра a.

Существуют различные методы определения величины b. Наиболее часто используется метод наименьших квадратов [28].

4. ВРЕМЕННОЙ АНАЛИЗ РИСКОВ БЕЗОПАСНОСТИ ИНФОРМАЦИОННЫХ СИСТЕМ

Прогноз времени отказа или появления неисправности в узлах системы является критически важным для обеспечения надежности и безопасности информационной системы в целом. Наиболее распространенным статистическим методом анализа времени отказов является анализ выживаемости, который изучает положительные случайные величины с цензурными наблюдениями для описания случаев до некоторых событий. Наибольшее распространение модели выживаемости получили в медицине, где событие может быть связано с развитием заболевания, реакции на лечение или смерть. Таким образом, изучение данных о выживании будет сосредоточено на прогнозировании вероятности реакции, выживания или среднего времени жизни, сравнивая распределения выживания экспериментальных животных или людей и выявление риска или прогностических факторов, связанных с реакцией, выживанием и развитием болезни.

Методы анализа выживаемости также подходят для применения в других областях, таких как технологии обеспечения надежности, социальные науки и бизнес. Примером анализа выживаемости является время жизни электронных устройств, а также системы и требования компенсации работникам (страхование) и их различные факторы, влияющие на риск. Существуют примеры применения анализа выживаемости для оценки надежности компьютерных сетей. Рассмотрим базовые понятия теории анализа выживаемости, а также примеры использования в области безопасности информационных систем.

Пусть T – неотрицательная случайная величина, представляющая собой время ожидания до наступления некоторого события.

Для простоты будем использовать терминологию анализа выживаемости, называя исследуемое событие «смертью», а время ожидания – временем «выживания», хотя изучаемые далее методы находят гораздо более широкое применение. Их можно использовать, например, для анализа времени работы системы до отказа, интервалов между сбоями системы или продолжительности жизни информационной системы.

4.1 Функция выживания

Предположим, что T – непрерывная случайная величина с функцией плотности распределения (ФПР) $f(t)$ и кумулятивной функцией распределения (КФР) $F(t) = P\{T \leq t\}$, дающей вероятность того, что событие наступило к моменту времени t.

Часто удобно работать с дополнением КФР, называемым функцией выживания:

$$S(t) = \mathbb{P}\{T > t\} = 1 - F(t) = \int_t^\infty f(x)dx \tag{4.1}$$

и дающим вероятность быть живым в момент времени t, или в более широком смысле, вероятность того, что исследуемое событие не наступило к моменту времени t.

4.2 Функция риска

Альтернативным способом охарактеризовать распределение величины T является функция риска, или мгновенная интенсивность осуществления события, определяемая как

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}\{t < T \leq t + dt | T > t\}}{dt} \tag{4.2}$$

Числитель этого выражения – условная вероятность того, что событие произойдет в интервале $(t, t + dt)$, если оно не произошло

ранее, а знаменатель – ширина интервала. Разделив одно на другое, получаем интенсивность осуществления события в единицу времени. Устремляя ширину интервала к нулю и переходя к пределу, получаем мгновенную интенсивность осуществления события.

Условную вероятность в числителе можно записать в виде отношения совместной вероятности того, что T принадлежит интервалу $(t, t + dt)$ и $T > t$ (что, конечно, совпадает с вероятностью того, что T принадлежит указанному интервалу), к вероятности условия $T > t$. Первая из них равна $f(t)dt$ для малого dt, а последняя – это $S(t)$, по определению.

Деление на dt и предельный переход дают следующий полезный результат:

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{4.3}$$

который некоторые авторы приводят в качестве определения функции риска. Содержательно, интенсивность осуществления события в момент времени t равна плотности событий в момент t, деленной на вероятность дожить до этого момента, не испытав событие ранее.

Заметим из уравнения (4.1), что $-f(t)$ – это производная $S(t)$. Тогда уравнение (4.3) можно переписать в виде

$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

Если теперь проинтегрировать обе части от 0 до t и ввести граничное условие $S(0) = 1$ (поскольку событие не может произойти к моменту времени 0), можно преобразовать приведенное выражение и получить формулу для вероятности дожить до момента времени t как функции от рисков во все моменты времени до t:

$$S(t) = \exp\left\{-\int_0^t \lambda(x)dx\right\} \tag{4.4}$$

Это выражение должно быть знакомо демографам. Интеграл в фигурных скобках в этом уравнении называют кумулятивным риском и обозначают как

$$\Lambda(t) = \int_0^t \lambda(x)dx \tag{4.5}$$

Можно рассматривать $\Lambda(t)$ как сумму всех рисков при переходе от момента времени 0 к *t*.

Приведенные результаты показывают, что функции выживания и риска дают альтернативные, но эквивалентные описания распределения величины *T*. Имея функцию выживания, всегда можно ее продифференцировать и получить функцию плотности, а затем найти функцию риска, используя уравнение (4.3). Имея функцию риска, всегда можно ее проинтегрировать и получить кумулятивный риск, а затем взять от нее экспоненту и найти функцию выживания, используя уравнение (4.4). Для закрепления введенных понятий рассмотрим пример.

Пример: Простейшее распределение времени жизни получается, если предположить постоянный риск, то есть

$\lambda(t) = \lambda$

для всех t. Соответствующая функция выживания имеет вид

$S(t) = \exp\{-\lambda t\}$.

Это экспоненциальное распределение с параметром λ. Функцию плотности можно получить, умножив функцию выживания на риск:

$f(t) = \lambda \exp\{-\lambda t\}$.

Математическое ожидание равно $1/\lambda$. Это распределение играет центральную роль в анализе выживаемости, хотя, возможно, оно является слишком простым, чтобы быть полезным в приложениях само по себе.

4.3 Ожидаемая продолжительность жизни

Пусть μ обозначает математическое ожидание *T*. По определению, значение μ можно подсчитать, умножив *t* на функцию плотности $f(t)$ и взяв интеграл, то есть

$$\mu = \int_0^\infty t f(t)dt.$$

Интегрируя по частям и используя тот факт, что $-f(t)$ – это производная S(t), удовлетворяющая граничным условиям $S(0) = 1$ и $S(\infty) = 0$, можно показать, что

$$\mu = \int_0^\infty S(t)dt \tag{4.6}$$

Иными словами, ожидаемая продолжительность жизни – это просто интеграл от функции выживания.

4.4 Цензурирование и функция правдоподобия

Вторая отличительная черта анализа выживаемости – цензурирование, то есть тот факт, что для некоторых объектов наблюдения исследуемое событие произошло, а значит, известно точное время ожидания, тогда как для других это событие не произошло, и все, что известно, – это то, что время ожидания превышает время наблюдения.

Существует несколько механизмов, способных генерировать цензурированные данные. При цензурировании типа I выборка из n объектов наблюдается в течении фиксированного времени τ. Число объектов, испытывающих событие, или число «смертей», случайно, но общая продолжительность исследования фиксирована. Тот факт, что продолжительность фиксирована, может быть важным практическим преимуществом при разработке последующего дополнительного исследования.

При простом обобщении этой схемы, называемом фиксированным цензурированием, каждый объект имеет максимально возможный период наблюдения τ_i, $i = 1, \ldots, n$, который может варьироваться от одного объекта к другому, однако фиксирован заранее. Вероятность того, что объект i будет жив в конце своего периода наблюдения, равна $S(\tau_i)$, а общее число смертей вновь является случайным.

При цензурировании типа II выборка из n объектов наблюдается так долго, сколько необходимо, чтобы d объектов испытали со-

бытие. В этой схеме число смертей d, которое определяет точность исследования, фиксировано заранее и его можно использовать в качестве параметра. К сожалению, в этом случае общая продолжительность исследования случайна и не может быть точно известна заранее.

При более общей схеме, называемой случайным цензурированием, каждый объект имеет потенциальный момент цензурирования C_i и потенциальную продолжительность жизни T_i, которые предполагаются независимыми случайными величинами. Наблюдается $Y_i = \min\{C_i , T_i\}$, то есть минимум из времени цензурирования и времени жизни, и переменная-индикатор, часто обозначаемая d_i или δ_i , которая указывает, закончено ли наблюдение в результате смерти или цензурирования.

Все эти схемы объединяет тот факт, что механизм цензурирования неинформативен, и все они, в сущности, ведут к той же самой функции правдоподобия. Наиболее слабое предположение, требуемое для получения этой функции правдоподобия, состоит в том, что цензурирование наблюдения не должно давать какой-либо информации относительно перспектив выживания этого конкретного объекта за пределами момента цензурирования. На самом деле базовое предположение, которому мы будем следовать, таково: все, что известно о наблюдении, цензурированном в момент времени t – это то, что время жизни для него превышает t.

4.5 Модели ускоренной жизни

До сих пор речь шла об однородной популяции, в которой продолжительность жизни каждого объекта характеризовалась одной и той же функцией выживания $S(t)$. Рассмотрим теперь третью отличительную черту моделей выживаемости – наличие вектора регрессоров, или объясняющих переменных, которые могут воздействовать на время жизни, – и обратимся к общей задаче моделирования этих эффектов.

Пусть T_i – случайная величина, представляющая собой (возможно, ненаблюдаемое) время жизни i-го объекта. Поскольку величина T_i должна быть неотрицательной, можно рассмотреть модель для ее логарифма, скажем, обычную линейную модель

$$\log T_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

где ε_i – надлежащий остаточный член, распределение которого будет специфицировано далее. Эта модель задает распределение логарифма времени жизни для i-го объекта как простой сдвиг стандартного, или базового, распределения, представленного остаточным членом.

Взяв экспоненту от этого уравнения, получаем модель собственно для времени жизни:

$$T_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}T_{0i},$$

где T_{0i} – экспонента от остаточного члена. Удобно также использовать обозначение γ_i для мультипликативного эффекта регрессоров, $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$.

Интерпретация параметров стандартна. Рассмотрим, например, модель с константой и фиктивной переменной x, отражающей бинарный фактор, скажем, принадлежность к группам 1 или 0. Предположим, соответствующий мультипликативный эффект $\gamma = 2$, так что коэффициент при x – это $\beta = \log(2) = 0{,}6931$. Тогда вывод состоит в том, что люди из первой группы живут вдвое дольше, чем из нулевой.

Существует интересная альтернативная интерпретация, которая объясняет название «модель ускоренной жизни». Пусть $S_0(t)$ обозначает функцию выживания для группы 0, которая будет контрольной группой, а $S_1(t)$ – для группы 1. Для этой модели

$$S_1(t) = S_0(t/\gamma).$$

Иными словами, вероятность того, что индивид из первой группы доживет до возраста t, в точности равна вероятности того, что индивид из нулевой группы доживет до возраста t/γ.

Для $\gamma = 2$ получим половину возраста, так что вероятность того, что индивид из первой группы доживет до 40 (или 60) лет будет равна вероятности того, что индивид из нулевой группы доживет до 20 (или 30) лет. Таким образом, можно рассматривать γ как параметр, воздействующий на протекание времени. В нашем примере люди в нулевой группе стареют «в два раза быстрее».

Заметим, что соответствующие функции риска связаны соотношением

$$\lambda_1(t) = \lambda_0(t/\gamma)/\gamma,$$

так что при $\gamma = 2$ в каждом данном возрасте люди из первой группы будут подвержены вдвое меньшему риску, чем вдвое младшие люди из нулевой группы.

Название «модель ускоренной жизни» происходит из промышленных приложений, когда предметы тестируются при гораздо худших условиях, чем встречающиеся в реальной жизни, чтобы тесты можно было выполнить за более короткое время.

Различные предположения о распределении остаточного члена приводят к различным видам параметрических моделей. Если ошибка i нормально распределена, получается логнормальная модель для T_i . Оценивание этой модели для цензурированных данных по методу максимального правдоподобия известно в эконометрической литературе как тобит-модель.

Если же ε_i имеет распределение экстремального значения с функцией плотности

$$f(\epsilon) = \exp\{\epsilon - \exp(\epsilon)\},$$

то T_{0i} имеет экспоненциальное распределение, и получается модель экспоненциальной регрессии, где T_i экспоненциально распределено с риском λ_i, удовлетворяющим логлинейной модели

$$\log \lambda_i = \mathbf{x}_i' \boldsymbol{\beta}.$$

Примером демографической модели, принадлежащей классу моделей ускоренной жизни, является модель Кола–Макнейла для

частоты первого брака, в которой доля индивидов, когда-либо состоявших в браке к возрасту a в данной популяции, записывается в виде

$$F(a) = cF_0\left(\frac{a - a_0}{k}\right),$$

где F_0 – модельное распределение долей женщин, состоявших в браке к определенному возрасту среди когда-либо состоявших в браке, на основе исторических данных по Швеции;

c – доля тех, кто в конечном счете вступают в брак, a_0 – возраст вступления в брак, а k – скорость протекания брака относительно шведского стандарта.

Модели ускоренной жизни по сути являются обычными моделями регрессии, примененными к логарифму времени жизни, и, не считая факта цензурирования данных, не представляют новых трудностей при оценивании. Как только выбрано распределение остаточного члена, оценивание осуществляется путем максимизации логарифмической функции правдоподобия для цензурированных данных.

4.6 Модели пропорциональных рисков

Большой класс моделей, впервые предложенный в Cox (1972), концентрируется непосредственно на функции риска. Простейший представитель этого класса – модель пропорциональных рисков, в которой риск в момент t для индивида с характеристиками x_i (не включая константу) имеет вид

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{7}$$

В этой модели $\lambda_0(t)$ – это базовая функция риска, которая измеряет риск для индивидов с $x_i = 0$, служащих точкой отсчета, а $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ – относительный риск, то есть пропорциональное увеличение или уменьшение риска, связанное с набором характеристик x_i. Заметим, что увеличение или снижение риска одинаково для всех моментов времени t.

В качестве иллюстрации рассмотрим пример с двумя выборками, когда имеется фиктивная переменная *x*, означающая принадлежность к первой или нулевой группе. В этом случае модель принимает вид

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{если } x = 0 \\ \lambda_0(t)e^{\beta} & \text{если } x = 1. \end{cases}$$

Таким образом, $\lambda_0(t)$ представляет собой риск в момент t в нулевой группе, а $\gamma = \exp\{\beta\}$ – это отношение риска в первой группе к риску в нулевой группе в любой момент времени t.

Если $\gamma = 1$ (или $\beta = 0$), риски одинаковы в обеих группах. Если $\gamma = 2$ (или $\beta = 0{,}6931$), риск для индивида из первой группы в каждый момент времени вдвое больше риска для индивида того же возраста из нулевой группы.

Заметим, что модель явным образом отделяет эффект времени от эффекта регрессоров. Логарифмируя, легко увидеть, что модель пропорциональных рисков – это простая аддитивная модель для логарифма риска:

$$\log \lambda_i(t|\mathbf{x}_i) = \alpha_0(t) + \mathbf{x}_i'\boldsymbol{\beta},$$

где $\alpha_0(t) = \log \lambda_0(t)$ – логарифм базового риска. Как во всех аддитивных моделях, предполагается что влияние регрессоров x одинаково для всех моментов времени, или возрастов, t.

Нельзя не отметить схожесть между этим выражением и стандартной моделью ковариационного анализа с параллельными прямыми.

Возвращаясь к уравнению (4.7), можно проинтегрировать обе его части от 0 до t и получить кумулятивные риски

$$\Lambda_i(t|\mathbf{x}_i) = \Lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

которые также пропорциональны. Взяв экспоненту от этого уравнения с противоположным знаком, получаем функции выживания

$$S_i(t|\mathbf{x}_i) = S_0(t)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}, \qquad (4.8)$$

где $S_0(t) = \exp\{-\Lambda_0(t)\}$ – базовая функция выживания. Таким образом, эффект регрессоров x_i на функцию выживания заключается в возведении ее в степень, равную относительному риску $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$.

В нашем примере с двумя группами и относительным риском $\gamma = 2$ вероятность того, что индивид из первой группы доживет до возраста t, равен квадрату вероятности того, что индивид из нулевой группы доживет до этого же возраста.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Лимончелли, Т. Системное и сетевое администрирование. Практическое руководство / Т. Лимончелли, К. Хоган, С. Чейлап ; пер. с англ. – 2-е издание. – Санкт-Петербург : Символ-Плюс, 2009. – 944 с.
2. Мюллер, А. Введение в машинное обучение с помощью Python: рук. для спец. / А. Мюллер, С. Гвидо. – Москва : Вильямс, 2016-2017. – 480 с.
3. Королев, В.Ю. Теория вероятностей и математическая статистика / В.Ю. Королев. – Москва :ТК Велби, Издво Проспект, 2008. – 160 с.
4. Положинцев, Б.И. Теория вероятностей и математическая статистика / Б.И. Положинцев. – Санкт-Петербург: СПбГПУ, 2010 – С. 90-93.
5. Кремер, Н.Ш. Теория вероятности и математическая статистика : учебник для вузов. / Н.Ш. Кремер. – 2-е издание. – Москва : ЮНИТИ. – ДАНА, 2006. – 573 с.
6. Овсянникова, С.Н. Статистика для студентов 2-го курса экономических специальностей : учебное пособие / С.Н. Овсянникова. – Москва : Экон-информ, 2011.– 126 с.
7. Белов, А.А. Теории вероятностей и математическая статистика : учебник / А.А. Белов, Б.А. Баллод, Н.Н. Елизарова. – Ростов-на-Дону : Феникс, 2009. – С. 53-54.
8. Долгов, Ю.А. Метод повышения точности вычисления параметров выборки малого объёма (метод точечных распределений) / Ю.А. Долгов, А.Ю. Долгов, Ю.А. Столяренко // Вестник Приднестровского университета. – 2010. – № 1(36) – С. 232-242.
9. Егорова, Л.Г. Базы данных. Операторы выборки данных : учебное пособие / Л.Г. Егорова, Ю.Б. Кухта. – Москва : МГТУ им. Г.И. Носова, 2017.
10. Горский, В.Г. Прикладная математическая статистика / В.Г. Горский // Заводская лаборатория. Диагностика материалов. – 2007.– Т 73, № 1.– С. 96-100.

11. Болдырев, И.В. «Ящик с усами» для анализа данных контроля точности / И.В. Болдырев // Контроль качества продукции. – 2014. – № 2 – С. 39-41.
12. Гергель, Н.И. Понятие о норме. Закон распределения нормальных величин Гаусса : учебное пособие для студентов медицинских вузов / Н.И. Гергель, И.А. Селезнева, Е.Е. Воронкова. – Самара : СамГМУ, 2011. – С.34-35.
13. Акимов, В.С. Классификация законов распределения и алгоритмизация процесса определения закона распределения вероятности / В.С. Акимов // Научное обозрение.– 2014. – № 1. – С. 86-90.
14. Орлов, А.И. Нечисловая статистика / А.И. Орлов. – Москва : МЗ-Пресс, 2004. – 513 с.
15. Чернова, Н.И. Теория вероятностей : учебное пособие / СибГУТИ. – Новосибирск, 2009. – 128 с.
16. Гусев, Л.А. О некоторых свойствах доверительных интервалов для неизвестных вероятностей / Л.А. Гусев // Автомат. и телемех. – 2007. – С. 70-84.
17. Леман, Э. Проверка статистических гипотез / Э. Леман. – 1964. – 500 с.
18. Арташесян, А.А. Алгоритм машинного обучения на основе анализа малых выборок / А.А. Арташесян, В.В. Дерюшев // Строительство и архитектура.– 2017 – С. 82-86.
19. Норман, Р.Д. Прикладной регрессионный анализ : Том 2 / Дрейпер Р. Норман, Гарри Смит. – Москва : Книга по Требованию, 2013. – 350 с.
20. Кривенко, М.П. Сравнительный анализ процедур регрессионного анализа / М.П. Кривенко // Информ. и ее примен. – 2014. – Т. 8, № 3. – С. 70-78.
21. Попова, О.А. Численный вероятностный анализ для агрегации, регрессионного моделирования и анализа данных / О.А. Попова // Информ. и связь. – 2015. – № 1 – С. 15-21.
22. Давнис, В.В. Предельный анализ регрессионных моделей дискретного выбора / В.В. Давнис, В.И. Тинякова // Эконом. анализ: теория и практика. – 2006. – № 10(67). – С. 4-13.
23. Яковлев, В.Б. Регрессионный анализ. Расчеты в Excel и Statistica / В.Б. Яковлев. – Москва : РУСАЙНС, 2018. – 178 с.

24. Сысоев, В.В. Парная линейная регрессия : учебное пособие / В.В. Сысоев. – Воронеж : Воронеж. Гос. Технол. Акад., 2003. – 66 с.
25. Гасников, А.В. Современные численные методы оптимизации. Метод универсального градиентного спуска : учебное пособие для студентов / А.В. Гасников. – Москва : МФТИ, 2018. – 166 с.
26. Филатова Л.Ф. Парная регрессия : учебное пособие / Л.Ф. Филатова. – Северск : Изд. СТИ НИЯУ МИФИ, 2013. – 75 с.
27. Губарев, В.В. Случайные функции с нелинейной регрессией и их применение / В.В. Губарев // Автометрия. – 2011. – Т. 47, № 6. – С. 39-50.
28. Сажин, Ю.В. Эконометрика : учебник / Ю.В. Сажин, И.А. Иванова. – Саранск : МГУ им. Н.П. Огарева, 2014. – 316 с.
29. Родригес, Г. Модели выживаемости / Г. Родригес // Квантиль. – № 5. – 2008. – С. 1-27.

Учебное издание

Сапрыкин Олег Николаевич

СТАТИСТИЧЕСКИЙ АНАЛИЗ РИСКОВ В СИСТЕМАХ КОМПЛЕКСНОЙ БЕЗОПАСНОСТИ

Учебное пособие

Редактор Л.Р. Дмитриенко
Компьютерная верстка Л.Р. Дмитриенко

Подписано в печать 21.12.2020. Формат 60x84 1/16.
Бумага офсетная. Печ. л. 4,5.
Тираж 120 экз. (1-й з-д 1-25). Заказ . Арт. – 5(РЗУ)/2020.

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)
443086, Самара, Московское шоссе, 34.

Издательство Самарского университета.
443086, Самара, Московское шоссе, 34.