Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Н. Б. ОСИПЕНКО, А. Н. ОСИПЕНКО

ПРОГРАММНЫЕ СРЕДСТВА ПЕРВИЧНОЙ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Практическое руководство

для студентов специальности
1-31 03 03-01 «Прикладная математика
(научно-производственная деятельность)»

Гомель
ГГУ им. Ф.Скорины
2015

УДК 004.62(076)
ББК 32.972я73
О–73

Рецензенты:
кандидат физико-математических наук А. И. Рябченко;
кафедра математических проблем управления
учреждения образования «Гомельский
государственный университет имени Франциска Скорины».

Рекомендовано к изданию научно-методическим советом
учреждения образования «Гомельский государственный
университет имени Франциска Скорины»

Осипенко, Н. Б.

О–73 Программные средства первичной статистической обработки экспериментальных данных : практическое руководство / Н. Б. Осипенко, А. Н. Осипенко ; М-во образования РБ, Гом. гос. ун-т им. Ф. Скорины. – Гомель: ГГУ им. Ф. Скорины, 2015. – 43 с.

ISBN 978-985-439-991-1

Практическое руководство предназначено для оказания помощи студентам в выполнении лабораторных работ по спецкурсу «Программные средства первичной статистической обработки экспериментальных данных» и усвоения знаний по данному спецкурсу. В него включены также контрольные вопросы.

Адресовано студентам специальности 1-31 03 03-01 «Прикладная математика (научно-производственная деятельность)».

УДК 004.62(076)
ББК 32.972я73

ISBN 978-985-439-991-1

Оглавление

Предисловие

Современный уровень развития компьютерных и информационных технологий характеризуется возрастающей сложностью не только отдельных физических и программных компонентов, но и лежащих в основе этих технологий концепций и идей. Целью практического пособия по спецкурсу «Программные средства первичной статистической обработки экспериментальных данных» является оказание помощи студентам в овладении основами технологии применения методов первичной статистической обработки данных и навыками работы с соответствующими прикладными пакетами.

Существует большое разнообразие прикладных пакетов, реализующих широкий спектр статистических методов, их также называют универсальными пакетами или инструментальными наборами. Так, в Microsoft Excel реализован широкий арсенал методов прикладной статистики, задания данного пособия выполняются на базе этого программного обеспечения. Следует заметить, что при использовании статистических методов, так же как и статистического программного обеспечения, – пользователю необходимы специальные навыки и знания.

В руководство включены как общеобразовательные, так и технологические аспекты изучения анализа и обработки экспериментальных данных. Главная задача пособия состоит в приобретении студентами теоретических и практических базовых знаний в области пакетов анализа и обработки данных.

Раздел прикладной статистики: первичная статистическая обработка данных – является одной из важнейших составных частей прикладной математики. В результате выполнения заданий по лабораторным работам студенты ознакомятся с технологиями применения методов первичной статистической обработки данных при исследовании наблюдений о сложных системах с помощью традиционного, а также нетрадиционного инструментария анализа данных; сформируют практические навыки планирования работы в процессе анализа данных на базе существующих и создаваемых программных средств; овладеют некоторыми традиционными и нетрадиционными методами и средствами методов первичной статистической обработки данных прикладной статистики.

Издание адресовано студентам специальности 1-31 03 03 – «Прикладная математика».

Тема 1. Краткая характеристика программного обеспечения прикладной статистики и базовых понятий анализа данных

1.1. Краткая характеристика программного обеспечения статистической обработки данных

Анализ тенденций развития программного обеспечения прикладной статистики (ПО ПС) показывает, что актуальной является задача создания проблемно-ориентированных средств ПО ПС, настроенных или легко адаптируемых на работу с конечным пользователем. Поэтому наибольшим спросом пользуется ПО ПС, позволяющее программисту адаптировать его под свою СУБД, статистику – организовать ассистирование решения задачи, конечному пользователю – иметь интерактивный доступ к ПО ПС для решения задачи и самостоятельного получения содержательных выводов.

Применение ПО ПС является важным этапом прикладных статистических исследований. Успешное статистическое исследование без него не возможно. ПО ПС поддерживает деятельность исследователя, направлено на получение содержательного результата и отражает представление исследователя о методах и средствах получения такого результата.

Приведем в связи с этим краткую характеристику программного обеспечения статистической обработки данных, изменяющиеся формы ПО ПС, тенденции развития ПО ПС и развития форм ПО ПС, наиболее перспективные при проведении статистического исследования.

При всём богатстве ПО ПС этапы решения статистических задач остаются неизменными – построение статистической модели, соответствующей задаче, выбор статистического метода (вместе с программой) и соответствующего ПО ПС, затем – реализация этого метода в конкретном ПО для решения задачи. Получение содержательно интерпретируемого результата не всегда означает окончание исследования. При использовании отобранных методов и средств для анализа информации может быть необходима непрерывная работа в течение нескольких лет. Поэтому выделим постоянно поддерживаемые (исследовательской группой) задачи и назовём эксплуатационными в отличие от хорошо изученных исследовательских задач. Использование ПО ПС планируется на всех этапах исследования, начиная от организации хранения данных и кончая обеспечением многократной прогонки заданий с различными параметрами.

Критерии оценки пакетов. *Понятность для пользователя***.** Ключевую роль при оценке пакета играет сопровождающая его документация. Ясное, короткое, хорошо организованное справочное руководство с алфавитным указателем должно описывать все возможности пакета. Руководство должно содержать не только синтаксические правила, но и наиболее вероятные ошибки. Процедуры должны быть описаны в общепринятых терминах, методы – со ссылкой на литературу и указанием значений стандартных значений параметров, задаваемых по умолчанию. Следует дать указание о действиях с отсутствующими значениями. Вывод результатов должен быть полным, компактным, не избыточным и содержать средства подавления лишней информации. Необходимы графические выводы (гистограмма, вероятностные графики и т. д.). Нужны надписи на графиках и возможность использовать разные шкалы. Должен быть алгоритм для определения стоимости и времени выполнения заданий. Язык управления заданиями следует привести со словарным запасом из той предметной области, на которую он ориентирован. Например, BMDP – для статистиков, SPSS – для специалистов по общественным наукам.

Статистическая эффективность. Пакет должен допускать динамичный и непрерывный процесс обработки. Для этого требуется удобная система файлов для подготовки данных, позволяющая выходу из процедуры служить входом в другую процедуру. Формулы для программ ПСП должны быть правильными, алгоритмы – устойчивыми в вычислительном отношении, правильно запрограммированными. Пакет должен позволять проконтролировать точность данных и процедур. Например, проверить точность обращения матрицы можно с помощью произведения исходной и обратной матрицы.

Удобство эксплуатации. Для удобства эксплуатации ПСП желательно иметь текст программ на исходном языке, как первичную документацию пакета. Пакет должен обладать способностью расширения (включения других программ).

Использование ПО ПС планируется на всех этапах исследования: от организации хранения данных до обеспечения многократной прогонки заданий с различными параметрами.

Коллективный опыт решения задач прикладной статистики породил три формы ПО прикладной статистики: библиотека; пакет; система анализа данных.

Для решения статистической задачи по отработанной методике, период эксплуатации которой несколько лет, эффективны библиотеки. При проведении исследовательского этапа работ в режиме проверки статистических гипотез наиболее эффективны пакеты прикладных программ. Для задач со сложными способами представления и использования данных,

требующих сложных способов организации вычислительного процесса, эффективны системы.

Решение задач средствами *библиотек* – удовольствие дорогое для исследовательской группы и эффективно при использовании нестандартных методов и режимов работы с ПО или при создании сложной многократно используемой трассы обработки (для этого нужен программист очень высокой квалификации).

Пакеты прикладных программ (ППП) наиболее эффективны, когда основные постановки задач анализа данных и режимы исследования совпадают с заложенной в ППП структурой.

Версия[1)] ППП BMDP обладает следующими возможностями: робастные (устойчивые) оценки, дополнительные статистики для таблиц сопряжённости признаков, обратный ход в регрессионном анализе, непараметрические статистические критерии, анализ повторных измерений; графический вывод, включая гистограммы, двумерные графики, графики нормального распределения, графики остатков и графики факторных нагрузок. Программы BMDP разбиваются на 6 категорий: дескриптивные (описание данных), анализ таблиц сопряженности признаков, многомерного анализа, регрессионные, специальные и дисперсионного анализа. Аналогом BMDP является СОМИ (статистическая обработка медицинской информации).

Другим популярным пакетом является Statistical Package for the Social Sciences (SPSS), разработанный Норманом и его сотрудниками из National Opinion Research Center at the University of Chicago. Этот пакет представляет собой комплекс программ, предназначенных для анализа данных общественных наук. Пользователю предоставляется возможность производить много типов анализа при большой гибкости форматов данных, преобразование данных и манипуляции с файлами. SPSS позволяет пользователю производить анализ при помощи управляющих операторов, формулируемых на языке, близком к естественному. Процедуры SPSS включают дескриптивный анализ, простую корреляцию, одномерную и многомерную классификацию, масштабирование Гутмана и другие многомерные процедуры. При использовании ПО ПС принятие решений остается за исследователем. Программа освобождает от рутинной вычислительной работы, но интерпретация полученных результатов зависит от его опыта

[1)] Одним из наиболее распространённых в 80–90-х годах являлся пакет статистических программ Biomedical Computer Programm, разработанный под руководством Диксона в ВЦ Медицинского центра Калифорнийского университета в Лос-Анжелесе. Первая версия пакета BMD появилась в 1961 г. и быстро развивалась за счёт дополнительных программ, улучшения средств и новых статистических методик. Появившаяся в 1975 г. новая версия пакета BMDP – фактически заменила предыдущую.

и знаний. Применение ПО ПС влечет за собой некоторые неудобства: исследователь должен привыкнуть к обозначениям и требованиям ПО ПС, часто не достаточно информации для интерпретации выходных данных; приходится ограничиваться численными методами, примененными в программах, возможно и не самыми эффективными; часто не предусмотрен вывод на печать всей информации, необходимой пользователю (например, точечные оценки без доверительных интервалов); ПО ПС базируется на стандартных статистических методах. При необходимости использовать нестандартный анализ исследователь должен написать свою программу.

Системы анализа данных. Используя методы прикладной статистики, система дает возможность манипулировать данными и получать ответы. В пределах одного задания можно создать набор данных, редактировать, корректировать, проводить статистический анализ, создавать новые наборы, размещать их для хранения на диске без привлечения языка управления заданиями. Остановимся на нескольких системах.

SAS (США) обладает широкими возможностями управления данными среди известного ПО ПС. Язык управления основан на двух типах предложений: DATA, PROC, и позволяет проводить матричные операции (преобразования выборочных данных), организовывать циклы из процедур обработки (т. е. использовать результаты предыдущего анализа).

Сейчас системы символьной математики[1)] (или компьютерной алгебры) выпускаются самого разного «калибра» – от рассчитанной «на всех» системы MathCad, поразительно компактной, быстрой и удобной для простых символьных вычислений системы Derive, и до компьютерных монстров Mathematica, Matlab и Maple, имеющих тысячи встроенных и библиотечных функций и изумительные возможности графической визуализации вычислений. Все эти системы работают на персональных компьютерах, оснащенных популярными операционными системами класса Windows, а также Linux, Unix, Mac и др.

Остановимся на одном из наиболее мощных программных средств по статистической обработке Statistica, которое является универсальной интегрированной программной системой, предназначенной для статистического анализа и визуализации данных, управления базами данных и разработки пользовательских приложений. Она содержит широкий набор процедур анализа для применения в научных исследованиях, технике, бизнесе, а также специальные методы добычи данных. Помимо общих статистических и графических средств, в системе Statistica имеются специали-

[1)] Системы символьной математики долгое время были ориентированы на большие компьютеры. С появлением ПК класса IBM PC и Macintosh и с ростом их возможностей эти системы были переработаны под них и доведены до уровня массовых серийных программных систем.

зированные модули, например, для проведения социологических или биомедицинских исследований, решения технических и, что очень важно, промышленных задач: карты контроля качества, анализ процессов и планирование эксперимента. Работа со всеми модулями происходит в рамках единого программного пакета, для которого можно выбирать один из нескольких предложенных интерфейсов пользователя.

С помощью реализованных в системе Statistica мощных языков программирования, снабженных специальными средствами поддержки, легко создаются законченные пользовательские решения и встраиваются в различные другие приложения или вычислительные среды. Она состоит из следующих основных компонент, объединенных в рамках одной системы: электронных таблиц для ввода и задания исходных данных, а также специальных таблиц для вывода численных результатов анализа; мощной графической системы для визуализации данных и результатов статистического анализа; набора специализированных статистических модулей, в которых собраны группы логически связанных между собой статистических процедур; специального инструментария для подготовки отчетов; встроенных языков программирования SCL (Statistica Command Language) и Statistica Basic, которые позволяют пользователю расширить стандартные возможности системы.

В ряде случаев для проведения законченного статистического исследования не требуется дополнительное программное обеспечение – все этапы статистического анализа, начиная от ввода исходных данных и их преобразований и заканчивая подготовкой отчета или написания собственных процедур обработки, можно выполнить, используя только систему Statistica.

Итак, для систем анализа данных характерно: ориентация на развитые средства манипуляции с данными; использование результатов предыдущего шага анализа; разделение анализа на этапы, последовательно обрабатываемые ведущей программой; ориентация на пользователя-профессионала (владеющего системой и знающего статистику).

Интерактивные средства ПО ПС. Применение интерактивных режимов упрощает обращение к ПО ПС и позволяет добиваться разной степени интеллектуализации, например: ассистирование; хранение и распространение опыта.

Опишем четыре пути создания диалоговых средств прикладного обеспечения прикладной статистики.

1. Диалоговый интерфейс к универсальным ППП. Цель диалога – получение от пользователя числовых параметров предложений входного языка пакета (ЯУП). Но такой диалог не очень эффективен с точки зрения поль-

зователя (хотя не обязательно знать конструкции языка и способы их представления, но необходимо знание пакетной документации) и с точки зрения программиста (диалог затянут и навязчив). Улучшение диалога связано с изменением цели: выяснение требований к задаче, сформулированных в содержательных терминах предметной области (например, вместо «количество признаков» – «количество параметров, описывающих изделие»).

2. Интерактивные режимы работы отдельных программ ПО ПС. Некоторые программы наиболее эффективны при многократной прогонке, например, при подборе параметров; получении некоторого предвидимого результата; при графическом представлении исходных данных и результатов.

3. Интерактивный ППП. Некоторые пакеты ПО ПС разработаны как интерактивные средства анализа данных. Так в ППП ОТЭКС используются эвристические методы, легкие для понимания после пояснения содержательного смысла метода (но не формального). В результате после диалога (определяется тип задачи, особенности данных) управляющая программа определяет последовательность модулей, организует связи и вызов.

4. Интерактивная система анализа данных.

Например, экспертная система (ЭС) CLAVESIN[1)] (оригинальные разработки в области анализа данных: классификация и описание данных) состоит из двух частей: генератор логического вывода OURSIN и пакета анализа данных SICLA. Работа ЭС напоминает работу эксперта, знающего возможности системы, правил обращения с процедурами, данными, командами SICLA. Пользователь ЭС – статистик, знающий статистическую терминологию, содержательный смысл статистических методов и обращающийся к системе как к эксперту, обеспечивающему решение с генератором логического вывода OURSIN. Интерфейс позволяет пополнить первоначальную базу данных, отражающую информацию о структуре и логике работы SICLA; собрать ответы пользователя, связанные со сценарием и поместить их в базу знаний. ЭС осуществляет действия в соответствии с правилами OURSIN, обращается к SICLA для выполнения сценария и опять переходит к диалогу. Таким образом, в течение всего сеанса осуществляется связь двух подсистем OURSIN и SICLA, т.е. CLAVESIN – качественно новый продукт на основе пакета анализа данных SICLA.

Здесь интерактивность реализована как ЭС, анализирующая задачу в содержательной постановке; выбирающая на основании знаний (о задаче и способах ее решения имеющимися средствами) наилучшую трассу обработки, сама ее реализующая и помогающая в интерпретации.

[1)] Данная система реализована в советское время для ЕС ЭВМ, сейчас не имеет практического применения, однако сама идея организации анализа данных до сих пор актуальна и не востребована в современных зарубежных системах анализа данных.

Выделяют следующие *типы диалога* (структурных и лингвистических средств, используемых для обмена сообщениями между пользователями и ЭВМ): команда; меню; заполнение бланка; вопрос, требующий ответа «да» или «нет»; взаимодействие на естественном языке.

Отметим некоторые проблемы, которые необходимо решить при проектировании диалога с точки зрения пользователя, с одной стороны, разработчика, с другой: простота диалога и быстрое достижение конечной цели – детерминированность (однозначность трактовки ответа системой); система действует по формальным правилам – включение неформальных знаний о задаче; выбор одного типа диалога – решение одним типом диалога не обеспечивается.

Для разрешения этих противоречий ограничивается класс решаемых задач.

При решении задач анализа данных выделяется два этапа: исследовательский и эксплуатационный. Первый – этап выработки и проверки статистических гипотез. Второй – использование полученных гипотез для построения вывода. Поэтому с целью оптимизации и внедрения диалоговых систем необходим исследовательский этап, когда пользователь получает инструментальные психологические навыки и знания о системе. Как показывает практика и анализ публикаций, наибольший эффект в результате внедрения диалоговых систем (поддержки пользователя) достигается при выборе одного из двух требований: пользователь является профессионалом, диалоговые средства – рабочим инструментом; диалоговые средства ориентированы на поддержание простейших форм диалога и достижение простых ясно формализуемых целей.

Диалоговое средство работы с ПО ПС намного упрощает статистическую обработку, позволяет организовывать удобный и эффективный режим общения с ЭВМ. Проиллюстрируем это на этапах статистической обработки.

Этап 1. Предварительный анализ реальной системы. Вырабатываются цели и средства исследования на содержательном уровне (терминология) и переводится в формализованную терминологию статистической постановки.

Этап 2. Составление детального плана сбора информации. Представление о данных фиксируется в виде структуры исследуемых выборок. Определяется форма представления данных в ЭВМ и возможности манипулирования.

Этап 3. Сбор данных и ввод в ЭВМ. Уточняются способы образования подвыборок и их анализа, способы содержательного изменения в хранении и манипуляции данными.

Этап 4. Первичная статистическая обработка. Сопоставляются постановка задачи и способы ее решения с возможностями ЭВМ и доступным ПО ПС. Здесь важна роль программиста, владеющего языком общения с ЭВМ с помощью языка управления заданиями, языка управления программами, а также программного обеспечения, разработанного с учетом требования задач.

Этап 5. Составление детального плана вычислительного анализа. Описывается блок-схема анализа с указанием используемых методов, продумывается интерпретация результатов на содержательном языке (для конкретной предметной области).

Этап 6. Вычислительная реализация статистической обработки данных. Формируются типовые задания для ППП, отражающие особенности обработки и хранения данных. Наиболее простой для оптимизации этап.

Этап 7. Подведение итогов исследования. Один из наименее формализованных этапов часто сливается с предыдущим, так как является промежуточным звеном анализа. Он может вызывать изменения способов представления данных, в статистическом аппарате, методах исследования, в содержательной постановке задач.

1.2. Базовые понятия анализа данных

Этапы работ, предшествующие обработке экспериментальных данных. Всех специалистов, профессионально занимающихся обработкой статистических данных, условно можно разделить на три категории: 1) приверженцы классической математической статистики (объектами их исследований обычно являются некоторые разделы биологии или физики); 2) представители школы обработки экспериментальных данных в рамках идеологии исследования операций (предметом их разработок чаще всего бывают результаты активных экспериментов над сложной технической системой); 3) специалисты по прикладной статистике и анализу данных, ориентированные на исследование естественных и социальных систем в таких, например, областях, как геология, медицина, экономика и социология. Характер данных и методологическое видение проблемного материала во всех трёх случаях столь различны, что в действительности эти три течения статистических исследований следовало бы признать самостоятельными. В настоящем пособии за основу принята концепция по отношению к прикладной статистике и анализу данных, окончательно сформировавшаяся к концу 80-х годов. Наиболее полно эта область прикладной математики изложена в трёхтомном справочном издании по прикладной статистике под редакцией С. А. Айвазяна. В текстах лекций

использована концепция стиля подачи материала упомянутого выше справочника.

Прикладная статистика. Целесообразность введения термина «прикладная статистика» наряду с привычным понятием «математическая статистика» объясняется тем, что для внедрения метода статистической обработки необходимо дополнительно провести сложную и наукоемкую работу. Условно разобьем её на ряд этапов: 1) адекватно «приложить» исходные модельные допущения к реальной задаче; 2) представить имеющуюся исходную информации (физические сигналы, геологические срезы и др.) в стандартной форме; 3) разработать вычислительный алгоритм и его программное обеспечение; 4) организовать удобный режим общения с ЭВМ в процессе решения задачи. Весь комплекс выше перечисленных действий и составляет содержание прикладной статистики

Исходя из вышесказанного, дадим определение, введенное в 1983 г. С. А. Айвазяном [1, стр. 19]. Прикладная статистика – это самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели предназначенные для организации сбора, стандартной записи, обработки статистических данных с целью их удобного представления (в том числе и на ЭВМ), интерпретации и получения научных и практических выводов.

Заметим, что некоторые специалисты, в частности, французские, вместо введенного термина «прикладная статистика» используют понятие «анализ данных», трактуя его в расширительном смысле.

Методологические принципы многомерного статистического анализа данных. Ниже приведены четыре принципа.

Принцип 1. Эффект существенной многомерности. Статистический анализ должен опираться одновременно на совокупность взаимосвязанных свойств объектов.

Принцип 2. Возможность лаконичного объяснения природы анализируемых многомерных структур. На нем построены такие важнейшие разделы математического аппарата многомерного статистического анализа данных, как метод главных компонент и факторный анализ, многомерное шкалирование, целенаправленное проецирование в разведочном анализе данных и др.

Принцип 3. Максимальное использование «обучения» в настройке математических моделей многомерного статистического анализа данных.

Принцип 4. Оптимизационная формулировка задач многомерного статистического анализа данных, в частности, классификации и снижения размерности.

Цели эксперимента в науке и промышленности. Экспериментальные методы широко используются как в науке, так и в промышленности, однако нередко с весьма различными целями. Обычно основная цель научного исследования состоит в том, чтобы показать статистическую значимость эффекта воздействия определенного фактора на изучаемую зависимую переменную. В условиях промышленного эксперимента основная цель обычно заключается в извлечении максимального количества объективной информации о влиянии изучаемых факторов на производственный процесс с помощью наименьшего числа дорогостоящих наблюдений. Если в научных приложениях методы дисперсионного анализа используются для выяснения реальной природы взаимодействий, проявляющейся во взаимодействии факторов высших порядков, то в промышленности учет эффектов взаимодействия факторов часто считается излишним в ходе выявления существенно влияющих факторов.

Указанное отличие приводит к существенному различию методов, применяемых в науке и промышленности. Если просмотреть классические учебники по дисперсионному анализу, то обнаружится, что в них, в основном, обсуждаются планы с количеством факторов не более пяти (планы же с более чем шестью факторами обычно оказываются бесполезными). Основное внимание в данных рассуждениях сосредоточено на выборе общезначимых и устойчивых критериев значимости. Однако если обратиться к стандартным учебникам по экспериментам в промышленности, то окажется, что в них обсуждаются, в основном, многофакторные планы (например, с 16 или 32 факторами), в которых нельзя оценить эффекты взаимодействия, и основное внимание сосредоточивается на том получении несмещенных оценок главных эффектов (или, реже, взаимодействий второго порядка) с использованием наименьшего числа наблюдений.

Подходы к статистическому анализу данных и причины малоэффективного использования машинных методов анализа данных. Развитие теории и практики статистической обработки данных шло в двух параллельных направлениях. Первое включает методы математической статистики, предусматривающие возможность классической вероятностной интерпретации анализируемых данных и полученных статистических выводов (вероятностный подход). Второе направление содержит статистические методы, которые априори не опираются на вероятностную природу обрабатываемых данных, т. е. остаются за рамками научной дисциплины «математическая статистика» (логико-алгебраический подход). Ко второму подходу исследователь вынужден обращаться лишь тогда, когда условия сбора исходных данных не укладываются в рамки статистического ансамбля, т. е. в ситуации, когда не имеется практической или хотя бы принципиально мысленно представимой возможности многократного

тождественного воспроизведения основного комплекса условий, при которых производились измерения анализируемых данных.

Типы реальных ситуаций с позиции выполнения требований статистического ансамбля. Выделяют три типа реальных ситуаций: с высокой работоспособностью вероятностно-статистических методов; с допустимостью вероятностно-статистических приложений (при этом нарушатся требования сохранения неизменными условий эксперимента); с недопустимостью вероятностно-статистических приложений (в этом случае идея многократного повторения одного и того же эксперимента в неизменных условиях является бессодержательной).

Причины малоэффективного использования машинных методов анализа данных. В последние десятилетия для решения многочисленных практических задач стали интенсивно использоваться машинные методы анализа данных. Не будучи математиком, специалист выбирает модель обработки либо по традиции, либо из доступного и легко интерпретируемого математического обеспечения. При этом он, как правило, не задумывается: соответствует ли его модель природе исходных данных? Подобная некомпетентность исследователя обусловлена рядом причин. Приведем наиболее важные из них.

1. Отсутствие подробных описаний алгоритмов программ, а также информации об ограниченности возможностей модельных алгоритмов и ориентиров по их применению (как в литературе, так и в сопроводительной документации к программному обеспечению).

2. Разделение труда специалиста и математика, появление ничейной зоны деятельности.

Математик ограничен рамками: «Есть множество объектов, описанных признаками...» – «В результате получены закономерности, которые неплохо согласуются с представлениями специалиста...». Он не задумывается о содержании предложенных признаков, о том, имеет ли смысл их совместный анализ, учтены ли все существенные факторы. С другой стороны, специалист не вникает в механизм обработки данных, не интересуется, на каком этапе происходит искажение информации. Интуитивные соображения привлекают только на этапе интерпретации результата, в котором ничего изменить нельзя.

3. Организационная разобщенность разработчиков алгоритмов и программ; отсутствие гибкой системы распространения программного обеспечения анализа данных.

Сегодня ничейная зона деятельности частично устраняется за счет разработки так называемых CASE[1)]-средств.

[1)] Computer Aided Software/System Engineering

CASE-технология представляет собой совокупность методологий анализа, проектирования, разработки и сопровождения сложных систем программного обеспечения (ПО), поддержанную комплексом взаимоувязанных средств автоматизации. CASE – это инструментарий для системных аналитиков, разработчиков и программистов, заменяющий им бумагу и карандаш на компьютер для автоматизации процесса проектирования и разработки ПО.

Основная **цель** CASE состоит в том, чтобы *отделить проектирование модели информационной системы или программного обеспечения от его кодирования* и последующих этапов разработки, а также скрыть от разработчиков все детали среды разработки и функционирования ПО. CASE-системы основаны на методологии структурного анализа и проектирования, обеспечивают строгое и наглядное описание проектируемой системы.

Структурные методологии зарождались как средства анализа и проектирования ПО. Сейчас CASE-технологии успешно применяются для моделирования практически всех предметных областей, преимущественно для бизнес-анализа (фактически, модели деятельности предприятий «как есть» и «как должно быть» строятся с применением методов структурного системного анализа и поддерживающих их CASE-средств) и системного анализа и проектирования (практически любая современная крупная программная система разрабатывается с применением CASE-технологий по крайней мере на этапах анализа и проектирования, что связано с большой сложностью данной проблематики и со стремлением повысить эффективность работ).

Краткая характеристика основных этапов обработки данных. Опишем общую логическую схему статистического анализа данных в виде семи этапов, перечислив их в хронологическом порядке (хотя они могут реализовываться в режиме итерационного взаимодействия).

Этап 1. Исходный (предварительный) анализ исследуемой системы. На этом этапе определяются: основные цели исследования на неформализованном, содержательном уровне; совокупность единиц (объектов), представляющая предмет статистического исследования; набор параметров-признаков (x^1,…, x^p) для описания обследуемых объектов; степень формализации соответствующих записей при сборе данных; время и трудозатраты, объем работ; выделение ситуаций, требующих предварительной проверки перед составлением детального плана исследований; формализованная постановка задачи; в каком виде осуществляется сбор первичной информации и введение в ЭВМ. Если обработка проводится с помощью существующего пакета статистической обработки,

то трудоемкость этого этапа бывает сравнима с суммарной трудоемкостью остальных шести этапов.

Этап 2. Составление плана сбора исходной информации. При составлении детального плана сбора первичной информации необходимо учитывать, как и для чего данные анализируются, т. е. учитывать полную схему анализа. Этот этап называют «организационно-методической подготовкой», так как на нем планируется: какой должна быть выборка – случайной, пропорциональной, расслоенной (если используется аппарат общей теории выборочных обследований); объем и продолжительность исследования; схема проведения активного эксперимента (в случае, если он возможен) с привлечением методов планирования эксперимента и регрессионного анализа для определения некоторых входных переменных.

Этап 3. Сбор исходных данных, их подготовка и введение в ЭВМ. Внесение в ЭВМ полного и краткого определения используемых терминов. Существует два вида представления исходных данных: матрица «объект-признак»: со значениями *k*-го признака, характеризующего *i*-й объект в момент *t* (числа, текст): $x_i^{(k)}(t)$, $t = t_1 \dots t_N$, $k = \overline{(1, p)}$, $i = \overline{(1, N)}$; и матрица «объект-объект» $\rho_{ij}(t)$ – характеристик попарной близости *i*-го и *j*-го объектов (при этом $m = N$) или признаков (при этом $m = p$) в момент *t*. Второй вид представления часто используется в социологии, где данные собираются с помощью специальных опросников, анкет. Примером характеристики попарной близости признаков может служить ковариационная матрица.

Этап 4. Первичная статистическая обработка данных. При первичной статистической обработке данных обычно решаются следующие задачи: отображение вербальных переменных в номинальную (с предписанным числом градаций) или ординальную (порядковую) шкалу; статистическое описание исходных совокупностей с определением пределов варьирования переменных; анализ резко выделяющихся переменных; восстановление пропущенных значений наблюдений; проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных; унификация типов переменных, когда с помощью различных приёмов добиваются унифицированной записи всех переменных; экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризация сведений о природе изучаемых распределений (эту разновидность первичной статистической обработки называют иногда процессом составления сводки и группировки); вычислительная реализация учета сложности задачи и возможностей ЭВМ; формулировка задачи на входном языке пакета статистической обработки.

Этап 5. Выбор основных методов и алгоритмов статистической обработки данных, составление детального плана вычислительного анализа материала. Определяются основные группы, для которых будет проводиться дальнейший анализ. Пополняется и уточняется тезаурус содержательных понятий. Описывается блок-схема анализа с указанием привлекаемых методов. Формируется оптимизационный критерий, по которому выбирается один из альтернативных методов.

Этап 6. Реализация плана вычислительного анализа исходных данных (непосредственная эксплуатация ЭВМ). Исследователь на этом этапе осуществляет управление вычислительным процессом, формирует задачу обработки и описания данных на входном языке пакета. Учитываются размерность задачи, алгоритмическая сложность вычислительного процесса, возможности ЭВМ и особенности данных (обусловленность операций, надежность используемых оценок параметров).

Этап 7. Подведение итогов. Строится формальный отчет о проведенном исследовании. Интерпретируются результаты применения статистических процедур (оценки параметров, проверки гипотез, отображения в пространство меньшей размерности, классификации). При интерпретации могут использоваться методы имитационного моделирования.

Если исследование проводится в рамках теоретико-вероятностного подхода, то выводы формируются в терминах оценок неизвестных параметров, или в виде отчета о справедливости гипотез с указанием количественной степени достоверности. В случае логико-комбинаторного подхода вероятностная интерпретация не делается. Работа завершается содержательной формулировкой новых задач, вытекающих из проведенного исследования.

Разведочный анализ данных. Этап разведочного (предмодельного) анализа данных (РАД) зачастую игнорируется или реализуется поверхностно в ходе прикладных статистических исследований. Одна из главных причин – отсутствие необходимой научно-методологической литературы. Большое внимание этим вопросам уделено в [1]. Основная цель РАД – построить некоторую статистическую модель в виде эмпирического описания структуры данных, которую необходимо будет потом в ходе статистического исследования верифицировать. Основная задача РАД – переход к компактному описанию данных при возможно более полном сохранении существенных аспектов информации, содержащихся в данных. Методы РАД направлены на «прощупывание» вероятностной и геометрической природы обрабатываемых данных и предназначены для формирования адекватных реальности рабочих исходных допущений, на которых строится дальнейшее исследование. РАД является необходимым и естественным

моментом первичной статистической обработки и применяется, когда отсутствует априорная информация о статистическом или причинном механизме порождения имеющихся у исследователя данных.

Важнейшим элементом РАД является широкое использование визуального представления многомерных данных. Его возможности возросли благодаря появлению динамических форм визуального представления. Для этого многомерные данные отображаются в пространство низкой размерности с сохранением существенных структурных особенностей. При этом структура данных может оказаться такой сложной, что небольшого числа проекций недостаточно для их представления. Тогда структуру описывают за счет агрегирования информации, содержащейся в большом числе низкоразмерных проекций.

К РАД относятся методы, дающие наглядное представление о структуре многомерных данных в пространствах малой размерности. В случае, если размерность пространства, куда отображаются данные, меньше или равна трем, то эти методы относятся к собственно разведочному анализу, когда по некоторому критерию при помощи вычислительной процедуры оптимизации ищут отображения, дающие наиболее выразительные проекции, а окончательное решение принимается визуально путем анализа (в одномерном случае – это гистограмма, на плоскости – диаграмма рассеивания). К РАД относятся также методы, связанные с линейным проецированием, упрощением описания с помощью компонентного анализа и многомерного шкалирования, кластер-анализа, анализа соответствий (для неколичественных переменных).

Модели структуры многомерных данных. Пусть данные заданы в виде матрицы данных. Объекты можно представить в виде точек в многомерном (*p*-мерном) пространстве. Для описания структуры этого множества точек в РАД используется одна из следующих статистических моделей:

– модель облака точек примерно эллипсоидальной конфигурации;

– кластерная модель, т. е. совокупность нескольких «облаков» точек, достаточно далеко отстающих друг от друга;

– модель «засорения» (компактное облако точек и при этом присутствуют дальние выбросы);

– модель носителя точек как многообразия (линейного или нелинейного) более низкой размерности, чем исходное: типичным примером является выборка из вырожденного распределения; в рамках этой модели можно рассматривать и регрессионную модель, когда соответствующее многообразие допускает функциональное представление $Y = F(X) + \varepsilon$, где Y – прогнозируемые, X –предсказывающие признаки, $F(X)$ – функция регрессии, ε – ошибка;

– дискриминантная модель, когда точки разделены на несколько групп и дана информация о их принадлежности к той или иной группе.

– эмпирический образ данных в виде покрытия выборочных точек многомерного признакового пространства сетью гиперпараллелепипедов с оцененной плотностью распределения (многомерный аналог гистограммы).

1.3. Анализ данных в Microsoft Excel

Microsoft Excel имеет большое число статистических функций. Некоторые из них являются встроенными, некоторые доступны после установки пакета анализа. Средства, включенные в пакет анализа данных, доступны через команду ***Анализ данных*** меню ***Сервис***. Если эта команда отсутствует, то в меню ***Сервис/Надстройки*** необходимо активировать пункт ***Пакет анализа***. Рассмотрим некоторые инструменты, относящиеся к первичной статистической обработке, включенные в Пакет анализа.

Описательная статистика (Descriptive statistics) – техника сбора и суммирования количественных данных, которая используется для превращения массы цифровых данных в форму, удобную для восприятия и обсуждения. Цель описательной статистики – обобщить первичные результаты, полученные в результате наблюдений и экспериментов.

Выбрав в меню ***Сервис Пакет анализа*** и инструмент анализа ***«Описательная статистика»,*** получаем одномерный статистический отчет, содержащий информацию о центральной тенденции и изменчивости или вариации входных данных. В состав описательной статистики входят следующие характеристики: среднее; стандартная ошибка; медиана; мода; стандартное отклонение; дисперсия выборки; эксцесс; асимметричность; интервал; минимум; максимум; сумма; счет. Рассмотрим, что же представляют собой характеристики описательной статистики.

Центральная тенденция. Измерение центральной тенденции заключается в выборе числа, которое наилучшим способом описывает все значения признака набора данных. Такое число имеет как свои достоинства, так и недостатки. Рассмотрим две характеристики этого измерения, а именно: среднее значение и медиану.

Главная цель ***среднего*** – представление набора данных для последующего анализа, сопоставления и сравнения. Значение среднего легко вычисляется и может быть использовано для последующего анализа. Оно может быть вычислено для данных, измеряемых по интервальной шкале, и для некоторых данных, измеряемых по порядковой шкале. Среднее значение рассчитывается как среднее арифметическое набора данных: сумма всех значений выборки, деленная на объем выборки. «Сжимая» данные таким образом, мы теряем много информации.

Среднее значение очень информативно и позволяет делать вывод относительно всего исследуемого набора данных. При помощи среднего мы получаем возможность сравнивать несколько наборов данных или их частей. Но при анализе данных средним не следует злоупотреблять, необходимо учитывать его свойства и ограничения. Известны характеристики «средняя температура по больнице» или «средняя высота дома», показывающие некорректность использования этой меры центральной тенденции для некоторых случаев.

Свойства среднего. 1. При расчете среднего не допускаются пропущенные значения данных. 2. Среднее может вычисляться для числовых и дихотомических шкал. 3. Для одного набора данных может быть рассчитано одно и только одно значение среднего.

Информативность среднего значения переменной высока, если известен ее доверительный интервал. Доверительным интервалом для среднего значения является интервал значений вокруг оценки, где с данным уровнем доверия находится «истинное» среднее популяции. Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Ширина доверительного интервала зависит от размера выборки и от разброса данных.

С увеличением размера выборки точность оценки среднего возрастает. С увеличением разброса значений выборки надежность среднего падает. Если размер выборки достаточно большой, качество среднего увеличивается независимо от выполнения предположения нормальности выборки.

Медиана – точная середина выборки, которая делит ее на две равные части по числу наблюдений. Обязательным условием нахождения медианы является упорядоченность выборки. Таким образом, для нечетного количества наблюдений медианой выступает наблюдение с номером $(n+1)/2$, где n – количество наблюдений в выборке. Для четного числа наблюдений медианой является среднее значение наблюдений $n/2$ и $(n+2)/2$.

Некоторые свойства медианы. 1. Для одного набора данных может быть рассчитано одно и только одно значение медианы. 2. Медиана может быть рассчитана для неполного набора данных, для этого необходимо знать номера наблюдений по порядку, общее количество наблюдений и несколько значений в середине набора данных.

Характеристики вариации данных. Наиболее простыми характеристиками выборки являются максимум и минимум. Минимум – наименьшее значение выборки. Максимум – наибольшее значение выборки. Размах – разница между наибольшим и наименьшим значениями выборки. Дисперсия – среднее арифметическое квадратов отклонений значений от их среднего. Стандартное отклонение – квадратный корень из дисперсии выборки – мера того, насколько широко разбросаны точки данных относительно их среднего.

Эксцесс показывает «остроту пика» распределения, характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение (пик заострен). Отрицательный эксцесс обозначает относительно сглаженное распределение (пик закруглен). Если эксцесс существенно отличается от нуля, то распределение имеет или более закругленный пик, чем нормальное, или, напротив, имеет более острый пик (возможно, имеется несколько пиков). Эксцесс нормального распределения равен нулю.

Асимметрия или асимметричность показывает отклонение распределения от симметричного. Если асимметрия существенно отличается от нуля, то распределение несимметрично, нормальное распределение абсолютно симметрично. Если распределение имеет длинный правый хвост, асимметрия положительна; если длинный левый хвост – отрицательна.

Выбросы (outliers) – данные, резко отличающиеся от основного числа данных. При обнаружении выбросов перед исследователем стоит дилемма: оставить наблюдения-выбросы либо от них отказаться. Второй вариант требует серьезной аргументации и описания. Полезным будет провести анализ данных с выбросами и без и сравнить результаты. Следует помнить, что при применении классических методов статистического анализа, которые, как правило, не являются робастными (устойчивыми), наличие выбросов в наборе данных приводит к некорректным результатам. Если набор данных относительно мал, исключение данных, которые считаются выбросами, может заметно повлиять на результаты анализа. Наличие выбросов в наборе данных может быть связано с появлением так называемых «сдвинутых» значений, связанных с систематической ошибкой, ошибок ввода, ошибок сбора данных и т. д. Иногда к выбросам могут относиться наименьшие и наибольшие значения набора данных.

Вопросы для самоконтроля

1. Создание какого прикладного обеспечения прикладной статистики является наиболее актуальной задачей и почему?
2. Как классифицируется ПО ПС?
3. Опишите этапы решения статистических задач с использованием и без использования ПО ПС.
4. Каковы критерии оценки пакетов? Дайте пример возможного измерения этих критериев.
5. Какие формы ПО прикладной статистики бывают? Приведите примеры.
6. Какие пути создания диалоговых средств прикладного обеспечения прикладной статистики существуют?

7. Назовите типы диалога.

8. Какие категории специалистов профессионально занимаются обработкой статистических данных?

9. Какие этапы работ для внедрения метода статистической обработки существуют?

10. Дайте определение прикладной статистики.

11. В чем смысл эффекта существенной многомерности?

12. В чем суть возможности лаконичного объяснения?

13. Что означает использование «обучения» в настройке математических моделей?

14. Какова цель научного исследования? Приведите примеры.

15. Какова цель промышленного эксперимента? Приведите примеры.

16. Каковы возможные подходы к статистическому анализу данных?

17. Какие типы реальных ситуаций с позиции выполнения требований статистического ансамбля существуют?

18. Какое явление в теории вероятностей определяется как случайное?

19. Приведите примеры подходов к статистическому анализу данных.

20. Дайте сравнение подходов к статистическому анализу данных.

21. Укажите причины малоэффективного использования машинных методов анализа данных.

22. Какова основная цель CASE?

23. Дайте краткую характеристику основных этапов обработки данных.

24. Какие вопросы решаются на этапе предварительного анализа исследуемой системы? В чем отличия объекта от предмета исследования?

25. Какие способы сбора первичной информации выделяют?

26. Назовите два вида представления исходных данных. Приведите примеры.

27. В чем особенность исходных данных в задачах многомерной статистики.

28. Что такое первичная статистическая обработка данных? Назовите её виды.

29. Какова основная цель разведочного анализа данных?

30. Приведите методы разведочного анализа данных.

31. Какие модели структуры многомерных данных выделяют?

32. Что такое унификация типа переменных? В чем её задачи.

33. Какие характеристики входят в состав описательных статистик?

34. Перечислите свойства описательных статистик: среднее; стандартная ошибка; медиана; мода; стандартное отклонение; дисперсия выборки; эксцесс; асимметричность; минимум; максимум.

Тема 2. Непараметрическая статистическая аппроксимация закона распределения

2.1. Типы задач аппроксимации распределений

Первичные данные, полученные при наблюдении, обычно трудно обозримы. Для того чтобы начать анализ, в них надо внести некоторый порядок и придать им удобный для исследователя вид. В частности, для начала желательно получить представление об одномерных распределениях случайных величин, входящих в данные.

Примечание – Визуализация – это инструментарий, позволяющий увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения. Визуализация данных может быть представлена в виде графиков, схем, гистограмм, диаграмм и т. д. Кратко роль визуализации можно описать такими ее возможностями: поддержка интерактивного и согласованного исследования; помощь в представлении результатов; использование глаз (зрения), чтобы создавать зрительные образы и осмысливать их. Но результаты визуализации иногда могут вводить пользователя в заблуждение. Допустим, мы имеем информацию о прибыли компании А за период с 2000 по 2005 год: 2000–1100, 2001–1101, 2002–1104, 2003–1105, 2004–1106, 2005–1107. Построим гистограмму в Excel по этим данным. Гистограмма представляет собой визуальное изображение распределения данных. Эта информация отображается при помощи серии прямоугольников или полос одинаковой ширины, высота которых указывает количество данных в каждом классе. Используя все значения построения графика, принятые по умолчанию, получаем гистограмму, приведенную на рисунке 2.1. Данный рисунок демонстрирует значительный рост прибыли компании А за период с 2000 по 2005 года. Однако, если мы обратим внимание на ось *y*, показывающую величину прибыли, то увидим, что эта ось пересекает ось *x* в значении, равном 1096. Фактически, ось *y* со значениями от 1096 до 1108 вводит пользователя в заблуждение. Изменив значения параметров, отвечающих за формат оси *y*, получаем график, приведенный на рисунке 2.2. Ось *y* со значениями от 0 до 2000 дает пользователю правильную информацию о незначительном изменении прибыли компании.

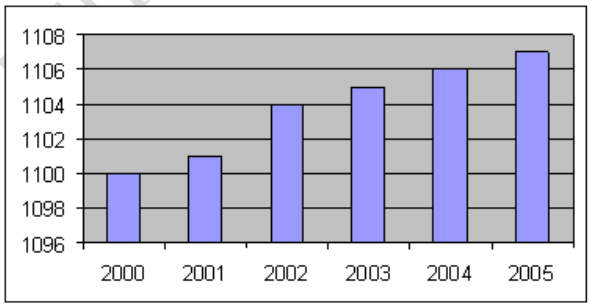

Рисунок 2.1 – Гистограмма, минимальное значение оси *y* равно 1096

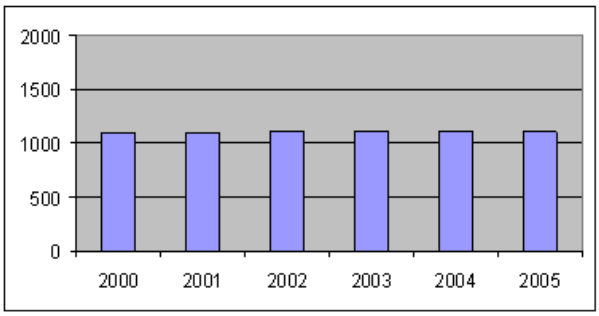

Рисунок 2.2 – Гистограмма, минимальное значение оси *y* равно 0

Существуют два типа задач аппроксимации распределений. Если вид функции распределения известен, но не известны ее параметры, тогда задача сводится к параметрическому оцениванию. Бывают ситуации, когда конкретный вид функции распределения неизвестен и о виде распределения можно сделать лишь самые общие предположения. При таких условиях аппроксимацию неизвестной функции распределения на основе выборки $(x_1, x_2, \ldots, x_N)$ называют непараметрической. Рассмотрим вначале этот случай.

2.2. Классические методы статистической аппроксимации

Классическими методами статистической аппроксимации функции плотности являются гистограмма (равноинтервальная и равнонаполненная) и полигон частот.

Выборочная функция плотности распределения $f_N(x)$ или гистограмма (*равноинтервальная*) строится следующим образом. Делим промежуток [*a,b*], на котором сосредоточены данные выборки на S интервалов $\Delta_1, \Delta_2, \ldots, \Delta_S$, равной длины *h = (b-a)/S*. Подсчитываем число наблюдений $m_1, m_2, \ldots, m_s$, , попавших в интервал $\Delta_1, \Delta_2, \ldots, \Delta_S$, соответственно. Полагаем

$$f_N(x) = m_i / (N \cdot h), \; x \in \Delta_i \qquad (2.1)$$

Полигон частот получают путем сглаживания гистограммы.

Выборочная функция плотности распределения $f_N(x)$ или гистограмма (*равнонаполненная*) строится исходя из предположения, что вся площадь под графиком оценки функции $f_N(x)$ разбивается точками $\tilde{x}_1, \ldots, \tilde{x}_k$ на *k* равных частей, где $\tilde{x}_1$ – стоящее на *N/k*-м месте в вариационном ряду, $\tilde{x}_2$ – стоящее на 2*(*N/k)* -м месте в вариационном ряду и т. д.. Тогда площадь каждой части равна $\Delta_i h_i = 1/k$, откуда $h_i = 1/(k \cdot \Delta_i)$. Для конкретной выборки рассчитываются длины интервалов $\Delta_1, \Delta_2, \ldots, \Delta_k$: Δ_1 – длина первого интервала [*a,* $\tilde{x}_1$] определяется как $\Delta_1 = \tilde{x}_1 - a$, используя которую можно посчитать высоту h_1 по формуле $h_1 = 1/(k \cdot \Delta_1)$ и т. д. На основании полученных значений длины основания и высоты каждого прямоугольника гистограммы получаем оценку $f_N(x)$.

2.3. Оценка плотности распределения вероятностей «ядерного» типа

Для малых выборок ($N < 30$) гистограмма и полигон частот оказываются обычно искаженными за счет тех или иных случайных локальных отклонений, связанных с отсутствием необходимого числа объектов. Одним из способов частично ликвидировать этот пробел явилась «ядерная» аппроксимация, которая путем «размазывания» имеющихся точек заполняет на гистограмме «впадины» и срезает «пики». Отметим, что «ядерное» сглаживание учитывает особенность функции плотности распределения $\int_a^b f(x)dx = 1$, и потому из всех методов сглаживания является наиболее корректным.

Оценка плотности распределения для большинства методов «ядерного» типа обобщенно может быть выражена линейной суммой двух компонент: априорной и эмпирической:

$$f(x) = a_0 f_0(x) + \frac{1-a_0}{N} \sum_{i=1}^{N} p(x - x_i), \tag{2.2}$$

где $f_0(x)$ – априорная компонента;

$p(x - x_i)$– составляющая эмпирической компоненты, связанная с *i*-ой реализацией выборки (заметим, что x_i играет роль параметра);

a_0 – вес априорной компоненты.

Различным методам исследования соответствуют разные значения $a_0 \in [0,1]$ и разные виды функции $p(x - x_i)$. Широко известны оценки *f*(*x*) типа (2.3) при значении $a_0 = 0$. В методе прямоугольных вкладов (МПВ):

$$a_0 = 1/(N+1), \tag{2.3}$$

$$f_0(x) = \begin{cases} 1/(b-a), & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}, \tag{2.4}$$

$$p(x - x_i) = \begin{cases} 1/d, & x \in [x_i - d/2, x_i + d/2] \\ 0, & x \notin [x_i - d/2, x_i + d/2] \end{cases}, \tag{2.5}$$

где [*a*,*b*] – интервал изменения случайной величины *x*;

d – ширина функции вклада.

В качестве *d* может быть взято, например: $d = (b-a)/s$, где $s \in [5,9]$.

Алгоритм ядерной аппроксимации функции плотности распределения имеет следующий вид.

Этап 1. Задается множество точек $\tilde{y}_j : \{x_i - d/2, x_i + d/2\},\ i = \overline{1, N}$;

Этап 2. Полученное множество точек $\tilde{y}_j$ упорядочивается по возрастанию и получаем вариационный ряд: $y_1, y_2, ..., y_{2N}$;

Этап 3. Определяется «ядерная» аппроксимация функции плотности распределения:

$$f(x) = \begin{cases} 0, & x < y_1, \\ \dfrac{s-1}{(N+1)(b-a)s} + \left(\dfrac{N}{N+1} \cdot b_j \right) \Big/ d, & y_j \le x < y_{j+1}, \\ 0, & x \ge y_{s+1}, \end{cases} \qquad (2.6)$$

где b_j – количество точек исходной выборки, попавших в интервал $[y_j, y_{j+1})$, а y_j ($j = \overline{1, s+1}$) – некоторое подмножество точек из множества $y_1, y_2, ..., y_{2N}$.

Вопросы для самоконтроля

1. Дайте определение функции плотности распределения вероятности.
2. Проведите геометрическую интерпретацию функции плотности распределения вероятности и её свойства.
3. Каковы возможные подходы к оцениванию функции плотности распределения вероятности?
4. Дайте определение гистограммы и полигона частот.
5. В чем заключается метод прямоугольных вкладов?
6. Поясните смысл априорной и эмпирической компонент в формуле оценки плотности распределения.

Лабораторная работа

***Цель работы*:** получение практических навыков построения непараметрической оценки функции плотности распределения вероятности.

***Материалы и оборудование*:** персональный компьютер.

Задание 1. Для исходных данных в соответствии с вариантом построить непараметрическую оценку функции плотности распределения вероятности в виде двух равноинтервальных гистограмм и полигонов частот для статистических данных с разбиением на 5 и 10 интервалов.

Задание 2. Для исходных данных в соответствии с вариантом построить непараметрическую оценку функции плотности распределения вероятности в виде двух гистограмм и полигонов частот с разбиением на

равнонаполненные интервалы для статистических данных по 10 % и 20 % выборочных значений.

Задание 3. Для исходных данных в соответствии с вариантом построить непараметрическую оценку функции плотности распределения вероятности методом прямоугольных вкладов в виде полигона частот.

Задание 4. Полученные варианты оценок распределения сравнить и сделать выводы.

Рекомендации по выполнению работы

Для выполнения задания 1 при работе в среде пакета Excel использовать функцию СЧЕТЕСЛИ, а также для построения полигона по равноинтервальной гистограмме, равнонаполненной гистограммы и соответствующего ей полигона частот необходимо использовать точечный график. Для этого нужно подсчитать координаты точек столбчатой диаграммы и задать их при построении точечного графика.

При выполнении задания 2 для исходных данных в соответствии с вариантом построить непараметрическую оценку функции плотности распределения вероятности методом прямоугольных вкладов в виде полигона частот. Для построения непараметрической оценки функции плотности распределения вероятности методом прямоугольных вкладов в виде полигона частот необходимо воспользоваться формулами (2.3–2.6), определить для графика 10 координат (x, $f(x)$), где в качестве x взять точки вариационного ряда исходной выборки с равным шагом, а $f(x)$ – определить по формуле (2.6).

Тема 3. Оценка функции распределения с помощью нормальной вероятностной бумаги

Пусть даны N наблюдений $x_1,...,x_N$, извлеченные из генеральной совокупности с функцией распределения $F(t)$. Пусть $x^{(1)},...,x^{(N)}$ – упорядоченный по возрастанию ряд наблюдений. Тогда за оценку $F(t)$ принимают

$$\hat{F}(t) = i/N \text{ , где } i \to \max_j x^{(j)};\ x^{(j)} \le t \ . \qquad (3.1)$$

В тех случаях, когда требуется проверить гипотезу о том, что случайная величина имеет функцию распределения $G(t)$, принадлежащую семейству вида $F((t-\mu)/\sigma)$, где $F(.)$ известная непрерывная функция распределения, при построении оценки $\hat{F}(.)$ часто используют специальную шкалу, откладывая по оси ординат вместо $\hat{F}(t)$ величину $V = F^{-1}\left(\hat{F}(t)\right)$, где F^{-1} – функция, обратная к F. В этом случае в координатах (t,v) график $G(t)$ превращается

в прямую линию, по положению которой можно легко оценить параметры μ и σ. Заметим, что наибольшее распространение на практике получила нормальная вероятностная бумага, для которой $V = \Phi^{-1}$, где $\Phi(.)$ – стандартная функция нормального распределения.

Алгоритм оценки параметров с помощью вероятностной бумаги. Опишем алгоритм оценки с помощью вероятностной бумаги параметров центра $\hat{\mu}$ и разброса $\hat{\sigma}$. Работа осуществляется в несколько этапов.

Этап 1. Строится вероятностная бумага. Для этого внизу окна графика на оси абсцисс (см. рисунок 3.1) откладывается интервал $[x_{\min}, x_{\max}] = [x^{(i)}, x^{(N)}]$. Масштаб подбирается так, чтобы интервал занял ширину окна, за исключением левого отступа 7–8 см. Ось ординат проводится с отступом от левого края 4–5 см. При этом пунктиром отделяется шкала величины $V = \Phi^{-1}$, которая равномерно изменяется от –3 до 3. Таким образом, точка V = –3,0 будет находиться на оси абсцисс, а точка V = 3,0 будет находиться в верхнем левом углу. Слева от пунктирной оси делаются отметки шкалы V, а между осью V и $\hat{F}(t)$ отметки вероятности p: 0,01; 0,05; 0,1; 0,25; 0,5; 0,75; 0,9; 0,95; 0,99.

Шкала вероятностей является неравномерной. Засечка вероятности осуществляется следующим образом. Берется вероятность p = 0,01. По таблицам нормальной функции распределения находится значение $V = \Phi^{-1}(p)$. Напротив полученного значения V ставится засечка 0,01 на шкале вероятностей. Далее берется вероятность p = 0,05 и т. д.

Этап 2. Исходная выборка значений упорядочивается по возрастанию. В результате получается последовательность $x^{(1)}$,..., $x^{(N)}$.

Этап 3. Для каждого значения $x^{(i)}$, $i = \overline{1, N}$ на плоскости $(t, \hat{F}(t))$ отмечается точка ($x^{(i)}$, i/N). Для того, чтобы определить расположение этой точки, находится значение $V_i = \Phi^{-1}(i / N)$, которое откладывается по оси V .

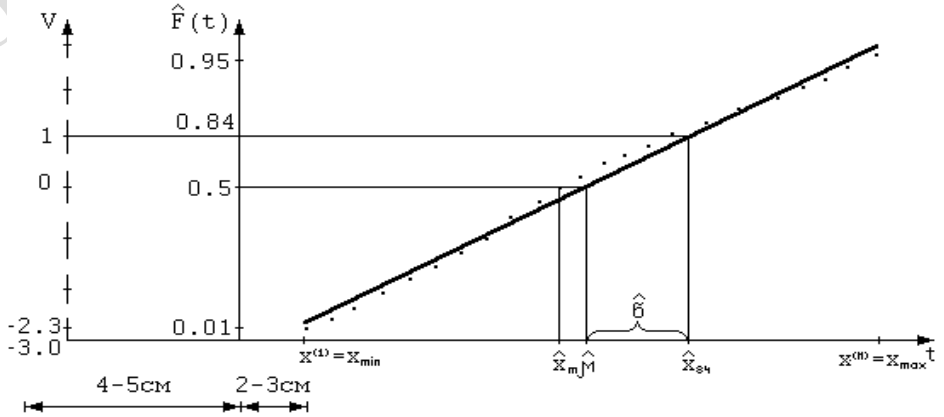

Рисунок 3.1 – График нормальной вероятностной бумаги

Этап 4. Если точки ($x^{(i)}$, *i/N*) в какой-то мере ложатся вдоль некоторой прямой, то можно предположить, что генеральная совокупность, из которой извлечена данная выборка, имеет нормальный закон распределения. В противном случае надо подыскать преобразование переменной, например, логарифмирование, извлечение корня и т. п., в результате которого выборка бы соответствовала нормальному распределению.

Этап 5. В случае принятия гипотезы о нормальности распределения осуществляется оценка параметров распределения. В качестве оценки центра $\hat{\mu}$ берется медиана выборки, которая соответствует вероятности p = 0,5, т. е. значение $\hat{x}_{50}$, где $\hat{x}_{50}$ – оценка 0,5 квантиля распределения, полученного при *V = 0*. Оценка стандартного отклонения $\hat{\sigma} = \hat{x}_{84} - \hat{\mu}$, где $\hat{x}_{84}$ – оценка 0,84 квантиля распределения, полученного при *V* =1.

Вопросы для самоконтроля

1. Приведите определение функции распределения.
2. Что такое геометрическая интерпретация функции распределения?
3. Каковы свойства функции распределения?
4. Как определяется квантиль уровня q?
5. Что такое геометрическая интерпретация квантиля уровня q (q = 0,1; 0,2 и т. д.) на графиках функции распределения и функции плотности распределения?
6. Что такое стандартная функция нормального распределения?
7. Что такое правило 3σ?
8. Что такое нормальная вероятностная бумага и как она строится?

Лабораторная работа

***Цель работы*:** получение практических навыков построения параметрической оценки функции плотности распределения вероятности с помощью нормальной вероятностной бумаги.

***Материалы и оборудование*:** персональный компьютер.

Задание. Для исходных данных в соответствии с вариантом построить параметрическую оценку функции плотности распределения вероятности с помощью нормальной вероятностной бумаги. Для этого проверить с помощью вероятностной бумаги гипотезу H_0 о том, что статистические данные $x_1,...,x_N$ – случайная выборка из нормального распределения (с параметрами $\hat{\mu}$ и $\hat{\sigma}$). Если гипотеза H_0 не верна, то воспользоваться преобразованием $y_i = ln(x_i)$ и для полученной выборки $y_1,...,y_N$ проверить с помощью вероятностной бумаги гипотезу H_0 о том, что статистические данные

$y_1,...,y_N$ – случайная выборка из нормального распределения (с параметрами $\hat{\mu}$ и $\hat{\sigma}$), т. е. статистические данные $x_1,...,x_N$ – случайная выборка из логнормального распределения. Оценить значения $\hat{\mu}$ и $\hat{\sigma}$.

Рекомендации по выполнению работы

Для построения графика, изображенного на рисунке 3.1, использовать по оси *x* значения исходной выборки, а по оси *y* полученные значения для стандартного нормального распределения.

Для выполнения задания при работе в среде пакета Excel использовать функцию НОРМСТОБР(.), которая возвращает обратное значение аргумента стандартного нормального распределения, при котором оно равно, например, значению 0,1: НОРМСТОБР(0,1) = –1,28155. Оценки значений вероятности посчитать по формуле (3.1).

Оценку $\hat{\mu}$ получить на пересечении графика тренда построенной кривой с осью абсцисс, а оценку $\hat{\sigma}$ получить как длину интервала от значения оценки $\hat{\mu}$ до значения аргумента, при котором значение соответствующего тренда равно 1.

Тема 4. Параметрическая статистическая аппроксимация закона распределения

Построение гистограммы, полигона частот и ядерной аппроксимации, рассмотренных в теме 2, основано на локальной интерполяции. Другой подход к аппроксимации заключается в интерполировании закона распределения на всем интервале [*a,b*]. К методам этого типа относится также аппроксимация с помощью системы кривых Пирсона [2]. Систему кривых Пирсона получают путем выравнивания дискретного гипергеометрического распределения непрерывной кривой. При этом для выбора подходящей кривой используют четыре первых момента выборочного распределения. Отметим, что практического распространения данный способ аппроксимации распределения не получил в связи с неустойчивостью моментов первого порядка и невозможностью интерпретации механизма генерации выборки полученного типа распределения.

Более популярными среди интегральных методов аппроксимации оказались параметрические методы оценки распределения путем проверки на согласие данного эмпирического распределения с конкретным теоретическим распределением, например, нормальным, экспоненциальным и т. д.

В реальной ситуации тип распределения часто бывает известен. Кроме того, просмотрев гистограмму или полигон частот, пользователь для себя уже принимает общепризнанную гипотезу H_0 о типе распределения (или, наоборот, отвергает ее из-за сильного засорения выборки, смешения в ней двух или более подвыборок из разных генеральных совокупностей). Математический аппарат в виде критерия согласия используется здесь с целью подтверждения и оформления решения пользователя.

Воспользуемся критерием согласия χ^2. Процедура проверки гипотезы H_0 в данном случае будет состоять из следующих этапов.

Алгоритм проверки гипотезы о типе распределения с помощью критерия согласия χ^2.

Этап 1. Область изменения выборки [*a*,*b*] делим на *s* интервалов, как при построении гистограммы (можно воспользоваться равноинтервальной и равнонаполненной). Если в каком-то интервале частота m_s слишком мала (меньше 5), то этот интервал объединяется с соседним интервалом. Таким образом, количество интервалов может уменьшиться и стать равным *s*′.

Этап 2. По выборке вычисляют оценки параметров теоретического распределения (тем самым теоретическое распределение будет полностью определено). Теперь по теоретическому распределению вычислим вероятности p_s того, что случайная величина *X* принимает значение из *s*-го интервала, при этом $\sum_{s=1}^{s'} p_s = 1$. Затем найдем теоретические частоты $n_s = N \cdot p_s$.

Этап 3. Гипотеза H_0 верна, если теоретические и эмпирические частоты n_s и m_s достаточно мало отличаются друг от друга. Для проверки гипотезы H_0 используем следующую статистику:

$$Q^2 = \sum_{s=1}^{s'} \frac{(m_s - n_s)^2}{n_s}. \qquad (4.1)$$

Этап 4. Случайная величина Q^2 имеет $\chi^2(\nu)$ распределение с числом степеней свободы $\nu = s' - r - 1$, где *s*′ – количество интервалов, *r* – количество параметров теоретического распределения, оценки которого вычислялись по выборке. Чем больше Q^2, тем хуже согласованы теоретическое и эмпирическое распределения. При достаточно большом значении Q^2 нужно отвергнуть гипотезу H_0. Поэтому используем только правостороннюю критическую область. *P*-значением является площадь области под функцией плотности распределения $\chi^2(\nu)$ справа от точки Q^2. Если $P < a$, то мы отвергаем H_0 и принимаем гипотезу H_1: теоретическое и эмпирическое распределения не согласованы. Здесь *a* – это уровень значимости, который обычно принимается равным 0,05.

Вопросы для самоконтроля

1. Что такое локальная интерполяция закона распределения?
2. Приведите интегральные методы аппроксимации закона распределения.
3. Что такое параметрическая аппроксимация закона распределения?
4. Что такое *P*-значение для критерия χ^2?
5. Как используют *P*-значение при принятии решения о согласии или несогласии эмпирического и теоретического распределений?

Лабораторная работа

***Цель работы*:** получение практических навыков проверки статистических данных на нормальность по критерию согласия χ^2.

***Материалы и оборудование*:** персональный компьютер.

Задание. Для исходных данных в соответствии с вариантом проверить статистические данные на нормальность с помощью критерия согласия χ^2.

Рекомендации по выполнению работы

Для выполнения работы воспользоваться результатами, полученными при построении непараметрической аппроксимации ф. п. р. в. в виде равнонаполненной гистограммы (20 %). С помощью функции плотности нормального распределения посчитать соответствующие теоретические частоты вероятности попадания в интервалы, задающие ширину 20 %-ных диапазонов гистограммы.

Определить значение Q^2 по формуле (4.1).

Для выполнения задания при работе в среде пакета Excel использовать статистическую функцию НОРМРАСП(A1;B1;C1;ИСТИНА). Она позволяет определить значение функции нормального распределения $N(x, \bar{x}, \sigma)$ в точке x, значение x содержится в ячейке A1 с математическим ожиданием $\bar{x}$, значение $\bar{x}$ содержится в ячейке B1, и средним квадратическим отклонением σ, значение σ содержится в ячейке C1.

Для определения p_s нужно воспользоваться соотношением:

$$p_s = P(x_{s-1} < X < x_s) = N(x_s, \bar{x}, \sigma) - N(x_{s-1}, \bar{x}, \sigma),$$

где x_{s-1} , x_s – начало и конец s-го интервала, соответственно.

Замечание. Для первого интервала $p_1 = P(-\infty < X < x_2) = N(x_2, \bar{x}, \sigma)$. Для последнего интервала $p_{s'} = P(x_{s'-1} < X < \infty) = 1 - N(x_{s'-1}, \bar{x}, \sigma)$.

С помощью $\chi^2(\nu)$ распределения с s' числом степеней свободы определить для значения Q^2 величину P. Проверить $P < a$ или $P > a$ и принять решение о «согласии» теоретического и эмпирического распределений.

Заметим, что P-значением является площадь области под функцией плотности распределения $\chi^2(\nu)$ справа от точки Q^2. Для определения P-значения нужно воспользоваться статистической функцией ХИ2РАСП(Q^2; k) с k степенями свободы.

Тема 5. Предварительный статистический анализ

5.1. Этап предварительного статистического анализа исследуемой системы

Общая логическая схема статистического анализа данных может быть представлена в виде семи этапов, реализуемых в режиме итерационного взаимодействия). Первый этап посвящен исходному предварительному анализу исследуемой системы. На этом этапе определяются: основные цели исследования на неформализованном, содержательном уровне; совокупность единиц (объектов), представляющая предмет статистического исследования; набор параметров-признаков для описания обследуемых объектов; степень формализации соответствующих записей при сборе данных (см. таблицу 5.1); время и трудозатраты, объем работ; выделение ситуаций, требующих предварительной проверки перед составлением детального плана исследований; формализованная постановка задачи; в каком виде осуществляется сбор первичной информации и введение в ЭВМ.

5.2. Семантическое моделирование в базах данных

Основные подходы к моделированию в базах данных. Первоначально в теории БД основное внимание уделялось средствам эффективной организации данных и манипулирования ими. В результате возникли три основные модели данных: иерархическая, реляционная и сетевая. При этом явно или неявно предполагалось, что предложенные средства достаточно универсальны для представления знаний или информации о любых предметных областях. Так, и сегодня приверженцы получившей наибольшее распространение реляционной модели зачастую утверждают, будто табличная форма представления данных является наиболее удобной и интуитивно понятной проектировщику.

Таблица 5.1 – Основные типы шкал измерения признаков и допустимые числовые операции с ними

Типы шкалы	Наименование шкалы	Множество допустимых преобразований $f(x)$	Отношения, отвечающие шкале	Допустимые числовые операции с измерениями	Примеры измерения
Качественные	Наименований (номинальная, классификационная)	Взаимно-однозначные преобразования	Эквивалентность	Сравнения: $x = y$, $x <> y$	Национальность, пол, профессия, вид оплаты труда
	Порядковая (ранговая, ординальная)	Монотонно-неубывающие функции	Квазипорядок (нестрогая ранжировка)	Сравнения: $x \le y$	В строгом смысле примеров шкалы нет. Условно: шкала твердости минералов, экспертные ранжировки, оценки предпочтений
Количественные	Разностей (балльная)	$F(x) = d+x$	Аддитивное метризованное	Сравнения: $x - y \le z - v$; $x + y$, $x - y$	Квалификационные разряды, балльные оценки
	Интервалов (интервальная)	$F(x) = d + kx$, $k > 0$	4-арное мультипликативное метризованное	$(x - y)/(z - v)$, $x + y$, $x - y$	Любые показатели, значения которых могут быть отрицательными: температура по Цельсию, летоисчисление, прибыль (при наличии убытков), высота над уровнем моря
	Отношений (относительная)	$F(x) = kx$, $k > 0$	Мультипликативное метризованное	x/y, xy, $x + y$, $x - y$	Температура по Кельвину, возраст, производительность труда

Однако проектирование базы данных в терминах этих моделей часто сводится к очень сложному и неудобному для проектировщика процессу, поскольку эти модели не содержат достаточных средств представления смысла данных. Семантика реальной предметной области должна независимым от модели способом представляться в сознании проектировщика. Такое положение вещей приводит к замедлению процесса разработки БД и является источником потенциальных ошибок.

По этой причине в последние годы получило развитие направление, являвшееся предметом активных исследований в конце 70-х – начале 80-х годов, – семантическое, или концептуальное, моделирование в базах данных. Его основная цель – организация интерфейса проектировщика, а также конечного пользователя с информационной системой на уровне представлений о предметной области (ПО), а не на уровне структур данных. Интерес к этому направлению возрос в связи с развитием средств автоматизированного проектирования БД на основе CASE-технологий.

В настоящее время определился основной подход к решению задач семантического моделирования в базах данных. Он заключается в выделении двух уровней моделирования: уровня концептуального моделирования ПО и уровня моделирования собственно базы данных.

На верхнем уровне осуществляется переход от неформализованного описания ПО и информационных потребностей конечного пользователя к их формальному выражению с помощью специальных языковых средств. На нижнем – преобразование концептуальной модели ПО в схему БД и нормализация схемы БД.

Предметная область и семантика предметной области. Понятие «предметная область» является базисным понятием в теории БД и поэтому не имеет строгого определения. Чтобы выяснить его смысл, обратимся к понятиям объект и предмет.

Объект – то, что существует вне нас и независимо от нашего сознания, явление внешнего мира, материальной действительности.

Объекты потенциально обладают огромным количеством свойств и находятся в потенциально бесконечном числе взаимосвязей друг с другом. Однако среди всего множества свойств и взаимосвязей между объектами имеет смысл выделять лишь существенные, важные с точки зрения потребителя информации.

Предмет – объект, ставший носителем определенной совокупности свойств и входящий в различные взаимоотношения, которые представляют интерес для потребителей информации. Один и тот же объект может восприниматься разными системами как разные предметы. Таким образом, предмет – это модель реального объекта.

Совокупность объектов, информация о которых представляет интерес для пользователей, образует объектное ядро предметной области.

Понятие «предметная область» соответствует точке зрения потребителей информации на объектное ядро, при которой выделяются только те свойства объектов и взаимосвязи между ними, которые представляют определенную прагматическую ценность и должны фиксироваться в базе данных. Таким образом, предметная область представляет собой абстрактную картину реальной действительности, определенная часть которой фиксируется в качестве модели фрагмента действительности.

В каждый момент времени ПО находится в одном из состояний, которое характеризуется совокупностью объектов и их взаимосвязей. Если объекты образуют объектное ядро, то совокупность взаимосвязей отражает структуру фрагмента действительности. С течением времени одни объекты исчезают, другие появляются, меняются свойства и взаимосвязи. Тем не менее возникающие новые состояния считаются состояниями одной и той же ПО. Таким образом, ПО целесообразно рассматривать как систему, переживающую свою историю, которая состоит из определенной последовательности состояний.

Введя пространство состояний, можно рассматривать в нем определенные траектории или последовательности состояний $S_0,S_1,...,S_t$, в которых находится ПО в моменты времени $0, 1,..., t$. Члены такой последовательности не могут быть совершенно произвольными, поскольку состояние S_t обычно каким-либо образом связано с предшествующими состояниями $S_0,S_1,...,S_{t-1}$. Поэтому ПО можно определить как класс всех действительно возможных последовательностей состояний. Такие последовательности называются траекториями ПО. Совокупность всех общих свойств траекторий называется семантикой предметной области.

Вопросы для самоконтроля

1. Назовите типы шкал измерения и примеры.
2. Каково множество допустимых преобразований для шкальных измерений?
3. Какие числовые операции допустимы с шкальными измерениями?
4. Приведите математический инструментарий статистического исследования зависимостей.
5. Поясните на примерах понятия предмета и объекта.
6. Что такое предметная область?

Лабораторная работа

***Цель работы*:** получение практических навыков постановки задачи по предварительной статистической обработке данных.

***Материалы и оборудование*:** персональный компьютер.

Задание. Для исходных данных в соответствии с вариантом для произвольной предметной области (приведенных в таблице 5.2) выполнить первый этап статистической обработки: предварительный статистический анализ данных.

Таблица 5.2 – Предполагаемый список предметных областей в соответствии с вариантом задания

Вариант	Предметная область
1	маркетинг
2	здравоохранение
3	производство
4	сбыт
5	Сельское хозяйство
6	Транспорт
7	Педагогика
8	Профилактика заболеваний
9	Военная промышленность
19	Судостроение
11	Растениеводство
12	Животноводство

Рекомендации по выполнению работы

В отчете представить следующие моменты:

– предметная область;

– цель исследования (ожидаемый результат);

– задача (фрагмент работы, позволяющий приблизиться к цели);

– объект (явление, процесс). Заметим, что объектом может быть, с одной стороны, например, клетка, ткань или отдел сердца, или, если мы изучаем причины разводов, то это может быть человек, семья, организация или в другой задаче статистической обработки, это уже может быть государство и т.п. С другой стороны, это может быть процесс или явление, в котором перечисленные выше объекты участвуют, например работа системы кровообращения отдельного человека и т. д.;

– признаки для описания объекта (их описание), шкалы измерения признаков (см. таблицу 5.3). Привести примеры всех наименований шкал,

перечисленных в таблице 5.1 (т. е. шкалы наименований; порядковой; разностей; интервалов и отношений):

Таблица 5.3 – Признаки для описания объекта

Наименование признака	Шкала	Единица измерения	Примеры
Вес	Количественная (относительная)	кг	1 кг

– содержательная постановка задачи. Уточнить способы и сроки сбора данных, ожидаемый результат, время и трудозатраты, объем работ; выделение ситуаций, требующих предварительной проверки перед составлением детального плана исследований; в каком виде осуществляется сбор первичной информации т. п.;

– математическая или формализованная постановка задачи и предполагаемый математический инструментарий исследования (статистическое описание: оценка функции плотности распределения вероятности, кластерный анализ; статистическое предсказание: корреляционный, регрессионный, дисперсионный и дискриминантный анализы). Прокомментировать краткую суть используемого алгоритма статистического описания или предсказания.

Тема 6. Восстановление пропущенных значений

Непараметрический подход к оценке пропусков в данных

Одной из важнейших задач первичной статистической обработки данных является восстановление пропущенных значений наблюдений в матрице исходных данных. Если объемы выборок не малы, то иногда удаляют все имеющиеся остальные значения признаков по этому объекту, и размерность матрицы уменьшается на число удаленных объектов, для которых хотя бы один признак оказался отсутствующим.

Иногда заменяют пропущенное измерение по признаку для некоторого объекта средним значением по всем имеющимся измерениям этого свойства. Такой способ используется, в частности, системой Statistica. Этот способ восстановления пропусков в данных далек от оптимального. Опишем ниже более приемлемый подход к восстановлению пропусков.

Непараметрический подход [2] к оценке пропусков в матрице данных. Наряду с подходом, требующим аналитического задания закона

распределения, существует и другой, основанный на использовании расстояния между параметрами объектов (в некоторой метрике), определяемого по значениям признаков, измеренных у обоих объектов. Постулируется, что, если два объекта близки в пространстве измеренных признаков, то они должны быть близки и в пространстве по неизмеренным признакам. Метрика и пороговое значение расстояния, определяющие близость объектов, вводятся в зависимости от условий задачи (шкалы, количества признаков).

Одним из локальных алгоритмов, предназначенных для заполнения пробелов, является *алгоритм ZET*. В основе алгоритма ZET лежат три предположения. Первое предположение (гипотеза избыточности) состоит в том, что реальные таблицы имеют избыточность, проявляющуюся в наличии похожих между собой объектов (строк) и зависящих друг от друга свойств (столбцов). Если избыточность отсутствует (как, например, в таблице случайных чисел), то предпочесть один прогноз другому невозможно.

Второе предположение (гипотеза аналогичности) состоит в утверждении, что, если некоторая пара объектов близка по значениям *(n-1)*-го свойств, то она близка и по *n*-му свойству.

Третье предположение (гипотеза локальной компетентности) заключается в том, что избыточность носит локальный характер: у каждого объекта есть свое подмножество объектов-аналогов и у каждого свойства есть свое подмножество свойств-аналогов. Если это так, то не имеет смысла привлекать к предсказанию значения некоторого элемента $x_i^{(j)}$ информацию, содержащуюся в строках, не похожих на *i*-ю строку, и в столбцах, не похожих на *j*-й столбец. В предсказаниях должны участвовать только так называемые «компетентные» строки и столбцы, которые выбираются для каждого предсказываемого элемента отдельно.

Рассмотрим схематично конкретизацию непараметрического подхода в алгоритме ZET [2]. Пусть у объекта X_i требуется оценить значение пропущенного признака $x^{(j)}$, т. е. оценить $x_i^{(j)}$ в матрице X. Для этого в X выделяется подмножество объектов, у которых измерено значение *j*-го признака. В этом подпространстве выделяется однородная группа объектов, наиболее близких к X_i в подпространстве признаков, полученном из исходного пространства исключением *j*-го признака. Неизмеренное значение $x_i^{(j)}$ заменяется средним значением по *j*-му признаку в выделенной группе объектов. Для оценки качества заполнения пропусков ввести формализованный критерий трудно. Приближенно его оценивают, например, так: из матрицы X случайным образом исключается часть измеренных значений, затем исключенные пропуски восстанавливаются. Мера качества заполнения определяется с помощью оценки удаления истинных значений от полученных с помощью непараметрического подхода.

Вопросы и задания для самоконтроля

1. Почему появляются пропуски в матрице исходных данных?
2. В чем суть непараметрического подхода восстановления пропущенных значений?
3. В чем суть метода ZET восстановления пропущенных значений?
4. Приведите примеры метрик.

Лабораторная работа

***Цель работы*:** изучить подходы к восстановлению пропущенных значений в исходных данных.

***Материалы и оборудование*:** персональный компьютер.

Задание 1. Разработать алгоритмическую схему восстановления пропущенных значений методом ZET.

Задание 2. Для исходных данных в соответствии с вариантом восстановить пропущенные значения методом ZET в матрице данных.

Задание 3. Оценить качество заполнения пропусков.

Рекомендации по выполнению работы

В отчете представить следующие моменты:
- алгоритм восстановления пропущенных значений методом ZET;
- матрица данных без пропусков;
- разработанный критерий качества заполнения пропусков;
- оценка качества заполнения пропусков.

Литература

1. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян [и др.] – М.: Финансы и статистика. 1989. – 605 с.

2. Айвазян, С. А. Прикладная статистика: основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика. 1983. – 472 с.

3. Айвазян, С. А. Прикладная статистика: исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика. 1985. – 488 с.

4. Афифи, А. Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзен. – М.: Мир, 1982. – 405 с.

5. Гаскаров, Д. В. Малая выборка / Д. В. Гаскаров, В. И. Шаповалов. – М: Статистика, 1978.

6. Мандель, И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статистика, 1988. – 172 с.

7. Осипенко, Н. Б. Пакеты анализа и обработки данных: тексты лекций / Н. Б. Осипенко. – Гомель: ГГУ им. Ф. Скорины, 2008. – 99 с.

8. Максимей, И. В. Обработка экспериментальных данных на ЭВМ: пособие для лабораторных занятий по спецкурсу / И. В. Максимей, Н. Б. Осипенко, А. Н. Осипенко. – Гомель: ГГУ, 1999. – 54 с.

9. Осипенко, Н. Б. Пакеты статистической обработки экспериментальных данных: тексты лекций / Н. Б. Осипенко. – Гомель: ГГУ им. Ф. Скорины, 2007. – 78 с.

10. Осипенко, Н. Б. Справочное пособие по первичной статистической обработке и исследованию зависимостей / Н. Б. Осипенко, А. Н. Осипенко. – Гомель: – ГГУ, 2000. – 85 с.

Производственно-практическое издание

Осипенко Наталья Борисовна,
Осипенко Александр Николаевич

ПРОГРАММНЫЕ СРЕДСТВА ПЕРВИЧНОЙ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Практическое руководство

для студентов специальности I–40 01 01
«Программное обеспечение информационных технологий»

Редактор *В. И. Шкредова*
Корректор *В. В. Калугина*

Подписано в печать 23.04.2015. Формат 60x84 1/16.
Бумага офсетная. Ризография. Усл. печ. л. 2,6.
Уч.-изд. л. 2,8. Тираж 25 экз. Заказ 235.

Издатель и полиграфическое исполнение:
учреждение образования
«Гомельский государственный университет
имени Франциска Скорины».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/87 от 18.11.2013.
Специальное разрешение (лицензия) № 02330 / 450 от 18.12.2013.
Ул. Советская, 104, 246019, Гомель.