Министерство образования Республики Беларусь

Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Н. Б. ОСИПЕНКО

ПРОГРАММНЫЕ СРЕДСТВА ПЕРВИЧНОЙ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

ТЕКСТЫ ЛЕКЦИЙ

для студентов
специальности 1-31 03 03 Прикладная математика (по направлениям)
(1-31 03 03-01 научно-производственная деятельность)

Гомель 2011

СОДЕРЖАНИЕ

РАЗДЕЛ 1 ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ

Тема 1 Базовые понятия анализа данных

1.1Этапы работ, предшествующие обработке экспериментальных данных
1.2 Прикладная статистика
1.3 Идеи и методологические принципы многомерного статистического анализа данных
1.4 Цели эксперимента в науке и промышленности

1.1 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.

1.2 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 выше сказанного, дадим определение, введенное в 1983г. 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.

1.3 Идеи и методологические принципы многомерного статистического анализа данных

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

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

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

Оптимизационная формулировка задач многомерного статистического анализа данных.

1.4 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-ю или 32-мя 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.

Тема 2 Подходы к статистическому анализу данных

2.1 Возможные подходы к статистическому анализу данных
2.2 Типы реальных ситуаций с позиции выполнения требований статистического ансамбля
2.3 Примеры подходов к статистическому анализу данных
2.4 Сравнение подходов к статистическому анализу данных

2.1 Возможные подходы к статистическому анализу 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.

2.2 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 условия эксперимента); с недопустимостью вероятностно-статистических приложений (в этом случае идея многократного повторения одного и того же эксперимента в неизменных условиях является бессодержательной).

2.3 Примеры подходов к статистическому анализу данных

Пример 1. Исследуется массовое производство. Контролируется брак на изделиях. Результаты фиксируются в выборке:

$$x_1,\ldots,x_N \qquad (2.1)$$

где $X_i=1$, если изделие дефектно, а иначе – $X_i=0$. Если производство отлажено и действует в стационарном режиме, то ряд наблюдений (2.1) естественно интерпретировать как ограниченную выборку из соответствующей бесконечной (генеральной) совокупности, которую бы мы имели, если бы осуществляли сплошной контроль изделий. В подобных ситуациях имеется принципиальная возможность многократного повторения наблюдения в рамках одинаковых условий. Такие ситуации могут быть описаны вероятностными моделями. Ряд (2.1) интерпретируется как случайная выборка из генеральной совокупности, т.е. как экспериментальные значения анализируемой случайной величины. Заметим, что в теории вероятностей под случайным явлением понимают явление, относящееся к классу повторяемых, обладающих свойством статистической устойчивости при повторении однородных опытов. Здесь для статистической обработки применяются классические математико-статистические методы. Если основные свойства и характеристики генеральной совокупности не известны исследователю, то они оцениваются по соответствующим свойствам и характеристикам выборок с помощью этих методов.

Пример 2. Исследуется совокупность средних городов России (с численностью [100; 500] тысяч человек) для выяснения типов городов, сходных или однородных по структуре уровня образования жителей, половозрастному составу и характеру занятости. Подробный анализ большого числа городов практически не реален, поэтому в фиксированном пространстве небольшого числа интегральных параметров города разделяются на типы, выделяются эталоны, а для них проводят подробный анализ с целью выявления наиболее характерных черт и закономерностей в социально-экономическом облике средних по величине типичных городов.

Так для N средних городов (например, для России их оказалось 74) $X_1 \ldots X_N$ были зарегистрированы 32 параметра

$$\begin{pmatrix} X_1 \\ \cdots \\ X_n \end{pmatrix} = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(2)} \\ \cdots & \cdots & \cdots \\ x_n^{(1)} & \cdots & x_n^{(2)} \end{bmatrix},$$

где $x^{\text{С}}$ - параметры, характеризующие среднее число жителей, приходящихся на 1000 человек населения города. Причем x^i, i=1…4 – параметры, характеризующие уровень образования (высшее, незаконченное высшее, среднее специальное, среднее); x^i, i=5…16 – 12 параметров, характеризующих половозрастной состав; x^i, i=17…21 – 5 параметров для описания социального характера занятости населения; x^i, i=22…32 – параметры, характеризующие занятость в материальном или нематериальном производстве и источники доходов.

Если допустить, что геометрическая близость двух точек – городов X_i и X_j в соответствующем 32-мерном пространстве означает их однородность (сходство) по анализируемым признакам и является основанием для их отнесения к одному типу, то для решения задачи надо привлечь методы кластер-анализа и снижения размерности. Математический аппарат этих методов предполагает вычисление средних, дисперсий, ковариаций, но эти характеристики описывают уже природу и структуру только реально анализируемых данных, т.е. статистически обследованную совокупность из n анализируемых городов.

2.4 Сравнение подходов к статистическому анализу данных

Основные отличительные особенности подходов на примере задачи классификации представим схематично в таблице 2.1.

Таблица 2.1– Отличительные особенности подходов

Составляющие	Первое направление	Второе направление
Цели исследования	Выделение классов, как инвариантов в потоке выборочных объектов	Выяснение распределения данных в системе
Объекты и признаки.	Независимы	Зависимость предполагается, ее нужно обнаружить
Выделяемые классы	Характеризуются эталоном и не пересекаются	Четко не выделяются, т.е. пересекаются
Аппарат исследования	Вероятностный - преобразование пространства признаков (даже в одномерную ось)	Логико-комбинаторный

Первое направление развития анализа данных, ориентированное на технические области знания, отстаивает идею простоты используемых моделей. В рамках этого направления неудовлетворительные результаты объясняют отсутствием информативных признаков.

Второе направление развития анализа данных ориентировано на социально-экономическую и социологическую информацию. При ее обработке появилось много новых идей, в частности, идея поэтапной группировки и коллектива решающих правил. Разработаны методы многомерного шкалирования, экспертных оценок.

В отличие от первого примера во втором примере невозможно: интерпретировать исходные данные в качестве случайной выборки генеральной совокупности (в связи с неприятием главной идеи понятия статистического ансамбля: идея многократного повторения одного и того же эксперимента в неизменных условиях теряет смысл); использовать вероятностную модель для построения и выбора наилучших методов статистической обработки; дать вероятностную интерпретацию выводам, основанным на статистическом анализе исходных данных.

Но в обоих случаях выбор наилучшего из всех возможных методов обработки данных производится в соответствии с некоторыми функционалами качества метода. Способ обоснования выбора этого функционала, а также его интерпретация различны. В первом случае выбор основан на допущении о вероятностной природе исходных данных и интерпретация тоже. Во втором случае исследователь не пользуется априорными сведениями о вероятностной природе исходных данных и при обосновании выбора оптимального критерия качества опирается на соображения содержательного (физического) плана - как именно и для чего получены данные. Когда критерий выбран, в обоих случаях используются методы решения экстремальных задач. На этапе осмысления и интерпретации каждый из подходов имеет свою специфику.

При выборе типа модели следует понимать, что всякая модель является упрощенным (математическим) представлением изучаемой действительности. Мера адекватности модели и действительности является решающим фактором работоспособности используемых затем методов обработки. А так как ни одна модель не может идеально соответствовать реальной ситуации, то желательна многократная обработка исходных данных для разных вариантов модели.

Тема 3 Общая характеристика пакетов прикладных программ статистической обработки данных

3.1 Классификация программного инструментария статистической обработки данных

3.2 Интерактивные средства программного обеспечения прикладной статистики

3.3 Формы программного обеспечения прикладной статистики

3.4 Пакеты статистических программ для ПЭВМ

3.1 Классификация программного инструментария статистической обработки данных

ГОСТ 19.101–77 ЕСПД подразделяет программы на следующие виды: компонент и комплекс. *Компонент* – программа, рассматриваемая как единое целое, выполняющая законченную функцию и применяемая самостоятельно или в составе комплекса. *Комплекс* – программа, состоящая из двух или более компонентов, выполняющих взаимосвязанные функции, и применяемая самостоятельно или в составе другого комплекса.

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

Программное обеспечение классифицируется

1 по реализуемым функциям;

2 по типу ОС, под управлением которой оно работает;

3 по способу управления.

По **функциям** ПС делятся на: обеспечивающие заданный режим обработки заданий ЭВМ, расширяющие возможности ОС (специфична для АСУ) и обеспечивающие решение задач пользователя (пакеты общего назначения). Классификация по **типу ОС** определяет возможность применения пакета в конкретной операционной обстановке.

По **способу управления** делятся на ПС с простой структурой и сложной обработкой.

ПС простой структуры – это набор модулей, обеспечивающих решение различных задач из предметной области, на которую ориентирован пакет. Существует две группы ПС **простой структуры**.

1. ПС, расширяющие библиотеки. Обращение к модулям осуществляется из прикладной программы пользователя, информационное содержание включённых в программу модулей друг с другом и с включающей программой происходит на уровне ОП. Ввод результатов, использование внешней памяти в качестве буфера обычно возложено на программиста. Обращение к модулям пакета – на алгоритмическом языке.

2. ПС с автономными программами. Управление пакетом – на языке управления заданиями (ЯУЗ) операционной системы, используются простейшие средства описания входных данных (описание форматов). Информационное сопряжение модулей – на уровне внешней памяти

Существует две группы ПС **сложной структуры.**

1. ПС с произвольной последовательностью обращения к модулям. Функции управления пакетом сосредоточены в специальных модулях, образующих управляющую программу. Межмодульный интерфейс организуется на уровне ОП и ВП. Порядок модулей и их информационное сопряжение осуществляет пользователь с помощью входного языка. Пакет не контролирует заданную последовательность обращений к обрабатывающим модулям.

2. ПС с фиксированной последовательностью обращения к модулям. Последовательность обращений к модулям фиксируется в виде графа предметной области.

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

Изучим в связи с этим изменяющиеся формы ПС, тенденции развития ПС и развитие каких форм ПС наиболее перспективное при проведении статистического исследования.

При всём богатстве ПС этапы решения статистических задач остаются прежними – построение статистической модели, соответствующей задаче, выбор статистического метода и (вместе с программой) соответствующего ПС, затем – реализация этого метода в конкретном ПС для решения задачи.

Но получение содержательно интерпретируемого результата не всегда означает окончание исследования. При использовании отобранных методов и средств для анализа информации может быть необходима непрерывная работа в течении нескольких лет. Поэтому выделим постоянно поддерживаемые (исследовательской группой) задачи и назовём эксплуатационными в отличие от хорошо изученных исследовательских задач. Использование ПС планируется на всех этапах исследования, начиная от организации хранения данных и кончая обеспечением многократной прогонки заданий с различными параметрами

Критерии оценки ПС

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 должны быть описаны со ссылкой на литературу и указанием значений стандартных значений параметров, задаваемых по умолчанию. Должно быть указание о действиях с отсутствующими значениями. Вывод результатов должен быть полным, компактным, неизбыточным и содержать средства подавления лишней информации и запроса дополнительного. Должна быть возможность графического вывода (гистограмма, вероятностные графики и т.д.). Должны быть надписи на графиках и возможность использовать разные шкалы. Должен быть алгоритм для определения стоимости и времени выполнения заданий. Язык управления заданиями должен быть 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 ПС желательно иметь листинги программ на исходном языке, как первичную документацию пакета. Пакет должен обладать способностью расширения (включения других программ) и легкого переноса на другие ЭВМ. Язык реализации пакета должен быть языком высокого уровня.

Интерактивные средства программного обеспечения прикладной статистики

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

Пути создания диалоговых средств

1. Диалоговый интерфейс к универсальным ПС. Цель диалога – получение от пользователя числовых параметров предложений входного языка пакета (ЯУП). Но такой диалог не очень эффективен с точки зрения пользователя (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 ПС. Некоторые программы наиболее эффективны при многократной прогонке, например, при подборе параметров; получении некоторого предвидимого результата; при графическом представлении исходных данных и результатов.

3. Интерактивные ПС. Некоторые ПС разработаны как интерактивные средства анализа данных. Так в ПС 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, ЭС CLAVESIN (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.

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

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

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

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

4)Первичная статистическая обработка. Сопоставляются постановка задачи и способы ее решения с возможностями ЭВМ и доступным 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.

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

6)Вычислительная реализация статистической обработки данных. Формируются типовые задания для ПС, отражающие особенности обработки и хранения данных. Наиболее простой для оптимизации 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.

3.3 Формы программного обеспечения прикладной статистики

Использование 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 ПС прикладной статистики: библиотека; пакет; система анализа данных.

Для решения статистической задачи по отработанной методике, период эксплуатации которой несколько лет, эффективны **библиотеки**: наиболее эффективны при проведении исследовательского этапа работ в режиме проверки статистических гипотез. Для задач со сложными способами представления и использования данных, требующих сложных способов организации вычислительного процесса эффективны системы.

Пример библиотеки: пакет научных программ БИМ (Институт математики АН БССР). Решение задач средствами библиотек удовольствие дорогое для исследовательской группы и эффективно при использовании нестандартных методов и режимов работы с ПС 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.

Одним из наиболее распространённых в 80-90-х годах является пакет статистических программ Biomedical Computer Programm, разработанный под руководством Диксона в ВЦ Медицинского центра Калифорнийского университета в Лос-Анжелесе. Первая версия этого пакета BMD появилась в 1961 и быстро развивалась за счёт дополнительных программ, улучшения средств и новых статистических методик. В 1975 г. новая версия – пакет BMDP – фактически заменила предыдущую. Версия 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 ПС принятие решений остается за исследователем. Программа освобождает от рутинной вычислительной работы, но интерпретация полученных результатов зависит от его опыта и знаний. Применение ПС влечет за собой некоторые неудобства: исследователь должен привыкнуть к обозначениям и требованиям 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); ПС базируется на стандартных статистических методах При необходимости использовать нестандартный анализ исследователь должен написать свою программу.

Упомянем также пакеты: пакет программ статистического анализа – ППСА (ЦЭМИ АН СССР), обработка таблиц экспериментальных данных – ОТЭКС (ИМ СО АН СССР).

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.

СОРРА (ИМ АН Лит. ССР) - система оперативной разработки распознающих алгоритмов - - предназначена для решения задач классификации при наличии обучающих выборок, статистического исследования зависимостей (линейная регрессия с возможной проверкой качества методом скользящего экзамена; непараметрическая регрессия;

пошаговая процедура отбора переменных). COPPA управляется с помощью языка пользователя, есть возможность работы в полу диалоговом режиме.

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).

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).

3.4 Пакеты статистических программ для ПК

При статистическом исследовании зависимостей применяются различные ПС. Характеристика разделов прикладной статистики, поддерживаемых в известных ППП приведена в таблице 3.1.

Таблица 3.1 - Статистическое исследование зависимостей в ППП

ППП	а		б						в	г		
	1	2	1	2	3	4	5	6		1	2	3
ПНП-БИМ	+	+	+	-	+	-	+	-	+	-	-	-
ППСА	+	+	+	-	+	+	+	-	-	+	-	+
ОТЭКС	-	-	-	-	-	-	-	+	-	-	-	+
COPPA	-	-	+	-	+	-	-	+	-	-	-	-
P-STAT	+	+	+	-	+	-	+	-	+	+	-	-
SPSS	+	+	+	-	+	-	-	-	+	+	-	-
SAS	+	+	+	-	+	-	+	-	+	+	-	+
BMDP	+	+	+	+	+	-	+	-	+	+	-	-

Комментарии к обозначениям столбцов таблицы 3.1:

а) Корреляционный анализ

1) Вычисление и анализ значимости частных и множественных коэффициентов корреляции;

2) Коэффициенты корреляции Спирмэна, Кендалля;

б) Регрессионный анализ

1) Линейно-множественная регрессия;

2) Отбор переменных в линейной регрессии методом полного перебора или «ветвей границ»;

3) Пошаговые процедуры отбора переменных;

4) Регрессия на главный компоненты;

5) Регрессия нелинейная относительно оцениваемых параметров;

6) Непараметрическая регрессия;

в) Дисперсионный анализ

г) Статистическое исследование зависимостей

1) Стандартный анализ таблиц сопряженности;

2) Логлинейный анализ таблиц сопряженности;

3) Выравнивание шкала (оцифровка, переход к бинарным переменным).

Большинство зарубежных пакетов программ (среди них BMDP, SPSS, P-STAT, SAS) в основном являются приспособлением к возможностям ПК версий пакетов для больших ЭВМ. В частности, по это причине, они в меньшей степени используют возможности ПЭВМ для создания интерактивного взаимодействия с помощью меню и возможности графики. Специально разработанные с учетом возможностей ПЭВМ пакеты, такие как STATA (full), SYSSTAT, STATGRAPHICS, предоставляет меньшие возможности собственно статистической обработки. Статистическое ПО ПЭВМ достаточно развито. Отметим, что наиболее полный набор статических процедур представлен BMDP (full). Самым гибким в отношении управления данными – P-STAT, SAS. STATGRAPHICS – сравнительно медленно работает и его лучше использовать на более мощных ЭВМ.

Выделим наиболее интересные программные средства, используемые при статистическом исследовании и охарактеризуем (см. таблицу 1.2) следующие характеристики: использование ресурсов; управление пакетом и данными; возможности статистической обработки.

1 **Использование ресурсов.**

В графе (таблицы 3.2) RAM (random access memory) приведен минимальный объем внутренней памяти, необходимой для работы пакета.

Таблица 3.2 - Характеристика ПС, используемых при статистическом исследовании

	RAM	Твердый диск	Сопроцессор	Max число		Способ управления	Импорт-экспорт	Манипулирование	Преобразование переменных	Пропуски взвешивание	Документация	Регрессионный анализ	Факторный анализ	Кластерный анализ	Дискриминантный анализ
				Переменных	Объектов										

ABC	256	-1	0	200	Д	М	.	-	.	+	-	.	-	-	-
BMDP/PC(basis)	640	+1-3	+	300	Д	К	+	-	+	.	.	+	-	-	-
BMDP/PC(full)	640	+1-20	+	300	Д	К	+	-	+	.	.	+	+	+	+
NCSS	196	-0,7	0	250	32К	М	.	.	.	.	.	+	.	-	.
PC-ISP	640	. 0,7	0	200	250К	К	+	.	+	.	+	+	+	.	.
P-STAT	640	+2-4	+	150	Д	К	+	+	+	+	.	+	.	-	-
SPSS/PC+(basis)	384	+1-3	0	200	Д	К	+	+	.	+	+	+	-	-	-
SPSS/PC+(full)	450	+2-6	0	200	Д	К	+	+	.	+	+	+	-	-	-
STATA(basis)	256	-0,3	0	RAM	250К	К	.	+	+	+	.	.	-	-	-
STATA(full)	256	-0,4	0	RAM	250К	К	.	+	+	+	.	.	-	-	-
STATGRAPHICS	384	+1	0	RAM	650К	М	.	.	.	.	.	+	.	.	.
SYSSTAT	256	+1-2	0	200	Д	К	+	+	+	.	+	+	.	+	.

В графе «Твердый диск» первый знак означает: «+» – необходимость диска; «.» – желательность; «–» – ненужность; 1-я цифра – min память на диске (Мб); 2-я цифра – max память на диске нужная некоторой программе (Мб).

«Сопроцессор» (Intel 8087): + – нужен; 0 – использование Intel 8087 носит опциональный характер. Сопроцессор ускоряет обработку в 3 раза.

«Мах число объектов»: Д – объектов может быть столько сколько поместится на диске; К – сколько поместится в области памяти такого объема

2 **Управление пакетом и данными**

«Способ управления»: М – меню; К – Команда.

«Импорт - экспорт» {Возможность взаимодействия по данным и др. пакетами (типа LOTUS) и базой данных (типа dbase II/III и др.)}: «+» – хорошо развитый и легко доступный; «.» – Удовлетворительный; «–» – возможность есть, но реализация трудоемка.

«Манипуляция» {Возможности на работе с файлами – слияние, разделение}.

«Пропуски» {Возможность на работе с пропусками, присвоение весов объектов.}

3 **Возможности статистической обработки**.

Возможность статистической обработки {Регрессионный анализ и др.}: «+» – есть; «.» – есть ограниченные возможности; «–» – нет.

Тема 4 Программный инструментарий статистической обработки данных

4.1 Краткая характеристика пакета анализа данных Excel

4.2 Краткая характеристика систем Mathematica, MATLAB и Maple

4.3 Краткая характеристика пакета Statistica

4.1 Краткая характеристика пакета анализа данных Excel

В Microsoft Excel представлено большое число статистических, финансовых и инженерных функций. Некоторые из них являются встроенными, другие доступны только после установки пакета анализа.

Обращение к средствам анализа данных. Средства, которые включены в пакет анализа данных, доступны через команду **Анализ данных** меню **Сервис**. Если этой команды нет в меню, необходимо загрузить надстройку **Пакет анализа**.

Дисперсионный анализ. Существует несколько видов дисперсионного анализа. Требуемый вариант выбирается с учетом числа факторов и имеющихся выборок из генеральной совокупности. **Однофакторный дисперсионный анализ.** Однофакторный дисперсионный анализ используется для проверки гипотезы о сходстве средних значений двух или более выборок, принадлежащих одной и той же генеральной совокупности. Этот метод распространяется также на тесты для двух средних (к которым относится, например, t-критерий). **Двухфакторный дисперсионный анализ с повторениями**. Представляет собой более сложный вариант однофакторного анализа с несколькими выборками для каждой группы данных. **Двухфакторный дисперсионный анализ без повторения.** Представляет собой двухфакторный анализ дисперсии, не включающий более одной выборки на группу. Используется для проверки гипотезы о том, что средние значения двух или нескольких выборок одинаковы (выборки принадлежат одной и той же генеральной совокупности). Этот метод распространяется также на тесты для двух средних, такие как t-критерий.

Корреляционный анализ. Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Коэффициент корреляции выборки представляет отношение ковариации двух наборов данных к произведению их стандартных отклонений. Корреляционный анализ дает возможность установить, ассоциированы ли наборы данных по величине, то есть, большие значения из одного набора данных связаны с большими значениями другого набора (положительная корреляция), или, наоборот, малые значения одного набора связаны с большими значениями другого (отрицательная корреляция), или данные двух диапазонов никак не связаны (нулевая корреляция). Для вычисления коэффициента корреляции между двумя наборами данных на листе используется статистическая функция КОРРЕЛ.

Ковариационный анализ. Ковариация является мерой связи между двумя диапазонами данных. Используется для вычисления среднего произведения отклонений точек данных от относительных средних. Ковариационный анализ дает возможность установить, ассоциированы ли наборы данных по величине, то есть, большие значения из одного

набора данных связаны с большими значениями другого набора (положительная ковариация), или, наоборот, малые значения одного набора связаны с большими значениями другого (отрицательная ковариация), или данные двух диапазонов никак не связаны (ковариация близка к нулю). Вычисления ковариации для отдельной пары данных производятся с помощью статистической функции КОВАР.

Описательная статистика. Это средство анализа служит для создания одномерного статистического отчета, содержащего информацию о центральной тенденции и изменчивости входных данных.

Экспоненциальное сглаживание. Применяется для предсказания значения на основе прогноза для предыдущего периода, скорректированного с учетом погрешностей в этом прогнозе. При анализе используется константа сглаживания *a*, по величине которой определяется степень влияния на прогнозы погрешностей в предыдущем прогнозе. Для константы сглаживания наиболее подходящими являются значения от 0,2 до 0,3. Эти значения показывают, что ошибка текущего прогноза установлена на уровне от 20 до 30 процентов ошибки предыдущего прогноза. Более высокие значения константы ускоряют отклик, но могут привести к непредсказуемым выбросам. Низкие значения константы могут привести к большим промежуткам между предсказанными значениями.

Двухвыборочный F-тест для дисперсии. Двухвыборочный F-тест применяется для сравнения дисперсий двух генеральных совокупностей. Например, F-тест можно использовать для выявления различия в дисперсиях временных характеристик, вычисленных по двум выборкам.

Анализ Фурье. Предназначается для решения задач в линейных системах и анализа периодических данных на основе метода быстрого преобразования Фурье (БПФ). Эта процедура поддерживает также обратные преобразования, при этом, инвертирование преобразованных данных возвращает исходные данные.

Гистограмма. Используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений. При этом рассчитываются числа попаданий для заданного диапазона ячеек. Например, необходимо выявить тип распределения успеваемости в группе из 20 студентов. Таблица гистограммы состоит из границ шкалы оценок и количеств студентов, уровень успеваемости которых находится между самой нижней границей и текущей границей. Наиболее часто повторяемый уровень является модой интервала данных.

Скользящее среднее**.** Скользящее среднее используется для расчета значений в прогнозируемом периоде на основе среднего значения переменной для указанного числа предшествующих периодов. Скользящее среднее, в отличие от простого среднего для всей выборки, содержит сведения о тенденциях изменения данных. Этот метод может использоваться для прогноза сбыта, запасов и других процессов.

Генерация случайных чисел. Используется для заполнения диапазона случайными числами, извлеченными из одного или нескольких распределений. С помощью данной процедуры можно моделировать объекты, имеющие случайную природу, по известному распределению вероятностей. Например, можно использовать нормальное распределение для моделирования совокупности данных по росту индивидуумов, или использовать распределение Бернулли для двух вероятных исходов, чтобы описать совокупность результатов бросания монеты.

Ранг и персентиль. Используется для вывода таблицы, содержащей порядковый и процентный ранги для каждого значения в наборе данных. Данная процедура может быть применена для анализа относительного взаиморасположения данных в наборе.

Регрессия. Линейный регрессионный анализ заключается в подборе графика для набора наблюдений с помощью метода наименьших квадратов. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных. Например, на спортивные качества атлета влияют несколько факторов, включая возраст, рост и вес. Регрессия пропорционально распределяет меру качества по этим трем факторам на основе его спортивных результатов. Результаты регрессии впоследствии могут быть использованы для предсказания качеств нового, непроверенного атлета.

Выборка. Создает выборку из генеральной совокупности, рассматривая входной диапазон как генеральную совокупность. Если совокупность слишком велика для обработки или построения диаграммы, можно использовать представительную выборку. Кроме того, если предполагается периодичность входных данных, то можно создать выборку, содержащую значения только из отдельной части цикла. Например, если входной диапазон содержит данные для квартальных продаж, создание выборки с периодом 4 разместит в выходном диапазоне значения продаж из одного и того же квартала.

Т-тест. Этот вид анализа используется для проверки средних для различных типов генеральных совокупностей. **Двухвыборочный t-тест с одинаковыми дисперсиями.** Двухвыборочный t-тест Стьюдента служит для проверки гипотезы о равенстве средних для двух выборок. Эта форма t-теста предполагает совпадение дисперсий генеральных совокупностей и обычно называется гомоскедастическим t-тестом. **Двухвыборочный t-тест с разными дисперсиями**. Двухвыборочный t-тест Стьюдента используется для проверки гипотезы о равенстве средних для двух выборок данных из разных генеральных совокупностей. Эта форма t-теста предполагает несовпадение дисперсий генеральных совокупностей и обычно называется гетероскедастическим t-тестом. Если тестируется одна и та же генеральная совокупность, используйте парный тест. **Парный двухвыборочный t-тест для средних.** Парный двухвыборочный t-тест Стьюдента используется для проверки гипотезы о различии средних для двух выборок данных. В нем не предполагается равенство дисперсий генеральных совокупностей, из которых выбраны данные. Парный тест используется, когда имеется естественная парность наблюдений в выборках, например, когда генеральная совокупность тестируется дважды — до и после эксперимента. Одним из результатов теста является совокупная дисперсия (совокупная мера распределения данных вокруг среднего значения), вычисляемая по следующей формуле.

Z-тест. Двухвыборочный z-тест для средних с известными дисперсиями. Используется для проверки гипотезы о различии между средними двух генеральных совокупностей. Например, этот тест может использоваться для определения различия между характеристиками двух моделей автомобилей.

4.2 Краткая характеристика систем Mathematica, MATLAB и Maple

Рост сложности решаемых задач по объективным причинам ведёт к сложности алгоритмов и их реализаций на алгоритмических языках Си, Паскаль, Фортран и др. Ещё больше времени уходит на отладку кода. Эти причины привели к созданию систем автоматизированного проектирования (САПР), в которые заложены некие алгоритмы. Такие системы появились достаточно давно, и были узкоспециализированными. Среди математических САПР наибольшую популярность приобрели MathCAD (MathSoft Inc.), Mathematica (Wolfram Research, Inc.), MATLAB (MathWorks Inc.), Maple V (Waterloo Maple Inc.).

Краткая характеристика Maple и Mathematica. К среднему уровню таких систем относятся интенсивно развиваемые системы класса Mathcad, имеющие (в дополнение к прекрасным средствам числовых вычислений) приобретенное по лицензии у фирмы Waterloo Maple Inc. (создателя систем Maple) ядро символьных вычислений. Ядро системы Maple используется и в другой маститой системе — MATLAB, придавая ей необычные для нее возможности символьной математики.

Одна из самых мощных и интеллектуальных систем компьютерной алгебры — Maple под Windows была создана группой ученых, занимающихся символьными вычислениями (The Symbolic Group), организованной Кейтом Геддом (Keith Geddes) и Гастоном Гонэ (Gaston Gonnet) в 1980 году в университете Waterloo, Канада. Вначале она была реализована на больших компьютерах и прошла долгий путь апробации, вобрав в свое ядро и библиотеки большую часть математических функций и правил их преобразований, выработанных математикой за столетия развития. Есть реализации программы на платформах ПК Macintosh, Unix, Sun и др. Вряд ли эта мощная математическая система, разделяющая претензии на мировое лидерство с системами Mathematica фирмы Wolfram Research Inc., нужна секретарше или даже директору небольшой коммерческой фирмы. Но, несомненно, любая серьезная научная лаборатория или кафедра вуза должны располагать подобной системой, если они всерьез заинтересованы в автоматизации выполнения математических расчетов любой степени сложности. Несмотря на свою направленность на серьезные математические вычисления, системы класса Maple необходимы довольно широкой категории пользователей: студентам и преподавателям вузов, инженерам, аспирантам, научным работникам и даже учащимся математических классов общеобразовательных и специальных школ. Все они найдут в Maple многочисленные достойные возможности для применения.

Сравнение системы Maple 7 с лидером среди систем компьютерной математики — системой Mathematica 4.1 — непродуктивно. У каждой программы есть свои достоинства и недостатки. А главное — у них есть свои приверженцы, которых бесполезно убеждать, что иная система в чем-то лучше. Все, кто всерьез применяют системы компьютерной математики, должны работать с несколькими системами, ибо только это гарантирует высокий уровень надежности сложных вычислений.

И все же надо отметить, что интерфейс Maple 7 более интуитивно понятен, чем у строгой Mathematica 4.1. Maple 7 на первый взгляд имеет несколько менее мощную графику, но простота управления параметрами и легкость подготовки графических процедур часто позволяет визуализировать решения математических задач с меньшими усилиями, чем при использовании системы Mathematica 4.1. Обе системы в последних реализациях сделали качественный скачок в направлении эффективности решения задач в численном виде, в частности за счет повышения скорости выполнения матричных операций.

Особенно эффективно использование Maple при обучении математике. Высочайший «интеллект» этой системы символьной математики объединяется в ней с прекрасными средствами математического численного моделирования и просто потрясающими возможностями графической визуализации решений. Применение таких систем, как Maple, возможно при преподавании и самообразовании от самых основ до вершин математики.

Практика показывает, что самым трудным является первый этап освоения системы. Первое знакомство с программой Maple многих пользователей просто подавляет — убедившись в невероятном множестве возможностей системы и не имея ее систематизированного описания (а оно поставляется в виде трех книг приличного размера, включая книги учебного характера), многие пользователи помещают систему в архив, где она «пылится» без дела.

Краткая характеристика пакета MATLAB. MATLAB является системой численных вычислений (хотя Math Works Inc. и закупил некоторые библиотеки символьных вычислений у Waterloo Maple Inc.), имеет смысл для некоторых «механических» расчётов использовать пакеты символьных вычислений, например Maple V.

Особенностью MATLAB является надстройка Simulink, которая позволяет решать многие задачи в режиме RAD — создать модель из отдельных блоков и запустить процесс.

Приведем в таблице 4.1 сравнительную характеристику системы MATLAB и Maple V:

Таблица 4.1 - Сравнительная характеристика системы MATLAB и Maple V

MATLAB	MAPLE
Ориентация на численные методы	Пакет символьных вычислений. Очень удобный инструмент для относительно несложных расчётов. Наиболее удачное применение совместно с MATLAB
Поддержка сценариев и включение новых алгоритмов	Отсутствует возможность включения новых алгоритмов

Сохранение результатов решения на диске, их загрузка в память в нужный момент, использование в других сценариях и т.д	Невозможно сохранить результаты решения на диске, т.е. необходимо каждый раз запускать сценарий заново
Решение систем дифференциальных уравнений (СДУ) в форме Коши. В моей практике MATLAB всегда находил решение СДУ. Имеется несколько методов, в том числе для жёстких систем	Решение систем дифференциальных уравнений и дифференциальных уравнений высшего порядка. Однако для жестких систем или сложных уравнений решение не всегда может быть получено или процесс займёт несравненно много времени. Можно использовать как численные, так и символьные методы. В случае символьных методов решение будет дано в общем виде. Решение ДУ с помощью преобразования Лапласа
Удобный внутренний язык описания сценариев	Несколько запутанный язык описания сценариев
Автоматизированная компиляция написанных вами функций, создание динамически загружаемых библиотек, исполняемых приложений	Отсутствует
Создание моделей объектов по технологии RAD в среде Simulink	Отсутствует
Создание графического пользовательского интерфейса	Отсутствует
Линкование с MS Word и, соответственно, у вас все возможности редактора. Принцип напоминает MathCAD. Однако, учитывая, что одновремення работа MATLAB и MS Word плохо сказывается на системных ресурсах компьютера, использование данной возможности весьма сомнительно	Верстка документа непосредственно в рабочей программе. Создание раскрывающихся уровней программы, что весьма удобно. Однако для документирования абсолютно не подходит
Работа со звуком, изображениями, анимация	Отсутствует (создание анимированных рисунков в формате GIF несколько иное приложение анимации, не имеет исследовательского применения)

4.3 Краткая характеристика пакета Statistica

STATISTICA является одним из наиболее мощных программных средств по статистической обработке - это универсальная интегрированная программная система, предназначенная для статистического анализа и визуализации данных, управления базами данных и разработки пользовательских приложений, содержащая 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 имеются специализированные 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.

STATISTICA представляет собой интегрированную систему статистического анализа и обработки данных. 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.

STATISTICA предоставляет пользователю уникальную среду экспериментирования, разведки, графического отображения и углубленного анализа данных, в которой статистическая обработка становится не рутинным занятием, а увлекательным исследованием с использованием новейших компьютерных технологий и современных приемов и методов.

Опишем основные модули пакета STATISTICA:

– модуль диалога

– модуль построения/выполнения технологических цепочек

– модуль управления данными.

– модуль статистических функций

– модуль визуализации

Модуль диалога. С помощью данного модуля пользователь производит выбор необходимого ему метода обработки информации или технологической цепочки методов и выполняет настройку соответствующих параметров. Данная компонента обеспечивает удобный интерфейс с пользователем в системе Windows. Предлагается широкий спектр диалоговых окон для настройки как параметров предоставляемого набора функций (статистической обработки, визуализации результатов и т.д.), так и параметров среды общения, что позволяет настроить модуль под конкретного пользователя. Также обеспечивается возможность справки по текущей ситуации.

Модуль построения/выполнения технологических цепочек. Данный модуль позволяет описывать часто используемую цепочку действий по обработке информации в виде пакета команд предлагаемого SCL-языка (STATISTICA Command Language), синтаксис которого очень похож на распространенный язык Basic или Pascal. Практически все возможности пакета STATISTICA продублированы соответствующими командами-функциями, на вход которых подаются продекларированные параметры настройки. Заметим, что возможности расширены вплоть до имитации работы пользователя (запись и воспроизведение команд пользователя и имитация работы органов управления посредством SCL-команд). Модуль реализован в виде двух Windows-программ: sta_com.exe (конструктор) и sta_run.exe (процессор). Созданные технологические цепочки хранятся в базе знаний в виде текстового файла с расширением SCL и могут выполняться посредством процессора sta_run из других Windows-приложений. Эта возможность позволяет строить проблемно-ориентированный программный инструментарий, который будет более понятен и прост в использовании для специалиста, что значительно повысит эффективность его работы.

Модуль управления данными. STATISTICA обеспечивает широкие возможности импорта/экспорта из различных стандартных типов баз данных как Windows, так и DOS версий (Symphony, Quattro, dBASE III+, dBASE IV, Paradox и ASCII формат). При импорте данные переводятся во внутренний формат (STA-формат), поддерживающий 32000 переменных (признаков) и обеспечивающий более быструю обработку данных. Предлагаются возможности верификации (попадание в интервал), различные варианты ранжирования, кодировки данных (соотнесение значений по заданным категориям), создание переменных по уравнению связи признаков, нормировки, смещения значений признаков, замена пропусков на медианное значение, а также стандартные возможности редактирования данных (выделения подвыборки, копирования, удаления, добавления, перемещения, транспонирования и сортировки). Для автоматизации выполнения небольшого объема действий по управлению данными имеется встроенная версия SCL-языка – QMML (Quick Megafile Manager Language). Имеется также возможности установления DDE-связи с другими Windows-приложениями (такие как Excel, MS Word, Ami Pro, Quattro Pro), а также поддержка OLE. Модуль реализован в виде Windows-программы: sta_dat.exe.

Модуль статистических функций. Здесь предлагается широкий спектр возможностей статистического анализа данных. Из-за большого объема информации по предлагаемым методам анализа ограничимся краткой характеристикой каждого из них:

– Basic Statistics – базовый статистический анализ: описательная статистика; описательная статистика для групп; t-тест для зависимых и независимых выборок; построение матрицы парных корреляций, частотных таблиц (гистограмм) и другое (реализован в виде Windows-программы sta_bas.exe);

– Nonparametrics – непараметрический анализ (внутри и межгрупповые различные тесты, корреляции), обычная описательная статистика (процентили, медиана и т.д. (sta_non.exe);

– Linear Regression – множественная линейная регрессия (различные методы определения), фиксированная нелинейная регрессия (полиномиальная) (sta_lin);

– NonLinear Estimation – построение нелинейной регрессии (определенного пользователем вида: фиксированной экспоненциальной, ломаной линейной и др.) с помощью различных аппроксимаций (Симплекс, Квази-Ньютона, Хук-Риверса и др.) (sta_log);

– Time Series and Forecasting – анализ временных рядов и прогноз, а также различного рода сглаживания, трансформации, определение сезонных колебаний и т.д. (sta_tim);

– Cluster Analysis – кластерный анализ, позволяющий выделять однородные группы с помощью метода К-средних, попарного объединения и иерархического метода (sta_clu).

– Factor Analysis – факторный анализ методом выделения главных компонент, максимального правдоподобия факторов, центроидный метод, метод главных аксцис (sta_fac);

– Canonical Analysis – оценка взаимосвязи признаков методом главных компонент (sta_can);

– Multidimensional Scaling – многомерное шкалирование, анализ расстояний или однородности/разнородности, восстановление расстояний (sta_mul)

– Reliability & Item Analysis – методы построения и анализа тестов, построения различных корреляций (Кронбач-альфа, часть-целое, множественные) (sta_rel);

– Discriminant Function Analysis – дискриминантный анализ (sta_dis);

– Survival Analysis – анализ процессов гибели и размножения: описание и сравнение развития; анализ таблиц развития; тест Каплан-Мейера и тесты для двух и более выборок, определение типа распределения (Вейбул, Гомпертц, ...), построение регрессионных моделей (лог-нормальная, экспоненцияльная и другие формы) (sta_sur);

– Quality Control – контроль качества, анализ различных диаграмм (X, R, S, Si, CUSUM, Парето и др.) (sta_qua);

– Process Analysis – анализ процессов: расчет плана по среднему, по пропорциям и Пуассоновским частотам, анализ совместности процессов и интервалов толерантности и другое (sta_pro);

– Experimental Design – планирование экспериментов, построение ДФП, ЦКП и др. (sta_exp);

Модуль визуализации. Данный модуль обеспечивает отображение результатов работы статистических процедур. Предлагается два режима отображения – табличный и графический. Для графического отображения предусмотрены следующие возможности:

– 2-мерная графика: гистограмма, XY-проекция, вероятностная бумага (нормальная, полунормальная), ящики с «усами», круговые диаграммы и другое;
– 3-мерная графика: проекции, гистограммы, ящики с «усами» и т.д.;
– различные 3-мерные проекции;
– отображение многомерной выборки в виде «лиц Чернова», графиков Кивиата, полигонов, профилей и т.д.

Имеется возможность сохранения результатов в виде файлов данных (для таблиц), STG-файлов (для графики) либо в виде твердой копии с помощью функций печати. Используя стандартные OLE-операции можно также вставлять результаты в документы и электронные таблицы (Word, PageMaker, Excel и другие приложения Windows, поддерживающие стандарт OLE).

Раздел 2 Первичная статистическая обработка

Тема 5 Этапы статистической обработки

5.1 Основные этапы статистической обработки экспериментальных данных
5.2 Основная цель разведочного анализа данных
5.3 Методы разведочного анализа данных
5.4 Модели структуры многомерных данных в разведочном анализе данных

5.1 Основные этапы статистической обработки экспериментальных 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 $(x^1,\ldots,x^p)$ 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 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. Сбор исходных данных и введение их в ЭВМ, а также 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): $x_i^{(k)}(t),\ t=t_1\ldots t_N,\ k=\overline{(1,p)},\ i=\overline{(1,N)}$; 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. Составление детального плана вычислительного анализа. 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 первого подхода (см. п.1.2), то выводы формируются в терминах оценок неизвестных параметров, или в виде отчета о справедливости гипотез с указанием количественной степени достоверности. В случае второго подхода вероятностная интерпретация не делается.

Работа завершается содержательной формулировкой новых задач, вытекающих из проведенного исследования.

5.2 Основная цель разведочного анализа данных

Этап разведочного 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 третьем томе справочника по прикладной статистике Айвазяна С.А. и др. 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.

5.3 Методы разведочного анализа данных

Методы разведочного (предмодельного) статистического анализа данных, 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 равно 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).

5.4 Модели структуры многомерных данных в разведочном анализе 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:

1- модель облака точек примерно эллипсоидальной конфигурации;

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

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

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

5- 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 соответствующие многообразие допускает функциональное представление $X_{11} = F(X_1) + \varepsilon$, где X_{11}- прогнозируемые, X_1-предсказывающие признаки, $F(X_1)$- функция регрессии, ε- ошибка.

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

.

Тема 6 Предварительный статистический анализ данных

6.1 Содержательная и математическая постановка задачи статистического описания
6.2 Содержательная и математическая постановка задачи статистического прогноза
6.3 Схема взаимодействия переменных при статистическом исследовании зависимостей
6.4 Математический инструментарий СИЗ

6.1 Содержательная и математическая постановка задачи статистического описания

Любое экспериментальное исследование содержит этапы постановки задачи, планирования и проведения эксперимента, а также анализа и интерпретация результатов. Главной трудностью на этапе постановки задачи является переход с языка специальности на язык планирования эксперимента, на язык математики.

Содержательная постановка задач статистического **описания и прогноза** является переходной формулировкой, позволяющей перейти к математической, на основании выявленной цели исследования. Математическая постановка задач статистического **описания и прогноза** предполагает то, что формулировка задачи будет сделана в терминах, используемых в конкретной формальной дедуктивной системе.

Математическая постановка задач статистического описания предназначена для описания структуры множества выборочных точек и для формирования адекватных реальности рабочих исходных допущений, на которых строится дальнейшее исследование. В вероятностно-статистическом подходе математическая постановка задач статистического описания может состоять в оценке закона распределения. В логико-комбинаторном подходе, или в РАД используется одна из первых четырех статистических моделей: модель облака точек, кластерная модель, модель «засорения» и эмпирический образ данных

В общем виде задачу классификации исследуемой совокупности N объектов $O=\{O_i\}$, $i=\overline{1,N}$, где для каждого объекта замерены значения p параметров, т.е. каждый объект O_i описан вектором $X_i = (x_i^1,...,x_i^p)$, можно сформулировать как задачу поиска такого разбиения S заданной совокупности на непересекающиеся классы $S_1,....,S_k$: $\cup S_j = O, S_i \cap S_j = \varnothing$, $i \neq j$, при котором функционал качества Q(S) достигает экстремального значения на множестве A допустимых правил классификации. В качестве Q(S) используют критерии, минимизирующие межгрупповое сходство и одновременно максимизирующее внутригрупповое сходство. Состав множества A зависит от предварительной (априорной) выборочной информации об этих классах. Итак, задача классификации формально сводится к нахождению разбиения S^*: $Q(S^*) = \min Q(S)$ для $S \in A$. Заметим, что при этом число k может быть и неизвестно. При любых трактовках кластеров и для различных методов классификаций неизбежно возникает проблема измерения близости объектов. С этой проблемой связаны следующие трудности: неоднозначность выбора способа нормировки и определения расстояния между объектами.

6.2 Содержательная и математическая постановка задачи статистического прогноза

Построение математической модели, например. Технологического процесса в зависимости от поставленной задачи может преследовать следующие цели: минимизировать расход материала на единицу выпускаемой продукции при сохранении качества, произвести замену дорогостоящих материалов на более дешевые или дефицитных на распространение; сократить время обработки в целом или на отдельных операциях, перевести отдельные режимы в некритические зоны, снизить трудовые затраты на единицу продукции и т.п.; улучшить частные показатели и общее количество готовой продукции, повысить однородность продукции, улучшить показатели надежности и т.п.; увеличить надежность и быстродействие управления, увеличить эффективность контроля качества, создать условия для автоматизации процесса управления и т.п. Прежде всего, необходимо выбрать зависимую переменную Y, которую обычно называют целевой функцией или параметром оптимизации, за который принимают один из показателей качества продукции либо по каждой технологической операции отдельно, либо по всему технологическому процессу сразу. Параметр оптимизации должен соответствовать следующим требованиям: параметр должен измеряться при любом изменении (комбинации) режимов технологического процесса; параметр должен быть статистически эффективным, то есть измеряться с наибольшей точностью; параметр должен быть информационным, то есть всесторонне характеризовать технологический процесс (операцию); параметр должен иметь физический смысл, то есть должна быть возможность достижения полезных результатов при соответствующих условиях процесса; параметр должен быть однозначным, т.е. должно минимизироваться или максимизироваться только одно свойство изделия.

Для достоверного отображения объективно существующих процессов необходимо выявить существенные взаимосвязи и не только выявить, но и дать им количественную оценку. Этот подход требует вскрытия причинных зависимостей. Под причинной зависимостью понимается такая связь между процессами, когда изменение одного из них является следствием изменения другого.

Сформулируем математическую постановку задачи статистического прогноза на примере задачи регрессионного анализа в п.3.

6.3 Схема взаимодействия переменных при статистическом исследовании зависимостей

Основная цель статистического исследования зависимостей (СИЗ) состоит в том, чтобы на основании частных результатов статистического наблюдения за показателями двух или трех различных явлений, происходящих с исследуемым объектом, выявить и описать существующие взаимосвязи. В случае численного выражения такие показатели называют переменными.

Рамки применения аппарата СИЗ определяются двумя условиями: - стохастичность интересующей нас взаимосвязи между переменными (т.е. реализация явления или события А одной переменной может повлечь за собой событие В другой переменной с вероятностью *p*); - взаимосвязь между переменными выявляется на основе статистических наблюдений по выборкам из соответствующих генеральных совокупностей событий.

Опишем функционирование изучаемого реального объекта набором переменных, среди которых выделим: $x^{(1)},\ldots, x^{(p)}$ - «входные» переменные, описывающие условия или причинные компоненты функционирования (поддаются контролю или частичному управлению); для них используются такие термины как факторы-аргументы, факторы-причины, экзогенные, предикторные (предсказательные), объясняющие; $y^{(1)},\ldots, y^{(m)}$ - «выходные», характеризующие поведение объекта или результат (эффективность) функционирования; обычно их называют отклики, эндогенные, результирующие, объясняемые, факторы-следствия, целевые факторы; $\varepsilon^{(1)},\ldots, \varepsilon^{(m)}$ - латентные (скрытые, не поддающиеся непосредственному измерению) случайные «остаточные» компоненты, отражающие влияние на $y^{(1)},\ldots, y^{(m)}$ неучтенных «на входе» факторов, а также случайные ошибки в измерении анализируемых показателей; остатки.

Используя введенный набор переменных, задача СИЗ может быть сформулирована следующим образом: по результатам *N* измерений

$$(x_i^{(1)},\ldots,x_i^{(p)},y_i^{(1)},\ldots,y_i^{(m)}),\ i=\overline{1,N}$$

исследуемых переменных на *N* объектах построить такую (векторно-значимую) функцию

$$f(x^{(1)},\ldots,x^{(p)})=\begin{pmatrix} f^{(1)}(x^{(1)},\ldots,x^{(p)}) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f^{(m)}(x^{(1)},\ldots,x^{(p)}) \end{pmatrix},$$

которая позволила бы наилучшим образом восстановить значения переменных $Y=(y^{(1)},\ldots,y^{(m)})'$ по заданным значениям объясняющих переменных $X=(x^{(1)},\ldots,x^{(p)})'$.

6.4 Математический инструментарий СИЗ

Методы СИЗ составляют содержание отдельных частей многомерного статистического анализа, которые можно определить как раздел математической статистики, посвященный построению оптимальных планов сбора, систематизации и обработки многомерных статистических данных, нацеленных на выявление характера и структуры взаимосвязей между компонентами *(X,Y)* и предназначенных для получения практических и научных выводов. Среди *p+m* компонент могут быть: количественные, порядковые (ординальные), классификационные (номинальные).

Методы СИЗ формировались с учетом специфики моделей, обусловленных природой изучаемых переменных. Схематично всю совокупность методов приведем в таблице 6.1.

Таблица 6.1 - Математический инструментарий СИЗ

Природа результирующих показателейY	Природа объясняющих переменных X	Названия обслуживающих разделов многомерного статистического анализа
Количественная	Количественная	Регрессионный и корреляционный анализ
Количественная	Одна количественная переменная, интерпретируемая, как время	Анализ временных рядов
Количественная	Неколичественная (ординальные или номинальные переменные)	Дисперсионный анализ
Количественная	Смешанная (количественные и неколичественные переменные)	Ковариационный анализ, модели типологической регрессии
Неколичественная (порядковые переменные)	Неколичественная (ординальные или номинальные переменные)	Анализ ранговых корреляций и таблиц сопряженности
Неколичественная (номинальные переменные)	Количественная	Дискриминантный анализ, кластер-анализ, расщепление смесей распределения
Смешанная (количественные и неколичественные переменные)	Смешанная (количественные и неколичественные переменные)	Аппарат построения логических решающих функций и эмпирического образа данных

Краткая характеристика математического инструментария

Корреляционный анализ оценивает степень тесноты статистической взаимосвязи и обосновывает целесообразность регрессионного анализа. Регрессионный анализ позволяет получить прогноз количественных значений результирующей переменной по значениям входных. Анализ временных рядов занимается исследованием поведения результирующих переменных во времени. Дисперсионный анализ выявляет наличие взаимосвязи между качественными показателями и результирующей переменной.

Тема 7 Оценка закона распределения. Непараметрический подход

7.1 Разновидности первичной статистической обработки
7.2 Параметрическое и непараметрическое оценивание закона распределения
7.3 Равноинтервальная гистограмма и полигон частот
7.4 Равнонаполненная гистограмма и полигон частот
7.5 Метод прямоугольных вкладов

7.1 Разновидности первичной статистической обработки

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.

7.2 Параметрическое и непараметрическое оценивание закона распределения

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.

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.

7.3 Равноинтервальная гистограмма и полигон частот

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

[ ], [ ]

Полигон частот $\varphi_N(x)$ получают путем сглаживания гистограммы

$$\varphi_N(x) = \frac{m_k + m_{k+1}}{2Nh} + (x - a_k)\frac{m_{k+1} - m_k}{Nh^2},\ x \in [x_k, x_{k+1}],$$

где $x_k (k = \overline{1,S})$ - середина промежутка Δ_k, a_k-правый конец промежутка Δ_k.

Очевидно, что $\varphi_N(x_k) = \frac{m_k}{Nh}, \varphi_N(x_{k+1}) = \frac{m_{k+1}}{Nh}$.

7.4 Равнонаполненная гистограмма и полигон частот

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 на *k* равных частей. Тогда площадь каждой части равна $\Delta_1 h_i = \Delta_2 h_2 = \cdots = \Delta_i h_i = \cdots = 1/s$, $h_i = 1/(s \cdot \Delta_i)$. Для конкретной выборки рассчитываются длины интервалов Δ_i, а затем по формуле $h_i = 1/(s \cdot \Delta_i)$, определяется h_i. На основании полученных значений длины и высоты каждого прямоугольника гистограммы получаем оценку $f_N(x)$.

7.5 Метод прямоугольных вкладов

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.

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

$$f(x) = \alpha_0 f_0(x) + \frac{1-\alpha_0}{N}\sum_{i=1}^{N} p(x - x_i),$$

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); α_0 - вес априорной компоненты.

Различным методам исследования соответствуют разные значения $\alpha_0 \in [0,1]$ и разные виды функции $p(x - x_i)$. Широко известны оценки «ядерного» типа для *f*(x) при значении $\alpha_0 = 0$.

В методе прямоугольных вкладов (МПВ)

$$\alpha_0 = \frac{1}{N+1},\ f_0(x) = \begin{cases} \frac{1}{b-a}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases},$$

,

где [*a*,*b*] -интервал изменения случайной величины *x*; d - ширина функции вклада.

В качестве d может быть взято, например: $d = \frac{1}{s}(b-a)$, 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: $y_{(1)}, y_{(2)}, \ldots, y_{(2N)}$;

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

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

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 $y_{(1)}, y_{(2)}, \ldots, y_{(2N)}$.

Тема 8 Оценка закона распределения. Параметрический подход

8.1 Нормальная вероятностная бумага

8.2 Параметрическое оценивание

8.3 Критерием согласия χ^2

8.1 Нормальная вероятностная бумага

Пусть даны *N* наблюдений $x_1,\ldots,x_N$, извлеченные из генеральной совокупности с функцией распределения *F(t)*. Пусть $x^{(1)},\ldots,x^{(N)}$- упорядоченный по возрастанию ряд наблюдений. Тогда за оценку *F(t)* принимают $\hat{F}(t) = i/N$, где $i \to \max_j x^{(j)}$; $x^{(j)} \le t$

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 $\hat{F}(t)$ часто используют специальную шкалу, откладывая по оси ординат вместо $\hat{F}(t)$ величину $V = F^{-1}(\hat{F}(t))$, 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 *Φ(.)*- стандартная функция нормального распределения.

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.рис. 2.2.1) 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 $x^{(1)},\ldots, x^{(N)}$.

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* .

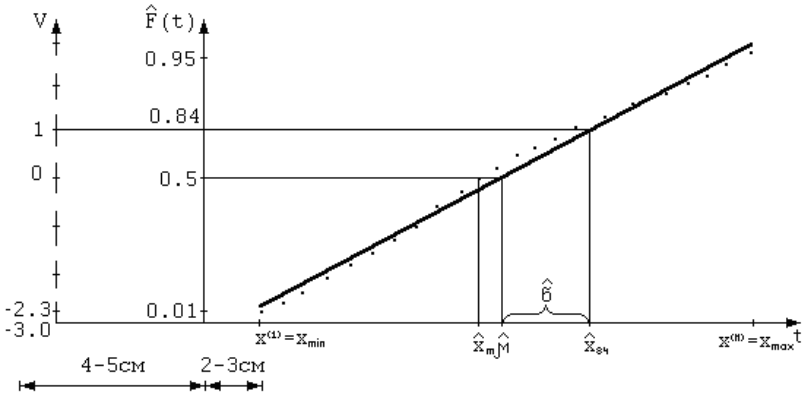

Рисунок 2.2.1 – Оценка параметров нормального распределения на нормальной вероятностной бумаге

Этап 4. Если точки ($x^{(i)}$,i/N) в какой-то мере ложатся вдоль некоторой прямой, то можно грубо считать генеральную совокупность, из которой извлечена данная выборка, нормальной. 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. Оценка стандартного отклонения $\hat{\sigma} = \hat{x}_{84} - \bar{\mu}$, где $\hat{x}_{84}$ - оценка 0.84 квантиля распределения, полученного при *V=1*.

8.2 Параметрическое оценивание

Построение гистограммы, полигона частот и ядерной аппроксимации основано на локальной интерполяции. Другой подход к аппроксимации заключается в интерполировании закона распределения на всем интервале [*a,b*]. К методам этого типа относится аппроксимация с помощью системы кривых Пирсона. 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.

8.3 Критерием согласия χ^2

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

Этап 1. Область изменения выборки [*a,b*] делим на S равных интервалов, как при построении гистограммы. 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:

$$Q^2 = \sum_{S=1}^{S'} \frac{(m_S - n_S)^2}{n_S}$$

Этап 4. Случайная величина Q^2 имеет $\chi^2(\nu)$ распределение с числом степеней свободы *v=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 (см. таблицу процентилей распределения χ^2). Если *P* < α, то мы отвергаем H_0 и принимаем гипотезу H_1: теоретическое и эмпирическое распределения не согласованы. Здесь α - это уровень значимости, который обычно принимается равным 0.05.

Тема 9 Восстановление пропущенных значений и анализ выбросов

9.1 Восстановление пропущенных значений
9.2 Алгоритм ZET
9.3 Анализ выбросов
9.4 Проверка гипотез

9.1 Восстановление пропущенных значений

Непараметрический подход 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).

9.2 Алгоритм ZET

Рассмотрим схематично конкретизацию этого подхода в известном алгоритме ZET. 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 по 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 заполняются. Мера качества заполнения определяется с помощью меры заполнения истинных значений от полученных.

9.3 Анализ выбросов

При наличии таких данных возникает вопрос: чем объяснить обнаруженные резкие отклонения в исходных данных? Например, объясняются ли они природой анализируемой генеральной совокупности? Если случайные

колебания выборочных значений обусловлены искажениями стандартных условий сбора статистических данных или прямыми ошибками регистрации и записи, то их надо исключить. Наиболее надежным способом решения вопроса об исключении данных из рассмотрения является изучение условий регистрации и сбора данных. Если невозможен анализ условий, при которых регистрировалось аномальное наблюдение, то обращаются к статистическим методам. Их общая логическая схема: исходя из исходных предложений о природе анализируемой совокупности данных, исследователь задается функцией $\Psi(X^*, X)$ (X - все имеющие наблюдения, X^* - подозрительные наблюдения), характеризующей степень аномальности, определяет значение Ψ и сравнивает с пороговым значением Ψ. При $\Psi > \Psi_0$ подозрительное наблюдения исключается, или для него определяется весовой коэффициент. В вероятностной постановке Ψ_0 определяется из стандартных статистических таблиц с учётом закона распределения статистики Ψ в предположении необоснованности относительно X^*. В других случаях Ψ_0 определяется из содержательных соображений.

9.4 Проверка гипотез

Статистические процедуры анализа резко выделяющихся наблюдений основаны на предположении однородности данных. При этом выбросы рассматриваются как наблюдения, нетипично удаляющиеся от центра распределения. Основная трудность при использовании имеющихся аналитических процедур состоит в том, что реальная доля «засорения» не известна, а оценивается по тем же данным, по которым проверяется значимость отклонения. Наиболее устойчивы к отклонениям от предположения нормальности основной части выборки графические процедуры. При использовании статистических методов выделения выбросов следует иметь в виду, что выбросы могут оказаться наиболее существенной частью выборки, проясняющей, например то, как собирались данные (например, изменение условий эксперимента, не замеченное исследователем). Данная задача распадается на два этапа: выделение подозрительных наблюдений; проверка статистической значимости отличий от основной массы данных. Оба этапа основываются на определенных предположениях о распределении основной (не засоренной) части наблюдений и выбросов (засорений). Обычно предполагают, что не засоренная часть наблюдений имеет одно или многомерное нормальное распределение с неизвестными параметрами $N(\mu, o^2)$, а засоренная: $N(\mu + d, o^2)$ или $N(\mu, \gamma o^2), \gamma \geq 1$.

Тема 10 Унификация признакового описания

10.1 Отношение, признаки, измерения
10.2 Типовые структуры признаков
10.3 Типы шкал
10.5 Унификация типа переменных

10.1 Отношение, признаки, измерения

Для описания разнородных задач первичной статистической обработки помимо обычного языка математической статистики удобно использовать терминологию теории бинарных отношений. Опишем кратко основные понятия.

Отношения. *Бинарное отношение* Р на множестве объектов A - подмножество упорядоченных пар объектов (а, в) декартового произведения A на A : $A \times A$.

У некоторых особо важных отношений есть специальные названия.

Отношение *эквивалентности* разбивает все множество объектов на не пересекающиеся классы, в каждом из которых объекты признаются тождественными, неразличимыми, а из разных классов – нетождественными.

Квазипорядок (нестрогий порядок) определяет отношение «быть не меньше». Если исключить из него возможность равенства элементов, то оно превратится в порядок.

Толерантностью называется отношение «похожести». В анализе данных оно имеет особую роль, так как объединение объектов происходит по похожести. Здесь в отличие от эквивалентности из а=в, в=с не следует а=с.

Метризованное отношение. Каждому отношению на множестве объектов $a_1, \ldots, a_n$ можно сопоставить матрицу $N * N$ из бинарных значений r_{ij}={0,1}, где $r_{ij} = 1,$ для $(a_i, a_j) \in P$, $r_{ij} = 0,$ иначе. Понятие «отношение « можно расширить, распространив его на количественные признаки. В 1977 Б. Г. Литваком введено понятие «метризованного отношения». «Метризованным отношением» называется пара <W(P),P> , где Р – отношение, W(P) – множество чисел (весов), характеризующих «степень принадлежности» пары к данному «метризованному отношению». Вместо булевских матриц (2.2) вводятся матрицы с вещественными элементами p_{ij}, которые определяются (для линейных отношений порядка).

$$p_{ij} = \begin{cases} W_{ij}, if(a_i, a_j) \in P \\ -W_{ij}, if(a_j, a_i) \in P \end{cases}$$

Признаки. Отношения определены на парах объектов. Признак – это свойство, измеренное на каждом объекте. Может случиться, что отношение существует, а измеримые признаки им не отвечают. Так, отношению толерантности нельзя сопоставить признак, определенный на каждом объекте.

Измерение. Рассмотрим способы измерения признаков. Обычно под процедурой измерения какого-либо свойства понимается приписывание некоторых числовых значений отдельным уровням этого свойства в определенных единицах. При этом важно знать в какой мере условность в выборе единиц измерения повлияет на значение показателя. Например, если стоимость продукции измерить в рублях, а потом в тысячах рублей, то изменится лишь число единиц измерения, суть же останется прежней. Здесь возможно умножение, деление на константу, т. Е. масштабирование. Бессмысленно задавать масштаб для температуры по Цельсию, так как мы не можем сказать во сколько раз $-5\,^{\circ}C$ меньше $+10\,^{\circ}C$. Таким образом разные типы признаков имеют разное множество допустимых преобразований $f(x)$ своих значений, которое определяет *тип шкалы*.

10.2 Типовые структуры признаков

Признаки, описывающие объекты получаются по-разному. В зависимости от того, как измеряют или оценивают значение признака, они могут быть первичными или вторичными. Замер берётся за значение признака. Можно выделить шесть типов признаков:

К первому типу относится прямое измерение, т.е. измерение с использованием приборов (например, измерение длины стола линейкой, измерение скорости машины спидометром, измерение температуры воздуха градусником, измерение силы тока амперметром, измерение глубины моря тахометром и т.д.) или при помощи счета (например, сосчитать количество книг на полке, количество фруктов в ящике, количество рыб в аквариуме и т.д.).

Ко второму типу относится прямое измерение с последующим аналитическим преобразованием, зависящим от параметров (они вносят случайный разброс в значение). Это измерение подразделяется на одноуровневое, т.е. измерение на объекте и двухуровневое – на группе объектов (например, измерение дозы облучения – человека помещают в некоторую камеру, где одновременно измеряется его вес, количество радиационных частиц, содержащихся в нем и получают представление о дозе внутреннего облучения).

К третьему типу относится аналитическая комбинация: $\sum x_i / n = \bar{x}$ нескольких первого типа или нескольких первого и второго типов (характеристика группы людей – имеется некоторое количество детей в группе, известен их вес, рост, нужно определить средние характеристики по группе, например, средний вес, процент девочек в группе).

К четвёртому типу относится прямая экспертная оценка (например, уровень подготовленности студента, пригодность продуктов для употребления, возможность использования природных ресурсов и т.д.).

К пятому – прямая экспертная оценка с последующим аналитическим преобразованием (например, в зависимости от компетентности эксперимента, т.е. от степени доверия к оценке, полученной экспертом, получается результирующая оценка путём умножения исходной оценки на некоторый коэффициент, который является функцией от компетентности).

К шестому – аналитическая комбинация экспертных оценок (например, берётся несколько экспертных оценок и у каждой есть своя компетентность, и вычисляется средняя оценка).

10.3 Типы шкал

Интегрированная информация о шкалах приведена в таблице 10.1.

Таблица 10.1 – Интегрированная информация о шкалах

Наименование шкалы	Множество допустимых преобразований F(x)	Отношения, отвечающие шкале	Допустимые числовые операции с измерениями	Примеры измерения
Качественная шкала				
Наименований (номинальная, классификационная)	Взаимно-однозначные	Эквивалентность	Сравнения: x=y, x<>y	Национальность, пол, профессия, вид оплаты труда
Порядковая (ранговая, ординальная)	Монотонно-неубывающие функции	Квазипорядок (нестрогая ранжировка)	Сравнения: x<=y	В строгом смысле примеров шкалы нет. Условно: шкала твердости минералов, экспертные ранжировки, оценки предпочтений
Количественная шкала				
Разностей (балльная)	F(x)=d+x	Аддитивное метризованное	Сравнения: x-y<=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, x*y, x+y, x-y	Температура по Кельвину, возраст, производительность труда

Шкалы. Отображение $\Psi: A \to R^1$, называется шкалой *наименований*, если его допустимым преобразованием является взаимно однозначное отображение $\eta_1: \Psi(A) \to R^1$. Шкальные значения играют роль имен объектов. Здесь определено отношение равенства, которое соответствует отношению эквивалентности. Оно индуцирует на А разбиение на непересекающиеся классы. Эти признаки называют классификационными или номинальными. Примеры: профессия, национальность, пол, место рождения.

Отображение $\Psi: A \to R^1$ называется шкалой *порядка*, если его допустимым преобразованием является монотонно возрастающее непрерывное отображение $\eta_2 : \Psi(A) \to R^1$. Определены отношения равенства и порядка. Первое соответствует эквивалентности объектов, второе - порядку. Отношение эквивалентности индуцирует разбиение А на классы, а отношение порядка задает линейный порядок на множестве классов эквивалентности. Соответствующее отношение порядка задает порядок на множестве различных значений признака x, которые называются градациями шкалы порядка. Эти признаки называют порядковыми или ординальными. В строгом смысле примеров шкалы нет. Условно примерами шкалы являются: сила ветра в баллах, образование, оценка на экзамене, шкала твердости минералов.

Отображение $\Psi(A) \to R^1$ называется *количественной шкалой*: а) интервалов; б) отношений; в) разностей; г) абсолютной, если допустимым преобразованием является положительное линейное преобразование вида:

$$\eta_3 : \psi(A) \to \alpha\psi(A) + \beta,$$

где для каждого подвида количественной шкалы а) $\alpha \in R^+, \beta \in R^1$; б) $\alpha \in R^+, \beta = 0$; в) $\alpha = 1, \beta \in R^1$; г) $\alpha = 1, \beta = 0$. Примеры: а) любые показатели, значение которых может быть отрицательным: температура по Цельсию, летоисчисление, убытки - прибыль; б) возраст, вес, длина; в) квалификационные разряды, балльные оценки; г) количество элементов некоторого множества, адрес в памяти ЭВМ.

10.4 Унификация типа переменных

Одна из сложностей автоматизированного анализа информации заключается в том, что среди признаков могут быть количественные и качественные (порядковые или классификационные), а большинство методов статистической обработки предполагают их однотипность. Поэтому и возникает вопрос об унификации записи единичного наблюдения.

1-й вариант решения. Наблюдение представляют в виде вектора размерности $m_1 + \ldots + m_p$, m_k - число градаций (интервалов группирования, уровней качества или однородных групп) признака x. Компоненты этого вектора принимают значение 0 или 1. Недостатки: субъективизм в выборе способов разбиения диапазонов количественных признаков, потеря информативности при переходе от индивидуальных к групповым значениям.

2-й вариант. Преобразование качественных переменных в количественные с помощью «оцифровки» (шкалирование).

3-вариант. Сведение классификационных и количественных данных к порядковым.