Министерство образования и науки Республики Казахстан
Костанайский государственный университет имени А. Байтурсынова

Кафедра программного обеспечения

Иванова И.В.

ТЕХНОЛОГИИ BIG DATA И АНАЛИЗ БОЛЬШИХ ДАННЫХ

Учебно-методическое пособие

Костанай, 2019

УДК 004.4
ББК 32.973.26-018.2я73
И 21

Составитель:
Иванова Ирина Владимировна, к.п.н., доцент кафедры программного обеспечения

Рецензенты:
Жунусов Куат Маратович, к.э.н., доцент кафедры информационных технологий Костанайского инженерно - экономического университета имени М.Дулатова
Салыкова Ольга Сергеевна, к.т.н., доцент кафедры программного обеспечения Костанайского государственного университета имени А. Байтурсынова
Исмаилов Арман Оразалиевич, к.т.н., доцент кафедры программного обеспечения Костанайского государственного университета имени А. Байтурсынова

Иванова, И.В.,
И 21 Технологии Big Data и анализ больших данных. Учебно-методическое пособие. – Костанай: КГУ имени А.Байтурсынова, 2019. - 98 с.

ISBN 978-601-7985-47-9

Учебно-методическое пособие охватывает содержание дисциплины «Технологии Big Data и анализ больших данных», которая читается докторантам специальности 6D070400 – Вычислительная техника и программное обеспечение. Учебно-методическое пособие включает в себя теоретический и практический материал для подготовки докторантов к работе с большими данными. Знания, полученные в результате освоения дисциплины, помогут при сборе и анализе огромных объемов структурированной или неструктурированной информации, при разработке моделей.

УДК 004.4

ББК 32.973.26-018.2я73
И 21

Утверждено и рекомендовано к изданию Учебно-методическим советом Костанайского государственного университета имени А.Байтурсынова от 29.04.2019 г. протокол № 3

ISBN 978-601-7985-47-9

СОДЕРЖАНИЕ

Тема1 Основные определения, термины, задачи анализа больших данных

Цель: рассмотреть вопросы безопасности BIG DATA; понятие Data Mining; задачи анализа больших данных.

План:

1.1 Вопросы безопасности BIG DATA

1.2 Понятие Data Mining

1.3 Задачи анализа больших данных

1.1 Вопросы безопасности BIG DATA

Традиционных механизмов безопасности, таких как брандмауэры и антивирусное программное обеспечение, устанавливаемое на компьютерах, недостаточно для эффективной защиты больших данных. Проблема состоит в том, что все это создавалось для защиты небольших объемов статической информации – файлов, сохраненных на жестких дисках, а не большого информационного потока, прибывающего из облака. Меры безопасности должны быть достаточно гибкими и оперативными, что позволит обеспечить бесперебойность получения данных и безопасность многочисленных «точек входа».

Безопасность вычисления в распределенных программных системах, которые выполняют несколько этапов вычислений, должно быть несколько уровней защиты: как минимум, один собственно для программ, другой для защиты данных от этих программ.

Безопасность нереляционных баз данных называемые на языке профессионалов NoSQL, активно развиваются. Поэтому необходимо развивать вместе с ними и соответствующие меры безопасности.

Безопасное хранение данных. В прошлом IT-менеджеры контролировали перемещение данных, но в эпоху big data ручное управление этими процессами является несостоятельным. Автоматическое перемещение данных по уровням требует дополнительных механизмов безопасности.

Проверка достоверности, как это происходит в случае сбора больших данных, получает миллионы вводных, надо убедиться, что каждый бит информации заслуживает доверия и является актуальным.

Мониторинг безопасности в режиме реального времени. Пока real-time security не может похвастать высоким уровнем выявления реальных угроз, потому специалисты получают тысячи ложных срабатываний системы ежедневно.

Data mining и аналитика сохраняют конфиденциальность. Big data – это возможность интенсивного сбора информации приватного характера без уведомления или согласия потребителей.

Шифрование управления доступом и обеспечение безопасности соединений, для полной безопасности данные должны быть зашифрованы от начала до конца, но при этом они должны оставаться эффективными и доступными для тех, кто в них нуждается.

Фрагментарный контроль доступа. Не все данные одинаково конфиденциальны, и компании должны иметь возможность сегментировать их по уровню секретности. Это даст возможность делиться максимальным количеством информации, сохраняя засекреченными только наиболее важные сведения.

Подробные аудиты. Чтобы узнать о нарушениях в системе безопасности, необходимы подробные аудиты. Однако в силу огромного объема больших данных такая отчетность должна соответствовать масштабу того или иного инцидента.

Происхождение данных. Решения безопасности обязательно должны быть направлены на то, чтобы контролировать и отслеживать источники, из которых приходят данные.

Предложения по повышению эффективности безопасности BIG DATA

У Говинда Раммурти, занимающего пост исполнительного директора компании eScan MicroWorld, есть несколько рекомендаций, которые помогут обеспечить безопасность данных. Все сводится к тому, что нужно фокусироваться на безопасности ресурсов и приложений, а не устройств, изолировать критически важные устройства и серверы, внедрять SIEM (средства управления информацией и событиями информационной безопасности) в режиме реального времени, а также обеспечивать баланс реактивной и проактивной защиты.

MicroWorld – разработчик передовых решений информационной безопасности, которые обеспечивают комплексную защиту от усложняющихся компьютерных угроз. Портфель продуктов компании включает решения MailScan и eScan, которые удостоены наград наиболее авторитетных лабораторий сравнительного тестирования, включая Virus Bulletin, West Coast Labs (Checkmark), AV-Comparatives, PCSL и ICSA labs.

Эксперты по облачным технологиям считают, что самым разумным проводником в вопросах улучшения безопасности больших данных является антивирусная индустрия. На протяжении десятилетий антивирусное программное обеспечение ведет борьбу с различными видами угроз. Есть множество поставщиков антивирусного ПО, предлагающих самые разные решения. И все они могут оказаться полезными, когда речь заходит о неприятных цифровых ошибках или серьезных угрозах.

К тому же эксперты всего мира высоко оценивают открытость антивирусной индустрии в отношении данных. Вместо блокировки своих секретов безопасности для получения конкурентного преимущества, производители антивирусного программного обеспечения (в том числе неправительственные организации, государственные учреждения, и даже частные предприятия) свободно обмениваются друг с другом данными об угрозах. Лидеры отрасли могут сотрудничать, чтобы бороться с новыми и опасными вредоносными программами во всем мире, обеспечивая максимальную безопасность big data. Такой своеобразный либерализм и отсутствие разрушительной конкуренции – именно то, что нужно для построения мощной и эффективной системы безопасности больших данных.

Такие организации, как Cloud Security Alliance, пытаются сотрудничать с игроками рынка ради защиты облачных систем. Однако пока индустрии не хватает взаимодоверия, чтобы обеспечить реальный прогресс в этом вопросе.

1.2 Понятие Data Mining

Сегодня на рынке представлено множество инструментов, включающих различные методы, которые делают Data Mining прибыльным делом, все более доступным для большинства компаний.

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, «извлечение зерен знаний из гор данных», раскопка знаний в базах данных, информационная проходка данных,

«промывание» данных. Понятие «обнаружение знаний в базах данных» (knowledge discovery in databases, KDD) можно считать синонимом Data Mining.

Понятие Data Mining, появившееся в 1978 г., приобрело высокую популярность в современной трактовке примерно с первой половины 90-х годов. До этого времени обработка и анализ данных осуществлялись в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных. О популярности Data Mining говорит и тот факт, что результат поиска термина «Data Mining» в поисковой система Google (на сентябрь 2005 года) – более 18 миллионов страниц.

Что же такое Data Mining?

Data Mining – мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др. (рисунок 1).

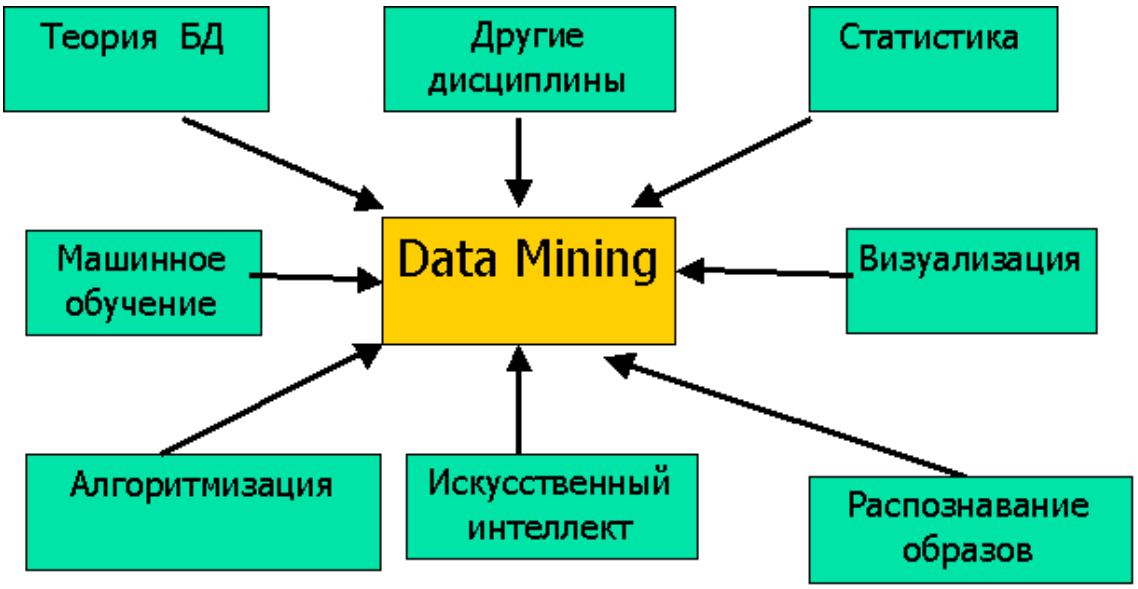


Рисунок 1 - Data Mining как мультидисциплинарная область

Понятие статистики

Статистика – это наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению.

Статистика является совокупностью методов планирования эксперимента, сбора данных, их представления и обобщения, а также анализа и получения выводов на основании этих данных.

Статистика оперирует данными, полученных в результате наблюдений либо экспериментов. Одна из последующих глав будет посвящена понятию данных.

Понятие машинного обучения

Единого определения машинного обучения на сегодняшний день нет. Машинное обучение можно охарактеризовать как процесс получения программой новых знаний. Митчелл в 1996 году дал такое определение: «Машинное обучение – это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы». Одним из наиболее популярных примеров алгоритма машинного обучения являются нейронные сети.

Понятие Искусственный интеллект

Искусственный интеллект – научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования видов человеческой деятельности, традиционно считающихся интеллектуальными. Термин интеллект (intelligence) происходит от латинского intellectus, что означает ум, рассудок, разум, мыслительные способности человека. Соответственно, искусственный интеллект (AI, artificial intelligence) толкуется, как свойство автоматических систем брать на себя отдельные функции интеллекта человека. Искусственным интеллектом называют свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека. Каждое из направлений, сформировавших Data Mining, имеет свои особенности. Проведем сравнение с некоторыми из них.

Сравнение статистики, машинного обучения и Data Mining:

Статистика

- Более, чем Data Mining, базируется на теории.
- Более сосредотачивается на проверке гипотез.

Машинное обучение

- Более эвристично.
- Концентрируется на улучшении работы агентов обучения.

Data Mining

- Интеграция теории и эвристик.
- Сконцентрирована на едином процессе анализа данных, включает очистку данных, обучение, интеграцию и визуализацию результатов.

Понятие Data Mining тесно связано с технологий баз данных и понятием данные.

Развитие технологии баз данных:

1960 гг.

В 1968 году была введена в эксплуатацию первая промышленная СУБД система IMS фирмы IBM.

1970 гг.

В 1975 году появился первый стандарт ассоциации по языкам систем обработки данных – Conference of Data System Languages (CODASYL), определивший ряд фундаментальных понятий в теории систем баз данных, которые и до сих пор являются основополагающими для сетевой модели данных. В дальнейшее развитие теории баз данных большой вклад был сделан американским математиком Э.Ф.Коддом, который является создателем реляционной модели данных.

1980 гг.

В течение этого периода многие исследователи экспериментировали с новым подходом в направлениях структуризации баз данных и обеспечения к ним доступа. Целью этих поисков было получение реляционных прототипов для более простого моделирования данных. В результате, в 1985 году был создан язык, названный SQL. На сегодняшний день практически все СУБД обеспечивают данный интерфейс.

1990 гг.

Появились специфичные типы данных «графический образ», «документ», «звук», «карта». Типы данных для времени, интервалов времени, символьных строк с двухбайтовым представлением символов были добавлены в язык SQL. Появились технологии Data Mining, хранилища данных, мультимедийные базы данных и Web- базы данных. Возникновение и развитие Data Mining обусловлено различными факторами, основные среди них:

- совершенствование аппаратного и программного обеспечения;
- совершенствование технологий хранения и записи данных;
- накопление большого количества ретроспективных данных;
- совершенствование алгоритмов обработки информации.

Понятие Data

Mining Data Mining – это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации).

Технологию Data Mining достаточно точно определяет Григорий Пиатецкий-Шапиро (Gregory Piatetsky-Shapiro) – один из основателей этого направления: Data Mining – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Неочевидных – это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных – это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных – это значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

Знания – совокупность сведений, которая образует целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д.

Использование знаний (knowledge deployment) означает действительное применение найденных знаний для достижения конкретных преимуществ (например, в конкурентной борьбе за рынок).

Приведем еще несколько определений понятия Data Mining. Data Mining – это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования.

Data Mining – это процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (определение SAS Institute).

Data Mining – это процесс, цель которого – обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образцов плюс применение статистических и математических методов (определение Gartner Group).

В основу технологии Data Mining положена концепция шаблонов (patterns), которые представляют собой закономерности, свойственные подвыборкам данных, кои могут быть выражены в форме, понятной человеку.

«Mining» по-английски означает «добыча полезных ископаемых», а поиск закономерностей в огромном количестве данных действительно сродни этому процессу.

Цель поиска закономерностей – представление данных в виде, отражающем искомые процессы. Построение моделей прогнозирования также является целью поиска закономерностей.

Перспективы технологии Data Mining

Потенциал Data Mining дает «зеленый свет» для расширения границ применения этой технологии. Относительно перспектив Data Mining возможны следующие направления развития:

– выделение типов предметных областей с соответствующими им эвристиками, формализация которых облегчит решение соответствующих задач Data Mining, относящихся к этим областям;

– создание формальных языков и логических средств, с помощью которых будет формализованы рассуждения и автоматизация которых станет инструментом решения задач Data Mining в конкретных предметных областях;

– создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать некие теории, опирающиеся на эмпирические данные;

– преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

Если рассматривать будущее Data Mining в краткосрочной перспективе, то очевидно, что развитие этой технологии наиболее направлено к областям, связанным с бизнесом. В краткосрочной перспективе продукты Data Mining могут стать такими же обычными и необходимыми, как электронная почта, и, например, использоваться пользователями для поиска самых низких цен на определенный товар или наиболее дешевых билетов.

В долгосрочной перспективе будущее Data Mining является действительно захватывающим, это может быть как поиск интеллектуальными агентами новых видов лечения различных заболеваний, так и нового понимания природы вселенной.

Однако Data Mining таит в себе и потенциальную опасность, ведь все большее количество информации становится доступным через всемирную сеть, в том числе и сведения частного характера, и все больше знаний возможно добыть из нее…

Например, крупнейший онлайновый магазин «Amazon» оказался в центре скандала по поводу полученного получения им патента «Методы и системы помощи пользователям при покупке товаров», который представляет собой не что иное, как очередной продукт Data Mining, предназначенный для сбора персональных данных о посетителях магазина. Новая методика позволяет прогнозировать будущие запросы на основании фактов покупок, а также делать выводы об их назначении. Цель данной методики – то о чем говорилось выше – получение как можно большего количества информации о клиентах, в том числе и частного характера (пол, возраст, предпочтения и т.д.). Таким образом, собираются данные о частной жизни покупателей магазина, а также членах их семей, включая детей. Последнее запрещено законодательством многих стран, сбор информации о несовершеннолетних возможен там только с разрешения родителей.

Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии. Области, где применения технологии Data Mining, скорее всего, будут успешными, имеют такие особенности:

•требуют решений, основанных на знаниях;

•имеют изменяющуюся окружающую среду;

•имеют доступные, достаточные и значимые данные;

•обеспечивают высокие дивиденды от правильных решений.

Существующие подходы к анализу

Достаточно долго дисциплина Data Mining не признавалась полноценной самостоятельной областью анализа данных, иногда ее называют «задворками статистики» (Pregibon, 1997).

На сегодняшний день определилось несколько точек зрения на Data Mining. Сторонники одного из них считают эту технологию миражом, отвлекающим внимание от классического анализа данных. Сторонники другого направления – это те, кто принимает Data Mining как альтернативу традиционному подходу к анализу. Есть и середина, где рассматривается возможность совместного использования современных достижений в области Data Mining и классическом статистическом анализе данных.

Технология Data Mining постоянно развивается, привлекает к себе все больший интерес, как со стороны научного мира, так и со стороны применения достижений технологии в бизнесе.

1.3 Задачи анализа больших данных

Задачи, связанные с анализом данных, возникали в самых разных областях исследования задолго до того, как появился сам термин. Но именно благодаря быстрому развитию компьютерной техники, появлению сети Интернет (в современном понимании этого термина) в конце 1980-х – начале 1990-х годов, сделавшему возможным сбор, хранение, передачу и обработку больших объёмов данных, анализ данных сформировался как самостоятельное научное направление. Термин «Анализ данных», или «Интеллектуальный анализ данных» - перевод английского термина «Data Mining», т.е. буквально «добыча данных» или даже «раскапывание данных». Сам термин «Data Mining», (а также термин «Knowledge Discovery in Data», KDD) был предложен в 1991 году Григорием Пятецким-Шапиро, выпускником Нью-Йоркского университета, который заинтересовался вопросом: «Возможно, ли автоматически находить правила, которые позволили бы ускорить выполнение запросов к большим базам данных?».

По определению Г.Пятецкого-Шапиро, «Data Mining — это процесс обнаружения знаний (в сырых данных), которые являлись бы:

- неизвестными ранее,
- нетривиальными,
- доступными для интерпретации;
- практически полезными;

необходимыми для принятия решений в различных сферах человеческой деятельности».

Определение понятия «Анализ данных» тесно связано с классификацией уровней информации (таблица 1):

Таблица 1 – Классификация уровней информации

Уровень информации	Описание
Сырые данные (raw data)	Необработанные данные, получаемые в результате наблюдения за объектами и отображающие их состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год, данные о ценах на рынке жилья, данные об абитуриентах, зачисленных на 1 курс)
Информация	Это либо: - сырые данные, но систематизированные, представленные в более компактном виде (например, результаты поиска – сведения об абитуриентах, поступивших в КГУ в этом году); - обработанные данные, имеющие информационную ценность для пользователя (например, сводные статистические характеристики – средний балл абитуриентов, поступивших в КГУ в этом году – его абсолютная величина и % по отношению к тому же показателю за предыдущий год).
Знания	Понятие «знания» включает: - скрытые взаимосвязи между объектами (признаками объектов); - некоторое ноу-хау, алгоритмы, методы решения задач. Знания обладают практической ценностью: «Знание – сила!» // Фрэнсис Бэкон.

В настоящее время выделяют следующие основные классы задач анализа данных:

Таблица 2- Основные классы задач анализа данных

Класс задач	Описание, примеры
Прогнозирование (Forecasting)	Нахождение будущих состояний объекта на основании предыдущих состояний (исторических данных). Примеры: - прогнозирование ситуаций на валютных рынках, - прогнозирование цен на рынке недвижимости, - прогнозирование демографических процессов, - прогнозирование климатических процессов….
Классификация (Classification)	Нахождение правила, позволяющее отнести объект к тому или иному классу (выбрать класс из числа известных заранее классов) на основе информации о том, к какому классу относятся другие объекты. Примеры: - Задачи распознавание образов (распознавание рукописного текста, фотографии (например, определение номера автомобиля по фото), идентификация личности по фото, голосу, видео…); - Задачи атрибуции (определение авторства / периода создания / страны происхождения …произведений искусства, археологических находок); - Задачи диагностики (в медицине и технике)
Кластеризация (Clusterization)	Нахождение правила для автоматического разделения имеющихся объектов на классы на основании сходства тех или иных характеристик (факторов) этих объектов. При этом ни сами классы, ни их количество заранее неизвестны. Примеры: - Сегментация рынка (разделение всех потенциальных потребителей на кластеры для последующего целевого воздействия, например, создания целевой рекламы); - Задачи разбиения множества индивидов на группы кластеры (в социологии, психологии, биологии и пр…)
Ассоциация (Associations)	Поиск устойчивых закономерностей между случайными событиями, наступающими одновременно. Пример: - Анализ покупательской корзины – поиск «устойчивых связей в корзине покупателя» (осуществляется с целью учёта их при планировании расположения отделов в супермаркете).
Последовательность (Последовательная ассоциация, Нахождение последовательных шаблонов) (Sequence, Sequential association, Sequential pattern)	Поиск устойчивых закономерностей между случайными событиями, связанными во времени, т.е. правил вида: после события X через время t происходит событие Y. Пример: - После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (Customer Lifecycle Management).
Визуализация данных (Data Visualization)	Графическое изображение данных (2D и 3D диаграммы, гистограммы, графики, облака точек…)
Анализ отклонений (Deviation Detection)	Обнаружение и анализ данных, наиболее отличающихся от общего множества данных. Примеры: - выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы; выявление мошенничества с кредитными карточками.

Таким образом, названные задачи анализа данных возникают в самых разных областях, в частности, в таких как:

- Розничная торговля
- Банковское дело
- Страхование
- Телекоммуникации
- Техника
- Медицина
- Молекулярная биология
- Молекулярная генетика
- Хемоинформатика

Рекомендуемая литература:

1. Введение в Data Mining [Электронный ресурс]/Режим доступа: http://files.pilotlz.ru/pdf/cB819-2-ch.pdf
2. Введение. Основные задачи анализа данных. [Электронный ресурс] / Режим доступа: https://edu.kpfu.ru/pluginfile.php/
3. Вопросы безопасности BIG DATA. [Электронный ресурс] / Режим доступа: http://rtbinsight.ru/articles/big-data-security.html

Тема 2 Методики сбора данных

Цель: провести обзор источников информации и методик анализа для Big Data

План:

2.1 Обзор источников информации для Big Data
2.2 Обзор методик анализа больших данных
2.3 Методики анализа больших данных

2.1 Обзор источников информации для Big Dat

20 свободных источников данных

- Data.gov
- US Census Bureau
- European Union Open Data Portal
- Data.gov.uk
- The CIA World Factbook
- Healthdata.gov
- NHS Health and Social Care Information Centre
- Amazon Web Services public datasets
- Facebook Graph
- Gapminder
- Google Trends
- Google Finance
- Google Books Ngrams
- National Climatic Data Center
- DBPedia
- Topsy
- Likebutton

- New York Times
- Freebase
- Million Song Data Set

2.2 Обзор методик анализа больших данных

Существует множество разнообразных методик анализа массивов данных, в основе которых лежит инструментарий, заимствованный из статистики и информатики (например, машинное обучение). Список не претендует на полноту, однако в нем отражены наиболее востребованные в различных отраслях подходы. При этом следует понимать, что исследователи продолжают работать над созданием новых методик и совершенствованием существующих. Кроме того, некоторые из перечисленных методик вовсе не обязательно применимы исключительно к большим данным и могут с успехом использоваться для меньших по объему массивов (например, A/B-тестирование, регрессионный анализ). Безусловно, чем более объемный и диверсифицируемый массив подвергается анализу, тем более точные и релевантные данные удается получить на выходе.

Рассмотрим наиболее часто используемые методики (таблица 3).

Таблица 3 - Методики анализа больших данных

Методика	Описание методики
A/B testing	Методика, ориентированная на поочередное сравнение контрольной выборки с другими. При этом выявляется наилучшая комбинация показателей, для достижения, скажем, желаемой ответной реакции потребителей на маркетинговое предложение. Статистическая достоверность результата достигается выполнением многочисленных итераций.
Association rule learning	Совокупность методик, направленных на выявление взаимосвязей, или ассоциативных правил между переменными величинами в больших массивах данных (применяется в Data Mining).
Classification	Совокупность методик, позволяющих спрогнозировать поведение потребителей в определенном сегменте рынка (применяется в Data Mining).
Cluster analysis	Метод классификации объектов по группам с выявлением заранее неизвестных общих признаков (применяется в Data Mining).
Crowdsourcing	Методика сбора данных их большого количества источников.
Data fusion and data integration	Совокупность методик, ориентированных на анализ комментариев пользователей социальных сетей и на сравнение их с результатами продаж в режиме реального времени.
Data mining	Совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. С помощью этих методов могут определяться категории потребителей, наиболее восприимчивых для продвигаемого продукта или услуги, выявляться качества успешных работников, прогнозироваться поведенческая модель потребителей.
Ensemble learning	Метод располагает широким спектром предикативных моделей, что делает прогнозирование максимально эффективным.
Genetic algorithms	Для данного метода характерно, что возможные решения выступают в виде «хромосом», способных комбинироваться и мутировать (по аналогии

	с процессом естественной эволюции выживает лишь наиболее адаптировавшаяся особь).
Machine learning	Направление, ориентированное на создание алгоритмов самообучения на основе эмпирических данных, - искусственный интеллект.
Natural language processing	Совокупность методик распознавания естественного языка человека, заимствованных из информатики и лингвистики.
Network analysis	Совокупность методик для анализа связей между узлами в сетях. В социальных сетях дает возможность исследовать взаимосвязи между отдельными пользователями, компаниями, сообществами и др.
Optimization	Совокупность численных методов для редизайна сложных систем и процессов, направленная на улучшение одного или нескольких показателей, поддерживающая принятие стратегических решений.
Pattern recognition	Совокупность методик с элементами самообучения для прогнозирования поведенческой модели потребителей.
Predictive modeling	Совокупность методик, нацеленных на создание математической модели, предваряющей заданный вероятный сценарий развития событий.
Regression	Совокупность статистических методов для обнаружения закономерности между изменением зависимой переменной и одной или несколькими независимыми. Применяется в прогнозировании и Data Mining.
Sentiment analysis	Методика оценки настроений потребителей, основанная на технологии распознавания естественного языка человека. Дает возможность выделить из общего информационного потока сообщения, относящиеся к интересующим предметам, и оценить полярность суждения.
Signal processing	Совокупность методик, взятая из радиотехники, направленная на распознавание сигнала на фоне шума и его последующий анализ.
Spatial analysis	Ряд методик (часть которых заимствована из статистики) анализа пространственных данных – топологии местности, географических координат, геометрии объектов. В данном случае источником больших данных являются геоинформационные системы.
Statistics	Наука о сборе, организации и интерпретации данных. Статистические методы нередко используются для оценочных суждений о взаимосвязях между различными событиями.
Supervised learning	Совокупность методик, базирующаяся на технологиях машинного обучения, для определения функциональных взаимосвязей в исследуемых массивах данных.
Simulation	Методики моделирования поведения сложных систем, основное назначение которых – прогнозирование и проработка возможных сценариев при планировании.
Time series analysis	Совокупность методов анализа (почерпнутых из статистики и цифровой обработки сигналов) повторяющихся последовательностей данных. Как правило, применяется для мониторинга рынка ценных бумаг или заболеваемости пациентов.
Unsupervised learning	Совокупность методик, базирующаяся на технологиях машинного обучения, для определения функциональных взаимосвязей в исследуемых массивах данных (прослеживается некоторая аналогия с Cluster Analysis).
Visualization	Ряд методов графического представления результатов анализа больших данных (используется для более легкой их интерпретации).

2.3 Методики анализа больших данных

При анализе информации часто сталкиваемся с тем, что теоретическое великолепие методов анализа разбивается о действительность. Ведь вроде все давно решено, известно множество методов решения задач анализа. Почему же довольно часто они не работают?

Дело в том, что безупречные с точки зрения теории методы имеют мало общего с действительностью. Чаще всего аналитик сталкивается с ситуацией, когда трудно сделать какие-либо четкие предположения относительно исследуемой задачи. Модель не известна, и единственным источником сведений для ее построения является таблица экспериментальных данных типа «вход – выход», каждая строка которой содержит значения входных характеристик объекта и соответствующие им значения выходных характеристик.

В результате они вынуждены использовать всякого рода эвристические или экспертные предположения и о выборе информативных признаков, и о классе моделей, и о параметрах выбранной модели. Эти предположения аналитика основываются на его опыте, интуиции, понимании смысла анализируемого процесса. Выводы, получаемые при таком подходе, базируются на простой, но фундаментальной гипотезе о монотонности пространства решений, которую можно выразить так: «Похожие входные ситуации приводят к похожим выходным реакциям системы». Идея на интуитивном уровне достаточно понятная, и этого обычно достаточно для получения практически приемлемых решений в каждом конкретном случае.

В результате применения такого метода решений академическая строгость приносится в жертву реальному положению вещей. Собственно, в этом нет ничего нового. Если какие – то подходы к решению задачи вступают в противоречие с реальностью, то обычно их изменяют. Возвращаясь к анализу данных, или, вернее, к тому, что сейчас называют Data Mining, следует обратить внимание еще на один момент: процесс извлечения знаний из данных происходит по той же схеме, что и установление физических законов: сбор экспериментальных данных, организация их в виде таблиц и поиск такой схемы рассуждений, которая, во-первых, делает полученные результаты очевидными и, во-вторых, дает возможность предсказать новые факты. При этом имеется ясное понимание того, что наши знания об анализируемом процессе, как и любом физическом явлении, в какой – то степени приближение. Вообще, всякая система рассуждений о реальном мире предполагает разного рода приближения. Фактически термин Data Mining – это попытка узаконить физический подход в отличие от математического к решению задач анализа данных. Что же мы вкладываем в понятие «физический подход»?

Это такой подход, при котором аналитик готов к тому, что анализируемый процесс может оказаться слишком запутанным и не поддающимся точному анализу с помощью строгих аналитических методов. Но можно все же получить хорошее представление о его поведении в различных обстоятельствах, подходя к задаче с различных точек зрения, руководствуясь знанием предметной области, опытом, интуицией и используя различные эвристические подходы. При этом мы движемся от грубой модели ко все более точным представлениям об анализируемом процессе. Слегка перефразировав Р. Фейнмана, скажем так: можно идеально изучить характеристики анализируемой системы, стоит только не гнаться за точностью.

Общая схема работы при этом выглядит следующим образом:

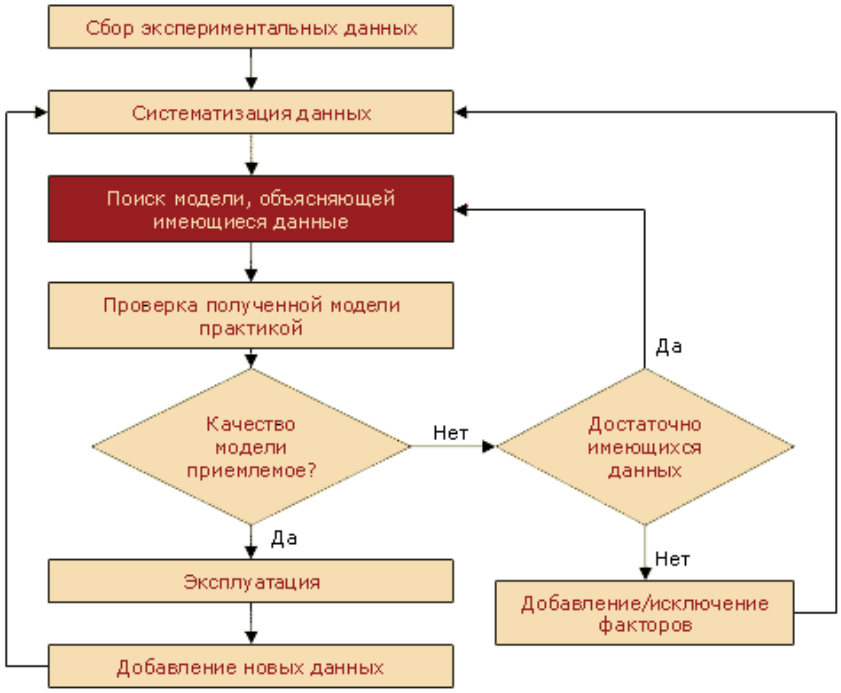


Рисунок 2 – Общая схема

Таким образом, данный подход подразумевает, что:

1. При анализе нужно отталкиваться от опыта эксперта.
2. Необходимо рассматривать проблему под разными углами и комбинировать подходы.
3. Не стоит стремиться сразу к высокой точности. Двигаться к решению нужно от более простых и грубых моделей ко все более сложным и точным.
4. Стоит останавливаться как только получим приемлемый результат, не стремясь получить идеальную модель.
5. По прошествии времени и накоплению новых сведений нужно повторять цикл – процесс познания бесконечен.

Пример работы

В качестве примера можно в общих чертах рассмотреть процесс анализа рынка недвижимости в г. Москве. Цель – оценка инвестиционной привлекательности проектов. Одна из задач, решаемых при этом, – построение модели ценообразования для жилья в новостройках, другими словами, количественную зависимость цены жилья от ценообразующих факторов. Для типового жилья таковыми, в частности, являются:

- местоположение дома (престижность района; инфраструктура района; массовая или точечная застройка; окружение дома (напр. нежелательное соседство с промышленными предприятиями, «хрущевками», рынками и т.д.); экология района (близость к лесопарковым массивам));
- местоположение квартиры (этаж – первые и последние этажи дешевле; секция – квартиры в торцевых секциях дешевле; ориентация квартиры по сторонам света – северная сторона дешевле; вид из окон).
- Тип дома (самая популярная серия П-44Т).
- Площадь квартиры.
- Наличие лоджий (балконов)
- Стадия строительства (чем ближе к сдаче дома, тем выше цена за кв.м).
- Наличие отделки («черновая» отделка, частичная отделка, под ключ. Большинство новостроек сдаются с черновой отделкой).

- Телефонизация дома.
- Транспортное сообщение (близость к метро, удаленность от крупных магистралей, удобный подъезд, наличие автостоянки около дома (наличие парковочных мест)).
- Кто продает квартиру («из первых рук» (инвестор, застройщик) или посредники (риэлтеры). Риэлтеры, как правило, берут за свои услуги – 3-6%).

Для начала из имеющейся истории продаж мы ограничились данными для одного района Москвы. В качестве входных факторов взяли ограниченный набор характеристик с точки зрения экспертов, очевидно влияющих на продажную цену жилья: серия дома, отделка, этаж (первый, последний, средний), готовность объекта, количество комнат, секция (угловая, обычная), метраж. Выходным значением являлась цена за квадратный метр, по которой продавались квартиры. Получилась вполне обозримая таблица с разумным количеством входных факторов.

На этих данных обучили нейросеть, то есть построили довольно грубую модель. При всей своей приблизительности у нее было одно существенное достоинство: она правильно отражала зависимость цены от учитываемых факторов. Например, при прочих равных условиях квартира в угловой секции стоила дешевле, чем в обычной, а стоимость квартир по мере готовности объекта возрастала. Теперь оставалось ее лишь совершенствовать, делать более полной и точной.

На следующем этапе в обучающее множество были добавлены записи о продажах в других районах Москвы. Соответственно, в качестве входных факторов стали учитываться такие характеристики, как престижность района, экология района, удаленность от метро. Так же в обучающую выборку была добавлена цена за аналогичное жилье на вторичном рынке. Специалисты, имеющие опыт работы на рынке недвижимости, имели возможность в процессе совершенствования модели безболезненно экспериментировать, добавляя или исключая факторы, т. к., напомню, процесс поиска более совершенной модели сводился к обучению нейросети на разных наборах данных. Главное здесь вовремя понять, что процесс этот бесконечен.

Это пример, как нам кажется, довольно эффективного подхода к анализу данных: использование опыта и интуиции специалиста в своей области для последовательного приближения ко все более точной модели анализируемого процесса.

Описанный подход позволяет решать реальные задачи с приемлемым качеством.

Рекомендуемая литература:

1. Большие_данные_(Big_Data) [Электронный ресурс]/Режим доступа: http://www.tadviser.ru/index.php/

2. 20 открытых источников данных. [Электронный ресурс]/Режим доступа: http://igorsubbotin.blogspot.com/2014/09/big-data-20-free-big-data-sources-everyone-should-know.html

Тема 3 Технологии хранения и обработки больших данных

Цель: провести обзор технологий хранения больших данных. Базы данных. Системы управления базами данных. Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных.

План:

3.1 Современные технологии обработки больших данных

3.2 Базы данных.

3.3 Системы для Больших Данных: конвергенция архитектур

3.4 Big Data Система управления большими данными

3.5 Модели данных.

В настоящее время компании создают огромные объемы данных в формате, плохо соответствующем традиционному структурированному формату баз данных, таких как веб-журналы, видеозаписи, текстовые документы, машинные коды или геопространственные данные. Все это хранится во множестве разнообразных хранилищ, зачастую за пределами организации. В результате корпорации могут иметь доступ к огромному объему своих данных и не иметь необходимых инструментов, чтобы установить взаимосвязи между этими данными и сделать на их основе значимые выводы. С учетом того, что данные сейчас обновляются все чаще и чаще, возникает ситуация, в которой традиционные методы анализа информации не могут «угнаться» за огромными объемами постоянно обновляемых данных, что в итоге и открывает дорогу технологиям больших данных.

В современных компаниях различные датчики, видеокамеры, интеллектуальные счетчики и другие подключенные устройства генерируют огромные объемы данных, которые добавляются к уже хранящейся информации. Настоящий предприниматель может разглядеть во всей этой «лавине» данных полезную информацию, однако исследование, проведенное недавно по заказу компании Cisco, показало, что ИТ-специалисты и компании пока с трудом извлекают пользу из поступающей информации. В ходе исследования под названием «Cisco® Connected World Technology Report», проведенного в 18 странах независимой аналитической компанией Insight Express, были опрошены 1800 студентов колледжей и такое же количество молодых специалистов в возрасте от 18 до 30 лет. Опрос проводился для того, чтобы выяснить уровень готовности ИТ-отделов к реализации проектов Big Data, получить представление о связанных с этим проблемах, технологических изъянах, выявить стратегическую ценность таких проектов.

Большинство компаний собирает, записывает и анализирует данные. Тем не менее, согласно отчету, многие компании в связи с Big Data сталкиваются с целым рядом сложных деловых и информационно-технологических проблем. Например, 60 % опрошенных признают, что Big Data могут усовершенствовать процессы принятия решений и повысить конкурентоспособность на мировом рынке, но лишь 28 % заявили о том, что уже получают реальные стратегические преимущества от накопленной информации.

Технологии Big Data может предоставить конкурентные преимущества тем, кто найдет новые, оригинальные способы использования данных. Назовем 5 стран, где специалисты в большей степени склонны полагать, что технологии Big Data станут их конкурентным преимуществом: Китай (90 %); Мексика (85 %); Индия (82 %); Бразилия (79 %); Аргентина (78 %).

Более двух третей опрошенных ИТ-руководителей считают, что в ближайшие 5 лет Big Data станет стратегическим приоритетом в их компаниях. Наибольшую уверенность в этом выразили респонденты из следующих стран: Аргентина (89 %); Китай (86 %); Индия (83 %); Мексика (78 %); Польша (78 %). Что для этого необходимо? 38 % опрошенных заявили, что хотя в их компаниях уже установлены решения Big Data, для раскрытия всех преимуществ им нужен стратегический план. По мнению опрошенных ИТ-руководителей, внедрению Big Data мешает ряд проблем, и прежде всего — обеспокоенность по поводу информационной безопасности, а затем — ограниченные бюджеты и дефицит специалистов. Безопасность данных и управление рисками считают главной проблемой 27 % респондентов. По их мнению, трудности с защитой данных в проектах Big Data связаны с большими объемами информации, разными методами доступа к ней и с малыми бюджетами, выделяемыми на информационную безопасность.

Более всех вопросами информационной безопасности обеспокоены респонденты из: Китая (45 %); Индии (41 %); США (36 %); Бразилии (33 %).

3.1 Современные технологии обработки больших данных

С приходом новых технологий, инструментов и средств коммуникаций, таких, как социальные сети, количество данных, производимых людьми, растет с каждым годом в геометрической прогрессии. Соотношение коэффициента полезности при этом уменьшается. Следовательно, вся генерируемая информация может быть использована для определенных целей только после предварительной и тщательной обработки.

Термин «Big Data» означает большие работы (коллекции, потоки) данных, которые не могут быть обработаны традиционными компьютерными техниками. Этот термин означает не само понятие «большие данные», а предмет исследования, который включает в себя различные инструменты, техники и платформы.

Большие данные включают в себя информацию, генерируемую различными системами и приложениями. Некоторые из сфер, которые попадают под определение «Big Data»:

– черный ящик: информационная составляющая часть вертолета, самолета, морского/космического корабля. Данные подобного рода включают в себя запись голосов экипажа (микрофоны и наушники), информацию о характеристиках объекта управления.

– социальные медиа: включают данные, распространяемые через социальные сети;

– фондовые биржи: хранение информации о сделках купли-продажи между копаниями-партнерами;

– энергосистемы: подобного рода данные содержат информацию о узлах и нагрузках энергетической сети;

– транспортные системы: модели, характеристики, расстояния - все информация о транспорте и дорожных сетях;

– поисковые системы: инженерный поиск информации различны базах данных;

Как следствие, термин «Big Data» включает большое объем, высокую скорость обработки и широкое разнообразие данных и делится на три типа:

– структурные данные – реляционные БД;

– полуструктурированные данные – XML-файлы;

– неструктурированные данные – файлы формата Word, PDF, Text, медиа-журналы.

Большие данные действительно имеют решающее значение для нашей жизни и становятся одной из самых важных технологий в современном мире. Например, использование информации, хранящейся в социальных сетях, маркетинговые агентства изучают обратную связь на свои кампании, акции, и другие рекламные носители. В свою очередь, использую информацию в социальных медиа-системах, таких как предпочтения и восприятие продукта потребителями, компании и розничные организации планируют свое производство. Касательно такой сферы, как медицина, применимость данных о предыдущей истории болезни пациентов способствует обеспечению лучшего и более быстрого обслуживания.

Большие технологии передачи данных играют важную роль в обеспечении детального анализа, который способствует принятию более точных решений, что в свою очередь приводит к повышению эффективности эксплуатации, снижению затрат и снижению рисков для бизнеса. Для использования возможностей больших данных требуется инфраструктура, которая может управлять и обрабатывать огромные объемы структурированных и неструктурированных данных в реальном времени и может защитить конфиденциальность и безопасность данных. Существуют различные технологии на рынке от различных поставщиков, включая такие компании, как Google, IBM, Microsoft, SAP и др.

Структуры хранения

Известно, что эффективные структуры хранения должны иметь иерархическую организацию. Это позволяет выстроить дерево областей памяти, обладающее следующими свойствами:

- Области нижнего уровня образуют область верхнего уровня.
- Области каждого уровня имеют специфические особенности, предназначенные для решения проблем, с которыми не удается справиться на других уровнях.
- На каждом уровне области, как правило, имеют несколько параметров, позволяющих оптимизировать их работу в зависимости от назначения.

Заметим, что терминология, применяемая в различных базах данных, различается существенно. Наша терминосистема ближе всего к применяемой в СУБД Oracle.

В типичном случае (рисунок 3) база данных состоит из одного или нескольких табличных пространств. Каждое такое пространство строится на одном или нескольких файлах данных.

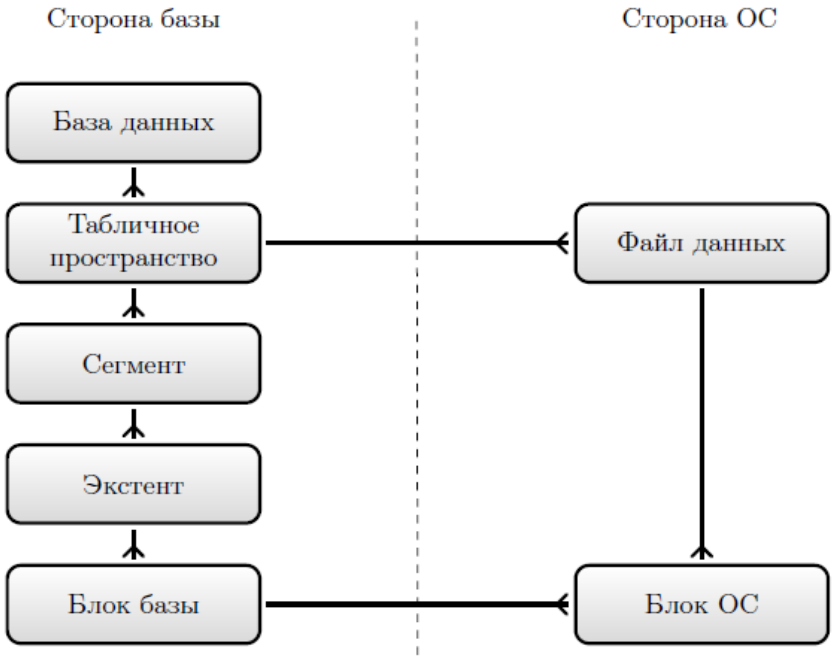


Рисунок 3 - Структуры базы данных

В одно табличное пространство стараются помещать объекты с одинаковым поведением. Например, для словаря базы можно выделить отдельное табличное пространство, обычно называемое системным. Пользовательские данные желательно помещать отдельно от словаря. Это уменьшит вероятность сбоя. Для индексов следует иметь свои табличные пространства. В некоторых СУБД можно отключать отдельные табличные пространства или делать их доступными только по чтению. Типичный пример — табличные пространства для хранения больших объемов очень редко меняющейся справочной информации. Для больших сортировок можно создавать временные табличные пространства, в которых объем данных может резко увеличиваться в размере и так же быстро уменьшаться.

Администратор должен выбрать состав, размеры табличных пространств и определить могут ли они расширяться, и какими порциями им будет предоставляться свободное пространство дисковой памяти.

Табличные пространства состоят из сегментов, содержащих хранимые объекты базы, например, таблицы, индексы, кольцевые буферы отката. Каждому такому объекту положено иметь свой сегмент, куда нет доступа данным других хранимых объектов.

Сегменты состоят из экстентов, представляющих наборы блоков данных базы, расположенных на диске непрерывно. Это ускоряет операции с блоками данных, входящими в состав экстента. Можно, например, при работе с любым элементом данных, читать сразу весь экстент, в надежде, что эти данные скоро понадобятся. Нетрудно догадаться, что сегмент увеличивается или уменьшается на целое число экстентов.

Блок базы, в другой терминологии — страница, — это минимальная единица хранения, которой база данных обменивается с диском. Блок базы образуется из нескольких блоков операционной системы.

Можно задаться вопросом — а почему не использовать блоки операционной системы в качестве блоков данных базы? Дело в том, что современные операционные системы стараются оптимизировать под целый ряд программ, для которых достаточно небольших блоков. Так что добавление больших блоков базы размером до 64 Кбайт, оптимальных для баз данных, неизбежно.

Можно выделить два режима работы базы данных. В первом режиме OLTP (Online Transaction Processing) информационная система использует большой поток транзакций, работающих с небольшими объемами данных.

Обычно это ввод первичных данных, их сохранение и не слишком сложная обработка информации.

Режим OLAP (Online Analytical Processing) используется аналитиками для подготовки сложных отчетов, для анализа информации. Связан с небольшим количеством транзакций, перерабатывающих большие объемы данных.

Установлено, что для работы в режиме OLTP, когда исполняется много сравнительно коротких транзакций, предпочтительнее небольшие блоки размером в 4-12 Кбайт. В режиме OLAP, когда исполняется сравнительно небольшое число длящихся долго транзакций, предпочтительнее блоки больших размеров.

Блоки базы

Для того, чтобы представить, как можно управлять пространством внутри блока, рассмотрим упрощенный вариант блока СУБД Oracle. В нем выделяются заголовок и область для размещения данных. Для блоков большинства сегментов определены два параметра — PCTFREE и PCTUSED (рисунок 5), определяемые в процентах от объема блока без заголовка. Область заголовка может изменяться во время обращения и манипуляций с данными за счет того, что каждая транзакция, обратившаяся к блоку, записывает в его заголовок свой номер SCN и другую информацию.

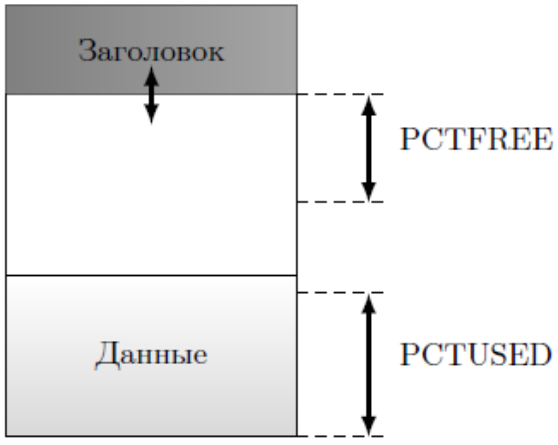


Рисунок 4 - Блок базы

Параметр PCTFREE определяет тот объем незанятого пространства блока, который необходимо оставить для того, чтобы с увеличением длины записей при выполнении инструкций UPDATE они поместились в своем блоке, а не мигрировали в другие блоки. Естественно, возникает вопрос: а как вычислить это значение? Ответ, наверное, не совсем ожидаемый: никак! Просто администратор может экспериментально подобрать некоторое, хорошее для текущего режима работы базы, значение.

В сегменте, выделенном для хранения таблицы, блоки, у которых свободного места меньше, чем PCTFREE, для записи не пригодны. Очевидно, СУБД необходим список блоков, пригодных для записи. Желательно, чтобы этот список был не один, так как транзакции будут конкурировать за доступ к нему при обращении к данным. Конечно, записи во всех таких списках должны быть одинаковыми.

Теперь можно разобраться с назначением второго параметра — PCTUSED, задающего момент включения блока данных в список блоков, пригодных для записи в своем сегменте.

Пусть объем данных в блоке увеличивается от 0 до величины, превышающей PCTUSED, но свободное пространство при этом не меньше PCTFREE. Блок остается в списках блоков, пригодных для записи. Как только свободное пространство станет меньше PCTFREE, блок удалится из всех списков блоков, пригодных для записи. После этого, при увеличении свободного пространства блока до величины большей, чем PCTFREE, он будет оставаться вне списков. И только когда занятое место станет меньше чем PCTUSED, блок вернется в списки блоков пригодных для записи.

3.2 Базы данных

Видов и типов баз данных очень и очень много и описать их все в данной публикации я просто не смогу, но самые распространенные виды хранения информации или виды баз данных я постараюсь описать. Понятно, что база данных хранит в себе информацию о каких-то объектах, например, информацию о товаре в интернет-магазине. Любой товар в базе данных – это объект с какими-то определенными параметрами и свойствами. Перейдем к конкретным примерам.

Иерархическая база данных, структура иерархических баз данных

Иерархическая база данных – каждый объект при таком хранение информации представляется в виде определенной сущности, то есть, у этой сущности могут быть дочерние элементы, родительские элементы, а у тех дочерних могут быть еще дочерние элементы, но есть один объект, с которого все начинается.

Следует сказать, что базы данных подобного вида оптимизированы под чтение информации, то есть, базы данных, имеющие иерархическую структуру умеют очень быстро выбирать, запрашиваемую информацию и отдавать ее пользователям. Но такая структура не позволяет столь же быстро перебирать информацию, тут можно привести пример из жизни, компьютер может легко работать с каким-либо конкретным файлом или папкой (которые, по сути являются объектами иерархической структуры) но проверка компьютера антивирусам осуществляется очень долго. Второй пример – реестр Windows.

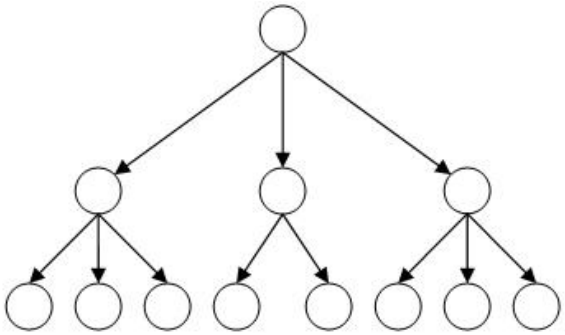


Рисунок 5 – Иерархическая структура базы данных

На рисунке 5 показана структура иерархической базы данных, в самом верху находится родитель или корневой элемент, ниже находятся дочерние элементы, элементы, находящиеся на одном уровне называются братьями, ну или соседними элементами. Соответственно чем ниже уровень элемента, тем вложенность этого элемента больше.

Сетевая база данных, структура сетевых баз данных

Сетевые базы данных, являются своеобразной модификацией иерархических баз данных. Сетевые базы данных отличаются от иерархических тем, что у дочернего элемента может быть несколько предков, то есть, элементов стоящих выше него. Для большей наглядности и понимания структуры сетевых баз данных обратите внимание на рисунок:

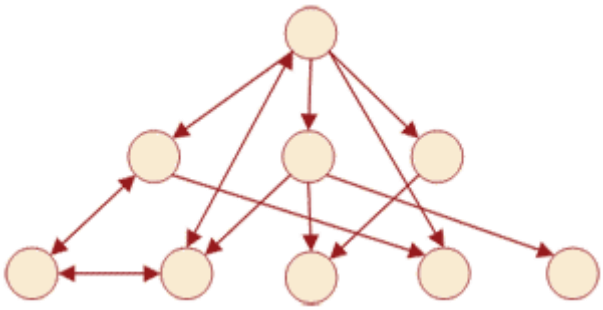


Рисунок 6 - Структура сетевых баз данных

Стоит заметить, что сетевые базы данных обладают примерно теми же характеристиками, что и иерархические базы данных.

Особенности реляционных баз данных

Главной особенностью реляционных баз данных является, то, что объекты внутри таких баз данных хранятся в виде набора двумерных таблиц. То есть, таблица состоит из набора столбцов, в котором может указываться: название, тип данных(дата, число, строка, текст и т.д.). Еще одной важной особенность реляционных БД является, то, что число столбцов фиксировано, то есть, структура базы данных известна заранее, а вот число строк или рядов в реляционных базах данных ничем не ограничено, если говорить грубо, то строки в реляционных базах данных и есть объекты, которые хранятся в базе данных.

На самом деле, базы данных – это абстрактное понятие, таблица – это всего лишь способ хранения информации, набор таблиц может быть связан логически и этот набор называют база данных. Поэтому неправильно говорить, что MySQL это база данных, база данных – это хранящаяся информация. А вот такое понятие, как СУБД – система управления базами данных, это и есть MySQL сервер, именно при помощи него мы управляем хранящимися данными. Или иначе MySQL – это программное воплощение математических идей.

3.3 Системы для больших данных: конвергенция архитектур

Происходящий последние десять лет экспоненциальный рост объемов данных привел к появлению нового класса систем, их обрабатывающих, а на переднем крае здесь оказались онлайн-компании, такие как Google и Amazon, работающие с огромными репозиториями данных.

Системы, работающие с большими объемами данных, традиционно строились на реляционных СУБД, которые по мере роста нагрузки и потребностей в хранении масштабируются в основном вертикально за счет более быстрых процессоров и накопителей. В связи с присущими таким базам ограничениями вертикального масштабирования появились новые продукты, нестрого выполняющие фундаментальные требования к СУБД. Жестко заданные модели данных, обеспечивающие надежные гарантии консистентности данных (требование, согласно которому все клиенты при считывании получают одни и те же данные) и строгое следование стандарту SQL, уступили место моделям, лишенным жестких схем и преднамеренно денормализованным, со слабой консистентностью и с проприетарным API.

Распределенные базы данных имеют фундаментальные ограничения с точки зрения качества обслуживания, определяемые теоремой CAP Эрика Брюера о консистентности, доступности (на каждый запрос возвращается отклик об успешном выполнении либо отказе) и устойчивости к разделению. Когда происходит разделение сети, вызывающее потерю связи между узлами кластера, система должна принести консистентность в жертву доступности. Дэниел Абади предлагает PACELC — практическую интерпретацию теоремы CAP: если происходит разделение (P), система обязана принести доступность (A) в жертву консистентности (C). Иначе (E) в обычной ситуации отсутствия разделения система обязана принести в жертву задержку (L) в пользу консистентности (C).

Трудности проектирования масштабируемых систем обработки больших объемов данных обусловлены тремя проблемами. Во-первых, достижение высокой масштабируемости и доступности требует наличия сильно распределенных систем на всех уровнях, от ферм веб-серверов и систем кэширования до систем хранения данных. Во-вторых, при большом масштабе трудно с помощью SQL-подобного языка запросов обеспечить абстрактную репрезентацию системы как единого целого — с транзакционными операциями записи и консистентным считыванием. Приложения должны: «знать» о существовании копий данных; компенсировать несоответствия, возникающие при конфликтующих обновлениях копий; продолжать работу, несмотря на неизбежные сбои процессоров, сетей и программных систем. В-третьих, при использовании любой NoSQL-системы приходится идти на определенные компромиссы, обычно выбирая между быстродействием, масштабируемостью, надежностью и консистентностью. Архитекторы должны тщательно оценивать имеющиеся технологии баз данных и выбирать те из них, которые в наибольшей степени подходят для конкретного приложения. Нередко прибегают к использованию разных технологий для хранения различных срезов данных в пределах одной и той же системы, чтобы удовлетворить соответствующие требования к атрибутам качества.

Особенности приложений для работы с большими данными

Приложения работы с Большими Данными проникают во многие области, и одна из них — аэронавигация. Современные коммерческие авиалайнеры генерируют по 500 Гбайт операционных данных за рейс, необходимых для диагностики неполадок в реальном времени, оптимизации расходов топлива и прогнозирования потребностей в ремонтно-техническом обслуживании. Для повышения безопасности и снижения затрат авиакомпаниям необходимо создавать масштабируемые системы с целью регистрации данных, их анализа и управления ими.

Еще одна область — здравоохранение, где, по некоторым оценкам, средства анализа Больших Данных могли бы принести экономию 450 млрд долл. (данные для США. — Прим. перев.). Анализ петабайтов данных по пациентам, накапливаемых в страховых компаниях, медицинских учреждениях и клинических исследованиях, поможет снизить затраты за счет повышения результативности лечения. Кроме того, аналитические системы позволят получать новые знания, помогающие в лечении и предотвращении заболеваний.

Для всех систем работы с Большими Данными характерны четыре общих требования, которые нужно учитывать при проектировании и которые в совокупности вынуждают существенно отклониться от архитектуры традиционных бизнес-систем, имеющих ограничения на рост объема данных и функциональности.

Во-первых, системы Больших Данных, от сайтов социальных СМИ до датчиков в энергосетях, должны справляться с большим объемом операций записи. Поскольку запись затратнее, чем считывание, можно пользоваться сегментированием данных (секционированием и распределением), чтобы разделять операции записи между накопителями, а для обеспечения высокой доступности можно прибегать к тиражированию. Однако операции сегментирования и тиражирования создают проблемы с доступностью и консистентностью, которые надо как-то компенсировать.

Второе требование — способность справляться с переменной нагрузкой. Уровень загрузки в коммерческих и государственных системах может сильно варьироваться: распродажи, экстренные ситуации, сдача налоговых деклараций и т. п. Чтобы избежать затрат на резервные ресурсы на случай подобных эпизодических скачков, облачные платформы делаются эластичными, позволяя приложениям подключать новые ресурсы для распределения нагрузки и высвобождать их, когда уровень загруженности падает. Для эффективного использования этого механизма нужны архитектуры, способные распознавать скачки нагрузки на различные приложения, быстро добавлять новые ресурсы и высвобождать их по мере необходимости.

Третье требование — возможность выполнения аналитики с большим объемом вычислений. Большинство систем Больших Данных должны справляться с комбинированными рабочими задачами, когда часть запросов требует быстрого ответа, а часть выполняется подолгу в связи со сложным анализом крупных срезов данных. Чтобы удовлетворить это требование, архитектуру ПО и данных специально оптимизируют с расчетом на то, что время обработки запросов будет варьироваться. Один из передовых образцов — система выдачи рекомендаций Netflix, которая способна параллельно выполнять быстрые запросы и сложный анализ больших коллекций данных, непрерывно улучшая качество генерируемых персональных советов.

Четвертое требование — высокая доступность. В горизонтально масштабируемых средах, состоящих из тысяч узлов, аппаратные и сетевые сбои неизбежны, поэтому распределенные архитектуры ПО и данных должны быть устойчивыми. Стандартные методы повышения доступности — тиражирование данных между географическими регионами, реализация сервисов без сохранения состояния и зависимость механизмов от конкретных приложений — позволяют при сбоях продолжать обслуживание, но со снижением качества.

Решения, предназначенные для удовлетворения всех перечисленных требований, реализуются на уровне архитектур распределения, данных и развертывания. Например, для обеспечения эластичности требуются: платформа выполнения, позволяющая резервировать дополнительные вычислительные мощности; политики и механизмы запуска и остановки сервисов в случаях изменения нагрузки на приложения; архитектура СУБД, продолжающая надежно выполнять запросы при возрастании нагрузки.

Подобная конвергенция архитектур, необходимая для обеспечения требуемого качества, типична для приложений, работающих с Большими Данными. Этот подход можно описать как архитектурную модель «4+1» с сильносвязанными процессным, логическим и физическим представлениями.

Пример объединения функциональностей

В Институте программной инженерии при Университете Карнеги — Меллона создается система агрегации данных из множества баз медицинских карт, емкостью в десятки петабайт каждая. Чтобы обеспечить высокую масштабируемость и доступность с малыми затратами, изучается возможность использовать для агрегации базы NoSQL. Для повышения доступности и сокращения задержки при обработке запросов от пользователей, находящихся в разных странах, применяются географически распределенные ЦОД.

Рассмотрим требования к консистентности для двух категорий данных — демографических сведений о пациентах (имя, страховщик и т. д.) и результатов анализов и исследований (рентгеновские снимки и др.). Записи, касающиеся демографии, обновляются нечасто, но должны немедленно отражаться там, где была произведена модификация (должна соблюдаться непротиворечивость чтения собственных записей), при этом допустима задержка отражения обновления в других местах (консистентность в конечном счете). Результаты исследований обновляются чаще, причем изменения должны немедленно отразиться везде (нужна сильная консистентность) — это необходимо для телемедицины и дистанционных консультаций.

3.4 Big Data Система управления большими данными

Big Data – это концепт работы с большими данными, своего рода определённая технология, которая позволяет работать с большими массивами данных быстро и без перегрузки существующей инфраструктуры.

На данный момент, идеология больших данных превратилась в маркетинговый ход крупных производителей программного обеспечения. Каждый, кто ранее продавал решение сбора, хранения, анализа информации … поспешил заявить, что именно его продукт соответствует данному концепту. Потребителю с каждым днём сложнее разобраться, что лож, а что истина данного подхода. Чтобы не возникало сомнений в понимании данного концепта, приведем пример.

Таблица 4 - Пример решения по технологии Big Data и без технологии Big Data

Big Data	**не Big Data**
Данные находятся в различных базах, на серверах отдельных организаций и персональных компьютерах сотрудников компании	Данные перекачиваются в промежуточную базу или создаётся отдельная база данных
Решение работает со всеми видами и типами существующих данных	Решение работает с определённым видом данных
Скорость работы высокая, от	Скорость сильно зависит от существующей инфраструктуры
	На качество информации решение не

существующей инфраструктуры зависит не сильно Чем больше решение работает, тем качественней становятся данные компании Поиск информации происходит мгновенно и не зависит от количества информации, от качества данных и их объёма Поиск информации происходит любым удобным способом, путём выделения карты, при помощи поисковой строки и т. п.	влияет вообще Поиск и скорость поиска зависит от качества и количества данных Поиск происходит только при помощи командной строки …

Big Data не заменяет существующие технологии. Например, почему данные с применением технологии Big Data будут быстрее и качественней анализироваться. Потому, что они будут быстрее найдены, источников информации будет больше и т. п. Система для анализа данных может быть одинаковая.

3.5 Модели данных

Иерархическая модель базы данных состоит из объектов с указателями от родительских объектов к потомкам, соединяя вместе связанную информацию. Иерархические базы данных могут быть представлены как дерево, состоящее из объектов различных уровней. Верхний уровень занимает один объект, второй — объекты второго уровня и т. д. Между объектами существуют связи, каждый объект может включать в себя несколько объектов более низкого уровня. Такие объекты находятся в отношении предка (объект более близкий к корню) к потомку (объект более низкого уровня), при этом возможно, когда объект- предок не имеет потомков или имеет их несколько, тогда как у объекта- потомка обязательно только один предок. Объекты, имеющие общего предка, называются близнецами.

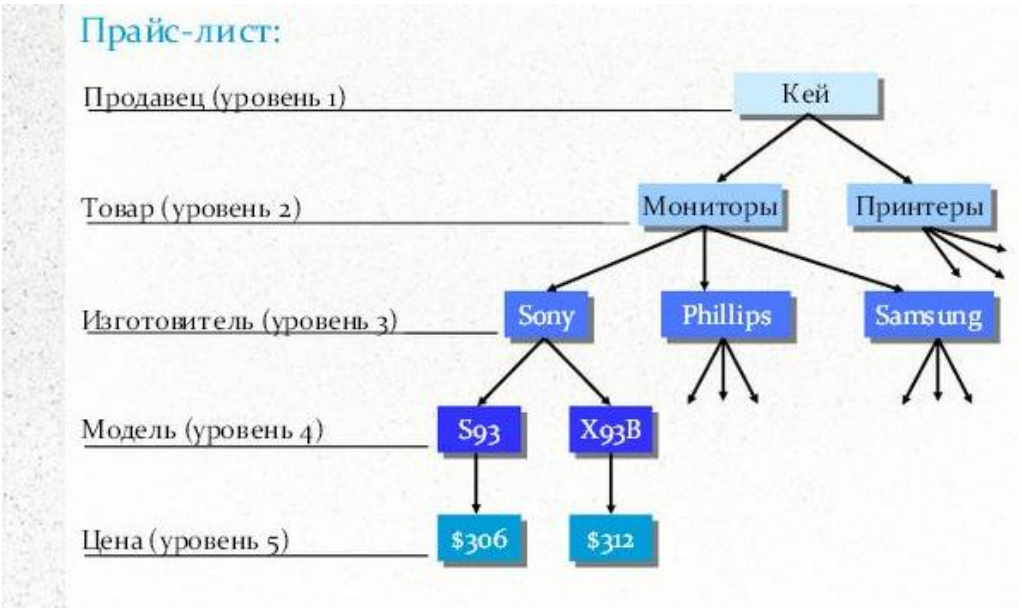


Рисунок 7 – Иерархическая модель базы данных

К основным понятиям сетевой модели базы данных относятся: уровень, элемент (узел), связь. Узел — это совокупность атрибутов данных, описывающих некоторый объект. На схеме иерархического дерева узлы представляются вершинами графа.В сетевой структуре каждый элемент может быть связан с любым другим элементом. Сетевые базы данных подобны иерархическим, за исключением того, что в них имеются указатели в обоих направлениях, которые соединяют родственную информацию. Несмотря на то, что эта модель решает некоторые проблемы, связанные с иерархической моделью, выполнение простых запросов остается достаточно сложным процессом. Также, поскольку логика процедуры выборки данных зависит от физической организации этих данных, то эта модель не является полностью независимой от приложения. Другими словами, если необходимо изменить структуру данных, то нужно изменить и приложение.

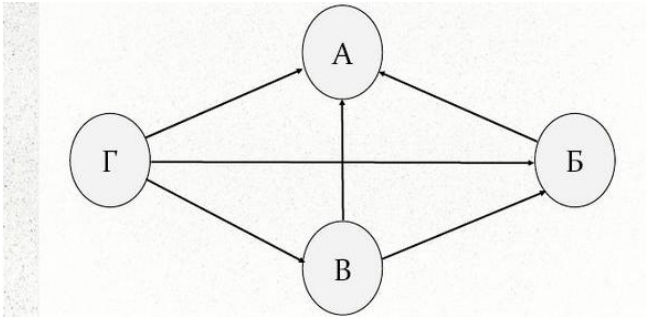


Рисунок 8 – Сетевая база данных

Реляционная база данных — база данных, основанная на реляционной модели данных. Термин «реляционный» означает, что теория основана на математическом понятии отношение (relation). В качестве неформального синонима термину «отношение» часто встречается слово таблица. Необходимо помнить, что «таблица» есть понятие нестрогое и неформальное и часто означает не «отношение» как абстрактное понятие, а визуальное представление отношения на бумаге или экране. Некорректное и нестрогое использование термина «таблица» вместо термина «отношение» нередко приводит к недопониманию. Наиболее частая ошибка состоит в рассуждениях о том, что РМД имеет дело с «плоскими», или «двумерными» таблицами, тогда как таковыми могут быть только визуальные представления таблиц. Отношения же являются абстракциями, и не могут быть ни «плоскими», ни «неплоскими».

Рекомендуемая литература:

1. D. Agrawal, S. Das, A. El Abbadi. Big Data and Cloud Computing: Current State and Future Opportunities // Proc.14th Int'l Conf. Extending Database Technology (EDBT/ICDT 11). — 2011. — P. 530–533.
2. W. Vogels. Amazon DynamoDB — a Fast and Scalable NoSQL Database Service Designed for Internet Scale Applications. Blog, 18 Jan. 2012. URL: http://www.allthingsdistributed.com/2012/01/amazon-dynamodb.html
3. F. Chang et al. Bigtable: A Distributed Storage System for Structured Data. ACM Trans // Computing Systems. — 2008. Vol. 26, № 2 (article 4).
4. Big Data Система управления большими данными URL: http://rzng.ru/index.php/big-data-sistema-upravlenija-bolshimi-dannymi.html
5. Системы для Больших Данных: конвергенция архитектур. URL: https://www.osp.ru/os/2015/03/13046898/

Тема 4 Стандарты в области больших данных

Цель: рассмотреть стандарты больших данных в ISO/IEC. Стандарты больших данных в ITU. Стандарты больших данных в NIST. Стандарты больших данных в BSI. Другие стандарты для больших данных.

План:
4.1 Стандарты больших данных в ISO/IEC.
4.2 Стандарты больших данных в ITU.
4.3 Стандарты больших данных в NIST.
4.4 Стандарты больших данных в BSI.
4.5 Другие стандарты для больших данных.

В настоящий момент, несколько основных институтов стандартизации вовлечены в работу по стандартам для больших данных. Основные игроки здесь – Международная организация по стандартизации и Международная Электротехническая комиссия (ISO/IEC), Международный Союз Электросвязи (ITU), Британский Институт Стандартов (BSI), Национальный Институт Стандартов и Технологии США (NIST).

4.1 Стандарты больших данных в ISO/IEC

Международная организация по стандартизации и Международная электротехническая комиссия (ИСО / МЭК) создали 3 рабочие группы, ориентированные на стандартизацию следующих технологий: большие данных (ISO/IEC JTC1/WG 9 «Big data»), интернет вещей (ISO/IEC JTC1/WG 10 «Internet of things») и умные города (ISO/IEC JTC1/WG 11 «Smart Cities»). На самом деле, это довольно общий тренд, увязывать большие данные и Интернет Вещей, так как измерения в IoT есть большие данные де-факто. В соответствии со стандартом ИСО, Рабочая группа по большим данным будет служить в качестве определяющей для главной темы большой программы стандартизации данных и выявления пробелов в области стандартизации. Она будет разрабатывать основополагающие стандарты - в том числе эталонной архитектуры и словарный запас. В настоящее время международная рабочая группа по стандартизации ISO/IEC JTC1/WG 9 «Big data» разрабатывает следующие проекты международных стандартов: комплекс стандартов на эталонную архитектуру больших данных (ISO/IEC 20547) и стандарт на термины и определения (ISO/IEC 20546). Руководитель этой рабочей группы также является цифровым консультантом данных для Национального института стандартов и технологий (NIST), поэтому мы можем ожидать, что в итоговых документах мы увидим многие идеи из NIST здесь. На момент написания статьи (август 2016 г.) рабочая группа выпустила обзор больших данных и словарь. Он был представлен как RFC – документ (запрос на комментарии). Это относительно небольшой документ. Он содержит основные определения модели данных (например, модели 4V для больших объемов данных - объем, скорость, разнообразие и изменчивость), а также небольшой набор терминов (например, кластерные вычисления, параллельные вычисления и т.д.). На самом деле, этот документ был переведен на русский язык и разослан на согласование по членам ТК098/РГ3 “Большие данные”. В то же время, необходимо прямо еще раз отметить, что это очень небольшой документ, и его явно недостаточно для практической работы (понимаемой именно как представление "наилучшей практики"). Точной информации о дорожной карте от ISO/IEC в этом вопросе у нас нет и, по-видимому, мы должны смотреть на то, что делает NIST в этом направлении. По нашему мнению, именно результаты NIST окажутся

в итоге в ISO/IEC. В разделе, который посвящен работам NIST, приводится его (NIST) собственная дорожная карта, составленная с помощью ISO.

4.2 Стандарты больших данных в ITU

В ITU можно отметить несколько областей активности, касающихся больших данных. В документах ITU указываются следующие области активности: - высоконадежная, гибкая и масштабируемая сетевая инфраструктура с высокой пропускной способностью и с низкой задержкой. Эта область (с точки зрения МСЭ) имеет много пересечений с разделом облачных вычислений. И именно в этой области, МСЭ представил свой документ как "первый стандарт Big Data". - Агрегирование и анонимизация наборов данных. Это направление очень важно для таких областей, как Умные Города (концепция города, управляемого данными или город – как сенсор), Интернет Вещей и электронные медицинские приложения. Именно эти работы могли бы послужить базов для отечественных работ в области стандартизации, тяготеющих к правовому регулированию. - Совместимые платформы. МСЭ планирует целевые вертикальные рынки (домашней автоматизации, электронное здравоохранение) со стандартами совместимости данных. Это должно представлять интерес для отечественных работ в области телемедицины. - Мультимедиа-аналитика. Мультимедийный контентный анализ IP-трафика, который позволит автоматически и быстро расшифровывать (выделять) события и другие мета-данные для видео-файлов и видео-потоков. - Стандарты для открытых данных. В соответствии с МСЭ, работа по стандартизации должны разработать требования к представлению данных и механизмов для публикации, распространения и открытия наборов данных. Это направление имеет большое пересечение с тематикой Умных Городов. Здесь сразу можно указать на стандарт BSI PAS-212:2016 для обнаружения (раскрытия) данных в умных городах. Опять таки, продвигаемый BSI как "первый стандарт Smart City". Теперь он передается ISO для будущего развития. PAS- 212 переведен на русский язык и его локализация в России обсуждается. Стандарты открытых данных, разрабатываемые ITU также должны включать элементы раскрытия данных. Это один из основных моментов – иначе как находить данные? Без кооперации с ISO мы рискуем увидеть несколько несовместимых стандартов. Как видно из приведенного списка, проводимые работы, на первый взгляд, далеки от того, что принято в России назвать большими данными. Поэтому первая задача для локальных стандартизаторов – это определиться с тем, что они будут рассматривать. В конце 2015 года, члены МСЭ договорились о международном стандарте для больших данных. Новый стандарт, рекомендация МСЭ-Т Y.3600 "Большие данные - требования и возможности на основе облачных вычислений", был разработан группой экспертов МСЭ-Т, ответственных за будущее сетей, облачных вычислений и сетевых аспектов мобильной связи. Стандарт описывает, как облачные вычислительные системы могут быть использованы для предоставления услуг Big Data. Главным образом, он описывает требования к облачным вычислениям на основе больших объемов данных (требования по сбору данных, предварительной обработки данных и требования к хранению данных, анализу, визуализации и управлению, безопасности данных и требования по защите, сбору и хранению данных). Кроме того, он содержит определения сервиса больших данных как услуги (BDaaS) - категории облачных сервисов, в которой возможности, предоставляемые клиенту облачных услуг, позволяют собирать, хранить, анализировать, визуализировать и управлять данными с использованием технологий больших объемов данных. На наш взгляд эта рекомендация заслуживает самого пристального внимания в России, поскольку предоставление “хостинга” больших данных – очень важная задача, хотя бы для учебных заведений, где построение учебных стендов собственными силами будет весьма затруднено.

4.3 Стандарты больших данных в NIST

На наш взгляд, NIST предлагает наиболее проработанный стек стандартов по большим данным. Этот так называемый NIST Big Data Interoperability Framework V1.0 включает в себя следующие документы: - Big Data Definitions - Big Data Taxonomies - Big Data Use Cases and Requirements - Big Data Security and Privacy - Big Data Architecture White Paper Survey - Big Data Reference Architecture - Big Data Standards Roadmap Определения более ориентированы на практику, чем у ISO и, фактически, посвящены моделям, а не словарным статьям. Документ по таксономии предлагает базовую архитектурную модель (рисунок 9).

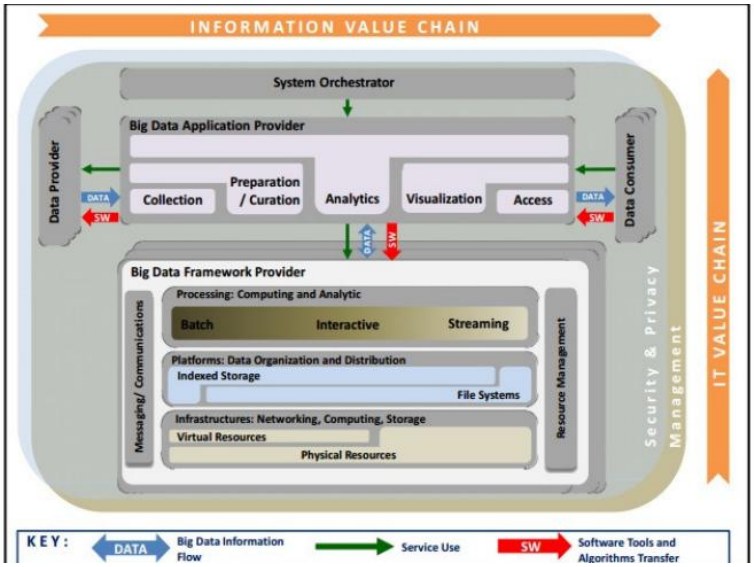


Рисунок 9 - Ссылочная модель Big Data от NIST

Документ с примерами содержит описание реальных применений (описание архитектур данных) из 9 областей: правительственные системы, коммерческие системы, военные применения, здравоохранение, социальные медиа, экология, астрономия и др. Этот раздел может вполне использоваться как учебник (или часть учебного материала).

Документ, касающийся безопасности и приватности также содержит практические примеры из таких областей как здравоохранение, телеметрия, маркетинг и др. Из технологий безопасности рассматриваются вопросы идентичности, авторизация, аудит, безопасность в сети, безопасность устройств. White Papers Survey рассматривает архитектуры данных и является одним из наиболее интересных элементов в списке NIST. предложения и справочные модели (для архитектур данных) от таких компаний, как IBM, Oracle и исследовательских групп в университетах. Big Data Reference Architecture представляет собой концептуальную модель высокого уровня, разработанную для того, чтобы служить в качестве инструмента для содействия открытому обсуждению требований, проектных структур и операций, присущих большим данных. Она не представляет системную архитектуру конкретной системы больших данных, а скорее является инструментом для описания, обсуждения и разработки архитектуры конкретной системы, используя общую точку (базу) отсчета. Эта модель не привязана к какой-либо конкретной продукции поставщиков. Последний документ - Дорожная карта описывает пересечения между NIST Framework и существующими стандартами (например, SQL). Но самая интересная часть этой дорожной карты - это список направлений развития, который

был построен с помощью Объединенного технического комитета 1 (ОТК1) ИСО/МЭК. Исследовательская группа по большим данным определила направления работ, которые могли бы служить в качестве потенциального руководства для ISO в их создании стандартов деятельности больших данных. Исследовательской группой были определены потенциальные пробелы в стандартизации темы больших данных. Эти результаты описывают достаточно широкие области, которые могут представлять интерес для разработчиков и исследователей. Фактически – это “горячие” темы для исследований в области больших данных: - Модели применения и ссылочные архитектуры. Платформы для больших данных. - Технические требования и стандартизация для метаданных. - Модели для приложений (например, пакетная обработка и анализ потоков) - Языки запросов, включая нереляционные запросы для поддержки различных типов данных (например, XML, RDF, JSON, мультимедиа) и больших объемов данных операций (например, матричных операций).

- Проблемно-ориентированные языки для задач больших данных. - Семантика для слабой согласованности и согласованности в конечном счете (eventual consistency). - Расширенные сетевые протоколы для эффективной передачи данных. - Общие и предметно-ориентированные онтологии и таксономии для описания семантики данных, включая интероперабельность между онтологиями. - Безопасность и контроль доступа. - Удаленные, распределенные и федеративные аналитики данных, включая методы обнаружения и обработки ресурсов, а также методы интеллектуального анализа данных. - Обмен и разделение данных. - Системы хранения данных (например, in-memory базы данных, распределенные файловые системы, хранилища данных). - Представление результатов анализа данных (сюда, очевидно, должны входить визуализация и объяснения). - Энергетическая эффективность работы с большими данными. - Интерфейс между реляционными (т.е. SQL) и нереляционными (т.е. NoSQL) хранилищами данных - Оценка качества и достоверности данных.

4.4 Стандарты больших данных в BSI

BSI и его работы в области стандартизации интересны, прежде всего, тесной связью с экономикой. Таким образом, решение BSI работать в стандартах больших данных также базируется на экономике и экономических данных. В соответствии с BSI, Big Data представляет собой весьма существенную и быстрорастущий возможность на рынке сегодня. Это принесет пользу экономике Великобритании на £ 216 млрд. к 2017 году и привести к созданию 58000 новых рабочих мест. На рисунке 10 показаны проблемы адаптации больших данных согласно BSI.

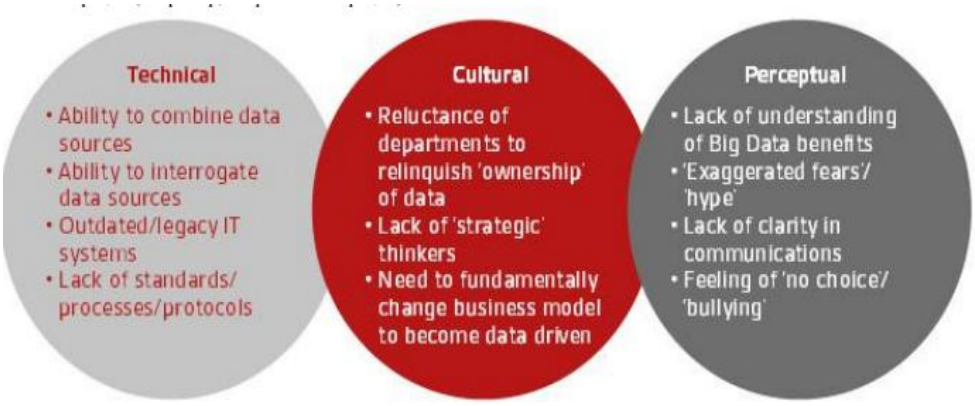


Рисунок 10 - Проблемы адаптации больших данных

В соответствии с BSI, следующие темы подлежат стандартизации в больших данных:

- Стандарт на метаданные. В мире больших данных, метаданные являются основой для аналитических отчетов. Множество аналитических отчетов базируются именно на метаданных. BSI предполагает определить наилучшую практику для сбора и хранения метаданных, в том числе, обеспечить качество и полноту собранной информации, а также регламентные процедуры (принципы) для долговременного хранения метаданных (например, как долго они должны храниться). В связи с этим, следует отметить еще раз стандарт для раскрытия (описания) данных в умных городах. Это типичный пример стандарта, связанного с метаданными.

- Стандарты на условия работы с данными. Это также полностью укладывается в идеологию построения «лучших практик». Построение общественного доверия имеет жизненно важное значение. Стандартизуя описание условий работы с данными, BSI собирается обеспечить наилучшую практику для обеспечения того, чтобы условия были просты для понимания и обеспечивали информированное общественное согласие. - Стандарты сбора данных. Этот стандарт должен быть нацелен именно на процесс сбора данных. Стандартизоваться должно представление того, как (с помощью чего) потребитель принимает решение о том, кто может использовать его данные. - Стандарты объяснения для проектов Big Data. Многие из инициатив в области больших данных не удается провести (реализовать) из-за общественного сопротивления. Этот стандарт должен определить наилучшую практику для того, как инициативы больших объемов данных должны быть объяснены. Он должен помочь предоставлять инициативы потребителям в ясной и однозначной форме.

- Руководство "Как сделать" для Big Data. На самом деле, это ближайший аналог Примеров использования из NIST. Согласно BSI, этот стандарт должен помочь предприятиям стартовать с проектами Big Data. Отметим, что аналогичные проекты на русском языке авторам неизвестны.

4.5 Другие стандарты для больших данных

Из других институтов, которые имеют инициативы, относящиеся к большим данным, отметим:

• Институт Инженеров Электротехники и Электроники (IEEE)

• Международную Электротехническую Комиссию (IEC)

• Инженерный Совет Интернета (The Internet Engineering Task Force - IETF)

• Консорциум Всемирной Паутины (World Wide Web Consortium -W3C)

• Открытый гео-консорциум (Open Geospatial Consortium - OGC)

• Глобальный консорциум промышленных стандартов (Organization for the Advancement of Structured Information Standards - OASIS)

• Пользовательская группа по облачным стандартам (The Cloud Standards Customer Council)

Направления IEEE Будущие Инициативы Big Data стремится объединить информацию о различных начинаниях, происходящих во всем мире, с тем, чтобы поддержать сообщество профессионалов в промышленности, научных кругах и правительствах, работающих над решением проблем, связанных с большими данными.

Международная электротехническая комиссия (МЭК) работает с ISO по стандартам больших данных. IETF разрабатывает и продвигает добровольные стандарты Интернет. Есть несколько проектов, связанных со стандартами больших данных. Например, это сеть телеметрии и анализа данных. Их развитие фокусируется на измерении сети (трафика) и анализа в сетевой среде. Он определяет сетевую телеметрию,

описывает архитектуру телеметрической сети, а затем исследует характеристики данных сети.

Работы W3C относятся к семантическим данным.

Проект Big Data Europe осуществляется в рамках программы Horizon 2020. Он будет предоставлять решения в следующих областях: гетерогенные связи и интеграция данных, биомедицинская семантическая индексация, крупномасштабная распределенная интеграция данных, мониторинг в режиме реального времени, потоковая обработка и анализ данных, поддержка систем принятия решений, потоковые сети сенсорных данных, гео-пространственная интеграция данных, мониторинг в режиме реального времени, анализ изображений.

Открытый Гео-консорциум создал рабочие группы с особым акцентом на пространственно-временные данные. OASIS также создал несколько комиссий, связанных с большими данными. Самая интересная, на наш взгляд, работа – это база данных OASIS Key-Value Application Interface (KVDB) TC. В ней определяется открытый программный интерфейс для управления и доступа к данным из систем баз данных на основе модели ключ- значение. Это одна из широко используемых моделей данных в NoSQL. Пользовательская группа по облачным стандартам публикует описания лучших технических практик для больших данных в облачной среде.

Рекомендуемая литература:

1. Намиот Д.Е., Шнепс-Шнеппе М.А. Об отечественных стандартах для Умного Города // International Journal of Open Information Technologies. 2016. -Т. 4. - №7. С.32-37.
2. BSI Big Data and standards market research, January 2016
3. ISO/IEC JTC 1 Forms Two Working Groups on Big Data and Internet of Things https://www.ansi.org/news_publications/news_story.aspx?menuid=7 &articleid=5b101d27-47b5-4540-bca3-657314402591 Retrieved: Aug, 2016
4. ITU-T LIAISON STATEMENT ISO/IEC JTC1/WG9 - ISO/IEC JTC 1/WG 9 N 201 http://www.itu.int/net/itu-t/ls/ls.aspx?isn=12493 Retrieved: Sep, 2016
5. ITU Big Data http://www.itu.int/en/ITU-T/techwatch/Pages/big-datastandards.aspx Retrieved: Sep, 2016.
6. Tene, Omer, and Jules Polonetsky. "Big data for all: Privacy and user control in the age of analytics." Nw. J. Tech. & Intell. Prop. 11 (2012): xxvii
7. Smith, John R. "Riding the multimedia big data wave." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.
8. Kupriyanovsky, Vasily, et al. "On Localization of British Standards for Smart Cities." International Journal of Open Information Technologies 4.7 (2016): 13-21.
9. ITU Y.3600 http://www.itu.int/itut/recommendations/rec.aspx?rec=12584 Retrieved: Aug, 2016
10. NIST Big Data http://www.nist.gov/itl/bigdata/bigdatainfo.cfm Retrieved: Aug, 2016
11. Big Data Taxonomies http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500- 2.pdf Retrieved: Sep, 2016
12. NBDRA http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf Retrieved: Aug, 2016
13. Big Data Standards Workshop http://www.bsigroup.com/en-GB/ourservices/events/2016/

14. Big Data Standards and Market Research http://shop.bsigroup.com/upload/275237/The-Big-Data-AndStandards-Market-Research-Report-By-BSI-And-Circle-Research.pdf Retrieved: Aug, 2016
15. IEEE Big Data http://bigdata.ieee.org/ Retrieved: Aug, 2016
16. Network Telemetry and Big Data Analysis https://datatracker.ietf.org/doc/draft-wu-t2trg-network-telemetry/ Retrieved: Aug, 2016
17. Big Data Europe https://www.big-data-europe.eu Retrieved: Aug, 2016
18. Big Data Working Group http://www.opengeospatial.org/projects/groups/bigdatadwg Retrieved: Aug, 2016
19. OASIS Key-Value Database Application Interface (KVDB) TC https://www.oasisopen.org/committees/tc_home.php?wg_abbrev=kvdb Retrieved: Aug, 2016
20. The Cloud Standards Customer Council http://www.cloudcouncil.org/resource-hub.htm Retrieved: Aug, 2016.

Тема 5 Основные понятия математической статистики

Цель: рассмотреть методы анализа данных: дескриптивная статистика, параметрические, непараметрические, номинальные методы (корреляционный, регрессионный, дисперсионный анализы, кластерный, дискриминантный, факторный анализы).

План:
5.1 Обзор методов статистического анализа данных

5.1 Обзор методов статистического анализа данных

Объектом исследования в прикладной статистике являются статистические данные, полученные в результате наблюдений или экспериментов. Статистические данные – это совокупность объектов (наблюдений, случаев) и признаков (переменных), их характеризующих. Например, объекты исследования – страны мира и признаки, – географические и экономические показатели их характеризующие: континент; высота местности над уровнем моря; среднегодовая температура; место страны в списке по качеству жизни, доли ВВП на душу населения; расходы общества на здравоохранение, образование, армию; средняя продолжительность жизни; доля безработицы, безграмотных; индекс качества жизни и т.д.

Переменные – это величины, которые в результате измерения могут принимать различные значения.

Независимые переменные – это переменные, значения которых в процессе эксперимента можно изменять, а зависимые переменные – это переменные, значения которых можно только измерять.

Переменные могут быть измерены в различных шкалах. Различие шкал определяется их информативностью. Рассматривают следующие типы шкал, представленные в порядке возрастания их информативности: номинальная, порядковая, интервальная, шкала отношений, абсолютная. Эти шкалы отличаются друг от друга также и количеством допустимых математических действий. Самая «бедная» шкала – номинальная, так как не определена ни одна арифметическая операция, самя «богатая» – абсолютная.

Измерение в номинальной (классификационной) шкале означает определение принадлежности объекта (наблюдения) к тому или иному классу. Например: пол, род

войск, профессия, континент и т.д. В этой шкале можно лишь посчитать количество объектов в классах – частоту и относительную частоту.

Измерение в порядковой (ранговой) шкале, помимо определения класса принадлежности, позволяет упорядочить наблюдения, сравнив их между собой в каком-то отношении. Однако эта шкала не определяет дистанцию между классами, а только то, какое из двух наблюдений предпочтительнее. Поэтому порядковые экспериментальные данные, даже если они изображены цифрами, нельзя рассматривать как числа и выполнять над ними арифметические операции 5 . В этой шкале дополнительно к подсчету частоты объекта можно вычислить ранг объекта. Примеры переменных, измеренных в порядковой шкале: бальные оценки учащихся, призовые места на соревнованиях, воинские звания, место страны в списке по качеству жизни и т.д. Иногда номинальные и порядковые переменные называют категориальными, или группирующими, так как они позволяют произвести разделение объектов исследования на подгруппы.

При измерении в интервальной шкале упорядочивание наблюдений можно выполнить настолько точно, что известны расстояния между любыми двумя их них. Шкала интервалов единственна с точностью до линейных преобразований ($y = ax + b$). Это означает, что шкала имеет произвольную точку отсчета – условный нуль. Примеры переменных, измеренных в интервальной шкале: температура, время, высота местности над уровнем моря. Над переменными в данной шкале можно выполнять операцию определения расстояния между наблюдениями. Расстояния являются полноправными числами и над ними можно выполнять любые арифметические операции.

Шкала отношений похожа на интервальную шкалу, но она единственна с точностью до преобразования вида $y = ax$. Это означает, что шкала имеет фиксированную точку отсчета – абсолютный нуль, но произвольный масштаб измерения. Примеры переменных, измеренных в шкале отношений: длина, вес, сила тока, количество денег, расходы общества на здравоохранение, образование, армию, средняя продолжительность жизни и т.д. Измерения в этой шкале – полноправные числа и над ними можно выполнять любые арифметические действия.

Абсолютная шкала имеет и абсолютный нуль, и абсолютную единицу измерения (масштаб). Примером абсолютной шкалы является числовая прямая. Эта шкала безразмерна, поэтому измерения в ней могут быть использованы в качестве показателя степени или основания логарифма. Примеры измерений в абсолютной шкале: доля безработицы; доля безграмотных, индекс качества жизни и т.д.

Большинство статистических методов относятся к методам параметрической статистики, в основе которых лежит предположение, что случайный вектор переменных образует некоторое многомерное распределение, как правило, нормальное или преобразуется к нормальному распределению. Если это предположение не находит подтверждения, следует воспользоваться непараметрическими методами математической статистики.

Корреляционный анализ

Между переменными (случайными величинами) может существовать функциональная связь, проявляющаяся в том, что одна из них определяется как функция от другой. Но между переменными может существовать и связь другого рода, проявляющаяся в том, что одна из них реагирует на изменение другой изменением своего закона распределения. Такую связь называют стохастической. Она появляется в том случае, когда имеются общие случайные факторы, влияющие на обе переменные. В качестве меры зависимости между переменными используется коэффициент корреляции (r), который изменяется в пределах от -1 до $+1$. Если коэффициент корреляции отрицательный, это означает, что с увеличением значений одной переменной значения

другой убывают. Если переменные независимы, то коэффициент корреляции равен 0 (обратное утверждение верно только для переменных, имеющих нормальное распределение). Но если коэффициент корреляции не равен 0 (переменные называются некоррелированными), то это значит, что между переменными существует зависимость. Чем ближе значение r к 1, тем зависимость сильнее. Коэффициент корреляции достигает своих предельных значений +1 или -1, тогда и только тогда, когда зависимость между переменными линейная. Корреляционный анализ позволяет установить силу и направление стохастической взаимосвязи между переменными (случайными величинами). Если переменные измерены, как минимум, в интервальной шкале и имеют нормальное распределение, то корреляционный анализ осуществляется посредством вычисления коэффициента корреляции Пирсона, в противном случае используются корреляции Спирмена, тау Кендала, или Гамма

Регрессионный анализ

В регрессионном анализе моделируется взаимосвязь одной случайной переменной от одной или нескольких других случайных переменных. При этом, первая переменная называется зависимой, а остальные – независимыми. Выбор или назначение зависимой и независимых переменных является произвольным (условным) и осуществляется исследователем в зависимости от решаемой им задачи. Независимые переменные называются факторами, регрессорами или предикторами, а зависимая переменная – результативным признаком, или откликом.

Если число предикторов равно 1, регрессию называют простой, или однофакторной, если число предикторов больше 1 – множественной или многофакторной. В общем случае регрессионную модель можно записать следующим образом:

$y = f(x_1, x_2, \ldots, x_n)$,

где y – зависимая переменная (отклик), x_i (i = 1,…, n) – предикторы (факторы), n – число предикторов.

Посредством регрессионного анализа можно решать ряд важных для исследуемой проблемы задач:

1). Уменьшение размерности пространства анализируемых переменных (факторного пространства), за счет замены части факторов одной переменной – откликом. Более полно такая задача решается факторным анализом.

2). Количественное измерение эффекта каждого фактора, т.е. множественная регрессия, позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, «что является лучшим предиктором для...». При этом, становится более ясным воздействие отдельных факторов на отклик, и исследователь лучше понимает природу изучаемого явления.

3). Вычисление прогнозных значений отклика при определенных значениях факторов, т.е. регрессионный анализ, создает базу для вычислительного эксперимента с целью получения ответов на вопросы типа «Что будет, если… ».

4). В регрессионном анализе в более явной форме выступает причинно-следственный механизм. Прогноз при этом лучше поддается содержательной интерпретации.

Канонический анализ

Канонический анализ предназначен для анализа зависимостей между двумя списками признаков (независимых переменных), характеризующих объекты. Например, можно изучить зависимость между различными неблагоприятными факторами и появлением определенной группы симптомов заболевания, или взаимосвязь между двумя

группами клинико-лабораторных показателей (синдромов) больного. Канонический анализ является обобщением множественной корреляции как меры связи между одной переменной и множеством других переменных. Как известно, множественная корреляция есть максимальная корреляция между одной переменной и линейной функцией других переменных. Эта концепция была обобщена на случай связи между множествами переменных – признаков, характеризующих объекты. При этом достаточно ограничиться рассмотрением небольшого числа наиболее коррелированных линейных комбинаций из каждого множества. Пусть, например, первое множество переменных состоит из признаков y1, …, yp, второе множество состоит из – x1, …, xq, тогда взаимосвязь между данными множествами можно оценить как корреляцию между линейными комбинациями a1y1 + a2y2 + ... + apyp, b1x1 + b2x2 + ... + bqxq,, которая называется канонической корреляцией.

адача канонического анализа в нахождении весовых коэффициентов таким образом, чтобы каноническая корреляция была максимальной.

Кластерный анализ

Кластерный анализ – это метод классификационного анализа; его основное назначение – разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры. Это многомерный статистический метод, поэтому предполагается, что исходные данные могут быть значительного объема, т.е. существенно большим может быть как количество объектов исследования (наблюдений), так и признаков, характеризующих эти объекты. Большое достоинство кластерного анализа в том, что он дает возможность производить разбиение объектов не по одному признаку, а по ряду признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходных данных практически произвольной природы. Так как кластеры – это группы однородности, то задача кластерного анализа заключается в том, чтобы на основании признаков объектов разбить их множество на m (m – целое) кластеров так, чтобы каждый объект принадлежал только одной группе разбиения. При этом объекты, принадлежащие одному кластеру, должны быть однородными (сходными), а объекты, принадлежащие разным кластерам, – разнородными. Если объекты кластеризации представить как точки в n-мерном пространстве признаков (n – количество признаков, характеризующих объекты), то сходство между объектами определяется через понятие расстояния между точками, так как интуитивно понятно, что чем меньше расстояние между объектами, тем они более схожи.

Дискриминантный анализ

Дискриминантный анализ включает статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками. Этот вид анализа является многомерным, так как использует несколько признаков объекта, число которых может быть сколь угодно большим. Цель дискриминантного анализ состоит в том, чтобы на основе измерения различных характеристик (признаков) объекта классифицировать его, т. е. отнести к одной из нескольких заданных групп (классов) некоторым оптимальным способом. При этом предполагается, что исходные данные наряду с признаками объектов содержат категориальную (группирующую) переменную, которая определяет принадлежность объекта к той или иной группе. Поэтому в дискриминантном анализе предусмотрена

проверка непротиворечивости классификации, проведенной методом, с исходной эмпирической классификацией. Под оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации. В общем случае задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является построение k-мерного случайного вектора X = (X1, X2, …, XK), где X1, X2, …, XK – признаки объекта. Требуется установить правило, согласно которому по значениям координат вектора X объект относят к одной из возможных совокупностей i, i = 1, 2, …, n. Методы дискриминации можно условно разделить на параметрические и непараметрические. В параметрических известно, что распределение векторов признаков в каждой совокупности нормально, но нет информации о параметрах этих распределений. Непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений. Если выполняются условия применимости дискриминантного анализа – независимые переменные–признаки (их еще называют предикторами) должны быть измерены как минимум в интервальной шкале, их распределение должно соответствовать нормальному закону, необходимо воспользоваться классическим дискриминантным анализом, в противном случае – методом общие модели дискриминантного анализа.

Факторный анализ

Факторный анализ – один из наиболее популярных многомерных статистических методов. Если кластерный и дискриминантный методы классифицируют наблюдения, разделяя их на группы однородности, то факторный анализ классифицирует признаки (переменные), описывающие наблюдения. Поэтому главная цель факторного анализа – сокращение числа переменных на основе классификация переменных и определения структуры взаимосвязей между ними. Сокращение достигается путем выделения скрытых (латентных) общих факторов, объясняющих связи между наблюдаемыми признаками объекта, т.е. вместо исходного набора переменных появится возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа взаимосвязанных переменных.

Рекомендуемая литература:

1Обзор методов статистического анализа данных. [Электронный ресурс]/Режим доступа: http://statlab.kubsu.ru/node/4

2 Решение задач анализа данных с помощью аналитических визуальных моделей http://graphicon.ru/html/2017/papers/pp116-120.pdf

3 Визуализация данных как средство принятия решений http://earchive.tpu.ru/bitstream/11683/37162/1/conference_tpu-2016-C04_V2_p145-146.pdf

Тема 6 Методика решения задач анализа данных при использовании аналитических визуальных моделей

Цель: рассмотреть методику решения задач анализа данных при использовании аналитических визуальных моделей

План:

6.1 Структура визуального анализа

6.2 Структурная единица визуального анализа
6.3 Типы структурных единиц
6.4 Комбинированная модель
6.5 Методический подход к использованию аналитических визуальных моделей
6.6 Исследование предметной области

6.1 Структура визуального анализа

Анализ является направленной процедурой, которую возможно представить в виде последовательности обоснованных шагов, направленных на получение исследователем ответа на основной вопрос задачи анализа. В простейшем случае, такая последовательность начинается с формулировки вопроса в терминах доступных средств исследования и продолжается поиском гипотез, позволяющих ответить на этот вопрос. Завершается последовательность этапом, на котором происходит проверка следствий одной или нескольких возможных гипотез и принятие полного ответа.

Правильность результата процедуры анализа определяется сочетанием нескольких независимых факторов. Характерные особенности этих факторов позволяют сформировать три функциональных группы. Улучшение общей результативности анализа достигается правильным выбором элементов в каждой такой группе. Таким образом, методика решения задачи анализа должна позволять исследователю делать наиболее результативный выбор факторов, характеризующих создание и использование средств анализа.

Разделение факторов, влияющих на ход анализа, создает предпосылки для построения классификации возможных решений. Возникает набор свойств, конкретизация значений которых при наличии постановки задачи анализа, позволяет выбрать наиболее перспективное направление решения. Результат анализа, в таком подходе, зависит от сути вопроса анализа, который определяется степенью предварительного понимания исследователем анализируемой информации, а также от формы этого вопроса. Форма, во многом, определяется языком, на котором поставлен вопрос, а также внутренней структурой, логикой и необходимостью взаимодействия с новыми источниками информации.

Анализ, осуществляемый на основании взаимодействия исследователя и визуального образа данных, можно представить аналогичным образом (рисунок 1). Созданный искусственный визуальный образ в момент начала анализа является моделью исходных данных, построенной так, чтобы спровоцировать зрителя на осознание существования проблемы. Вместе с формированием гипотезы решения и на основании ее верификации у пользователя формируется вопрос дальнейшего исследования или уверенность в достижении цели анализа.

Возникает последовательность состояний визуальной модели или циклически повторяемая процедура анализа, для которой характерно одновременное существование поисковой и верифицирующей составляющих. Все действия, совершаемые наблюдателем, направлены на построение и оценку гипотезы ответа на вопрос очередного этапа. Эти действия возможны лишь на основании информации, представленной в визуальной модели и, вероятно, при использовании собственных знаний пользователя. Следовательно, многошаговый последовательный процесс визуального анализа состоит в переходе от визуального образа-вопроса, являющегося формулировкой решаемой задачи, к мысленному образу-ответу, интерпретируемому пользователем как закономерность в исследуемых данных, соответствующая ответу на общий вопрос (рисунок 11).

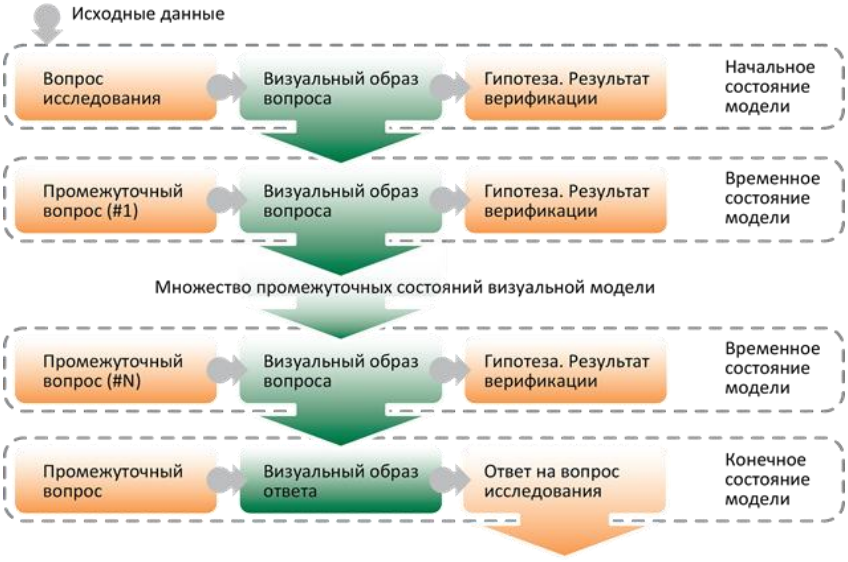


Рисунок 11 - Структура визуального анализа

6.2 Структурная единица визуального анализа

Цикличность процедуры визуального анализа приводит к появлению понятия о элементарном структурном объекте, повторение которого формирует ход решения. Структурной единицей визуализации (визуальным элементом) будем считать состояние визуальной модели, интерпретация которого предоставляет наблюдателю объем информации, необходимый для последующего получения общего результата проводимого анализа. Структурная единица может быть самостоятельной визуальной моделью, либо входит в состав более сложного объединения, являющегося средством визуального анализа применительно к поставленной задаче.

Согласно приведенному выше описанию процедуры анализа, структурная единица визуального анализа является визуально воспринимаемым образом, интерпретируемым как однозначный ответ на один из промежуточных вопросов. Сложность и структура такого вопроса определяется ограниченностью времени, предоставляемого процедурой анализа, а также ресурсоемкостью процессов создания структурной единицы и верификации сведений, полученных пользователем, благодаря ее изучению.

В терминах системного анализа, структурная единица визуального анализа представляет собой управляемую систему *S* с обратной связью *R(t)*. Объем исследуемых данных поступает на неуправляемый вход *V(t)* структурной единицы (Рис. 2). Результатом является информация, изменяющая состояние выхода Y(t). Особенностью такой системы является возможность изменения состояния управляемого входа системы *U(t)* в зависимости от полученных результатов и благодаря наличию обратной связи R(t). Пользователь *P* является, с точки зрения комплексного подхода к визуальному анализу, обязательным участником структурной единицы. Влияние пользователя проявляется в изменении состояний управляемых входов и выходов, а также в регулировании работы системы S и принятии решения о завершении ее функционирования.

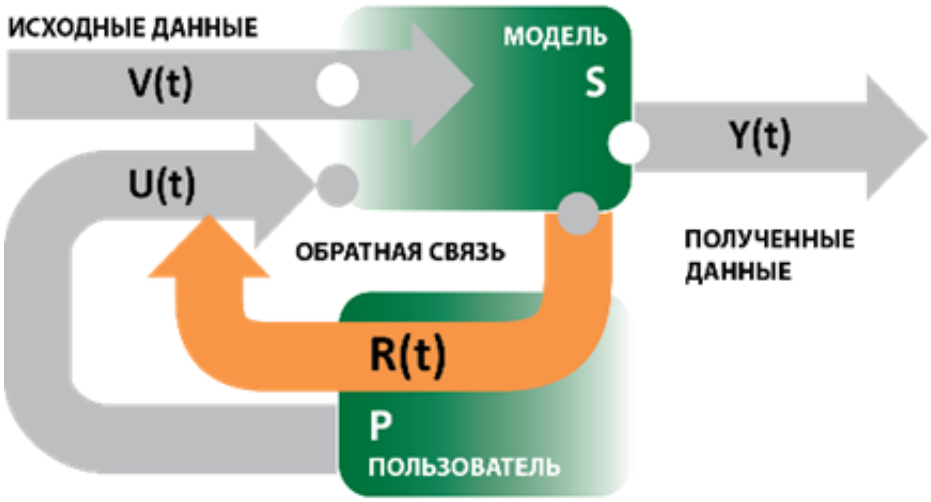


Рисунок 12 - Структурная единица визуального анализа

6.3 Типы структурных единиц

Структурную единицу визуальной модели можно представить простой логической схемой (рисунок 13А). Исходные данные, возникшие в условии задачи или ставшие ответом на предыдущий промежуточный вопрос, подаются на неуправляемый вход (*IN*) и формируют первоначальный образ данных, т.е. определяют состояние модели *S*. В результате управляющего взаимодействия пользователя с образом, формируется новый образ, соответствующий не исходным данным, а пониманию их смысла исследователем. Измененное состояние визуальной модели соответствует объему новых данных, передающихся на выход (*OUT*). Если полученные сведения являются необходимым ответом на общий вопрос анализа, то процесс заканчивается, а в остальных случаях происходит передача данных на вход следующего элемента визуальной модели.

В процессе формирования образа ответа формируется запрос на привлечение дополнительных сведений, которые поступают на управляемый вход (*IN2*). Кроме того, осмысление даже отдельного визуального факта способно создать ассоциативные гипотезы, являющиеся ответами на вопросы, не относящиеся к теме исследования. Это данные, которые могут фиксироваться в виде воспринимаемых фрагментов образа данных и запоминаться пользователем, поступая на выход структурной единицы (*OUT2*), изменяя собственную внутреннюю информированность пользователя и оказывая влияние на поиск новых ответов.

Логическая схема элемента позволяет декларировать отличия между структурными единицами на основании различий в уровне информационной активности существующих входов и выходов. Введение параметра информационной активности в качестве характеристики элементов в логической схеме структурной единицы создает возможность определения типов визуальных моделей. Отличия структурных единиц между собой определяют, в свою очередь, их возможности, назначение, правила функционирования и, следовательно, применимость для решения различных видов задач:

1. В одном из простейших случаев, визуальная модель данных выполняет функцию индикатора, оповещающего наблюдателя о состоянии изучаемой системы. Целью такой модели данных является информирование исследователя о наступлении ожидаемого события без предоставления излишних сведений. Таким образом, визуальная модель такого типа, отличается высокой активностью основного входа, отсутствием необходимости в использовании дополнительных связей и наличием дискретного спектра

значений на основном выходе (рисунок 13Б). Основной областью применения этих моделей становятся задачи визуального информирования.

2. В следующем случае, использование внешних связей структурного элемента визуальной модели принимает более сложный характер. Построение визуальной модели происходит с целью передачи наблюдателю наиболее полной информационной картины. В упрощенном виде, задачей визуальной модели является передача информации, достоверность которой не подвергается сомнению пользователем. Вторым требованием становится создание условий для адекватной интерпретации образа. В большинстве случаев, представление о выразительных средствах, необходимых для достижения этих целей, существует заранее. Таким образом, управление восприятием наблюдателя становится процессом, регулирование которого возлагается на пару дополнительных связей (*IN2-OUT2*). Модели этого типа используются в задачах обучения и характеризуются высокой скоростью информирования пользователя (рисунок 13В).

3. Визуальные модели данных получили широкое применение в системах поддержки принятия решений. В этом случае, задача анализа требует быстрого осмысления поступающих данных. Определяющим условием для моделей данных является активное использование знаний и опыта наблюдателя для получения выводов, оказывающих влияние на дальнейшее существование исследуемой системы. Для этого типа визуального анализа характерно активное использование дополнительных сведений, знаний пользователя (*IN2*) и формулирование новых промежуточных вопросов (*OUT2*), которые играют роль управляющего воздействия на модель. Информационный вход модели (*IN*) используется для предоставления сведений, необходимых для построения визуального образа в форме, согласующейся с формой управляющего вопроса (рисунок. 13Г). Это означает, что воспринимаемые компоненты модели становятся объектами, инициирующими построение гипотез ответа, и потому выбор типа визуального представления имеет принципиальное значение.

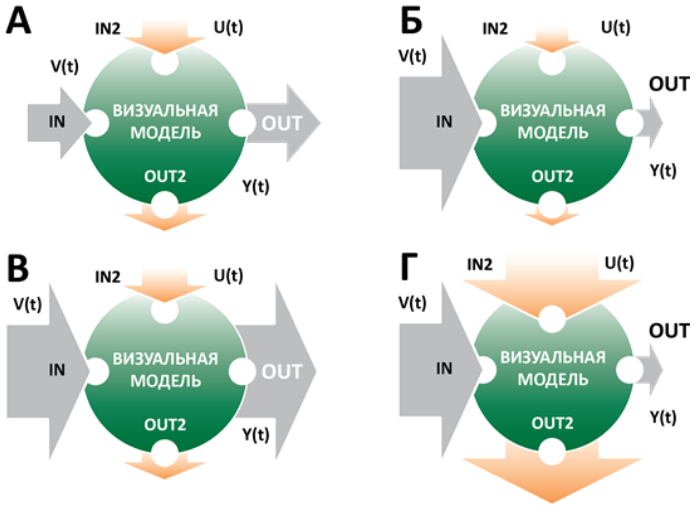


Рисунок 13 - Схемы различных структурных единиц

Кроме этого, комбинирование структурных единиц, в том числе, различного вида, позволяет проектировать средства визуального анализа, позволяющие решать задачи более сложных видов, для которых, например, не существует точно сформулированного вопроса анализа или отсутствуют однозначные критерии выбора правильного ответа. Представление о структуре визуального анализа и возможность выбора свойств для разрабатываемой визуальной модели, позволяет анализировать и, что намного важнее, прогнозировать эффективность использования средств визуального анализа.

6.4 Комбинированная модель

При исследовании сложных и объемных данных, также как и при изучении особенностей изменяющихся состояний, возможности одной визуальной модели могут оказаться недостаточными. Например, это происходит в тех случаях, когда визуальный образ не может, по любым причинам, содержать в себе весь объем необходимых сведений, либо, когда технически это возможно, но адекватная интерпретация вызывает у пользователя затруднения. Решением, позволяющим найти выход из этого затруднения, может быть построение набора визуальных моделей, каждая из которых соответствует некоторому фрагменту исходного объема данных, либо отвечает на вопрос анализа лишь частично. В результате, происходит построение сложной визуальной модели, имеющей более широкие возможности (рисунок 4).

Объединение структурных единиц в функционирующую систему, приводит к построению визуальной модели, когнитивное значение которой превосходит результативность исследования отдельных образов. Эмерджентность визуальной модели обеспечивается связями составляющих ее элементов, поэтому особое значение приобретает выделение и описание именно этих свойств структурных единиц. Условно, они могут быть разделены на две функциональные группы: информативные и управляющие связи.

Для объединения отдельных структурных элементов в единую систему необходимо наличие правила, обеспечивающего достижение результата анализа в пределах сформулированных граничных условий задачи анализа. Для изменяющихся данных, таким правилом может быть принцип хронологической последовательности, позволяющий наблюдать и сравнивать динамические характеристики элементов изучаемой системы. Однако, использование в качестве объединяющего правила естественных закономерностей, связанных с происхождением данных, не является обязательным.

Управляющие связи структурных единиц, ответственные в визуальных моделях за верификацию появляющихся гипотез, могут, а иногда и обязаны учитывать при выполнении своих функций всю информацию, полученную пользователем на уже пройденных этапах анализа. С точки зрения повышения результативности проводимого анализа, это обстоятельство может стать негативным фактором. Более сложная процедура проверки связана с дополнительными затратами времени, являющимся одним из основных критериев результативности визуальной модели. Это потребует уточнения разработанной системы образов для достижения сбалансированного решения.

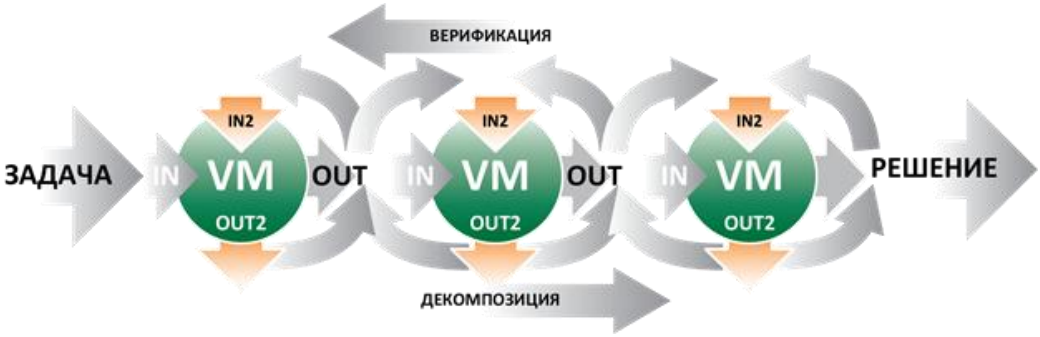


Рисунок 14 - Схема комбинированной модели

В описанной таким образом логической структуре визуальной модели наряду с последовательностью наблюдаемых пользователем образов данных формируются еще два функциональных объекта, участвующих в визуальном анализе. Первым является цепочка промежуточных вопросов, руководящих процессом исследования данных и

ответственных за его логичность и обоснованность. Вторым объектом становится управляющая последовательность, состоящая из процедур верификации, активно использующих возможности визуального восприятия и эмоциональной оценки наблюдаемых исследователем образов. Успешность применения модели для решения задачи анализа достигается сбалансированным и полноценным использованием всех логических элементов, образующих структуру функционирования визуальной модели.

6.5 Методический подход к использованию аналитических визуальных моделей

Структура визуальной модели, предназначенной для проведения анализа данных в задачах с низким уровнем формализации, определяет последовательность и логику рассуждений пользователя. Таким образом, свойства визуальной модели оказывают прямое влияние на получаемое решение. Систематизация правил, регулирующих процедуры создания визуальных моделей и использования их при решении задач анализа, позволит увеличить результативность анализа и определить направления развития для применения визуальных средств исследования данных [9].

Следует обратить внимание на то обстоятельство, что результативность каждой из составляющих визуальной модели (функциональной, информативной и управляющей) зависит от использования определенных возможностей пользователя. В случае сложной модели, последовательное приближение к решению задачи приводит к непрерывному и обязательному присутствию пользователя в процессе анализа. Таким образом, полнота логической схемы функционирования визуальной модели не может быть достигнута без включения в нее четвертой обязательной составляющей – возможностей и действий конкретного пользователя, влияющих на результат анализа на протяжении прохождения всей последовательности этапов визуального анализа.

Основной идеей, обеспечивающей увеличение когнитивной результативности визуального анализа, становится направленное использование возможностей пользователя, реализуемое благодаря свойствам визуальной модели. Это создает возможность привлечения дополнительного ресурса для решения поставленной задачи анализа. Перспективными направлениями в этом случае становятся привлечение не только информированности пользователя, но и особенностей его мышления, как для формирования гипотезы решения, так и для принятия решения о ее достоверности.

Наличие интерактивного управления визуальным образом обеспечивает прямое участие пользователя в манипулировании образом. Наличие системы оперативного управления создает условия для постановки в ходе анализа новых вопросов исследования и быстрого получения ответов. В свою очередь, появление в распоряжении пользователя новых сведений не только ускоряет достижение цели анализа, но и формирует у исследователя представление о дальнейших действиях. Таким образом, дополнение визуальной модели управляющим интерфейсом, соответствует предложенной структуре визуального анализа и позволяет сделать его инструментом аналитического решения любой поставленной задачи.

В случае исследования данных, относящихся к вопросам, не имеющим полноценного формального описания структуры соответствующих знаний, использование визуальных моделей становится не только способом обнаружения новых закономерностей, но и формирует понимание их смысла. Следовательно, приобретение визуальной моделью статуса аналитического инструмента возможно лишь в том случае, когда созданы условия для интерактивного управления способом представления данных с ее помощью. Визуальная модель, предлагающая наблюдателю на первоначальном этапе исследования образ не только исходных данных, но и обнаруженных закономерностей, может использоваться в качестве *формальной визуальной модели* области знаний.

Основными критериями результативности использования визуальной модели выступают два параметра: достоверность полученного ответа и время, затраченное на его формулирование. Наличие в структуре процедуры анализа нескольких функциональных элементов и их взаимное влияние приводит к необходимости определения правил, соблюдение которых необходимо для получения наиболее эффективной комбинации. С учетом динамичного характера проведения процедуры анализа, соблюдение и адаптация этих правил к конкретным условиям задачи становятся еще одной целью интерактивного управления визуальной моделью.

Таким образом, аналитический потенциал визуальной модели определяется выбором способов визуального моделирования изучаемых данных, максимально возможной скоростью построения зрительного образа при сохранении всех необходимых, с точки зрения функциональности, нюансов визуального представления. Кроме этого, возможно, наиболее значимым элементом модели становится ее интерфейс, необходимый для использования информативного и когнитивного потенциала самого пользователя.

На основании сделанных утверждений можно сформулировать ряд методических рекомендаций, позволяющих системно использовать преимущества визуальных моделей при решении задач исследования данных:

1. Предварительный анализ условий задачи необходим для определения возможного типа визуальной модели. Если проведенная оценка уровня сложности задачи позволяет говорить о необходимости создания аналитической визуальной модели, то процедура решения задачи начинается с декомпозиции основного вопроса исследования. Декомпозиция проводится до получения последовательности связанных элементарных вопросов, т.е. вопросов, на которые может быть найден однозначный ответ. Для построения общего решения необходимо выделение начальной формулировки проблемы исследования.

2. Выявление ограничений, накладываемых на поиск желательного решения, формирует объем данных, оказывающих влияние на характеристики визуальной модели. Авторы средств визуального анализа должны обеспечить возможность участия этих данных в процедуре поиска решения задачи анализа на каждом этапе. Целью такого участия данных, не имеющих непосредственного отношения к задаче анализа, является выбор характеристик и управление визуальной моделью. В числе подобных ограничений рассматриваются, помимо особенностей задачи, допустимые затраты времени и ресурсов.

3. Определение особенностей восприятия пользователя или группы пользователей влияющих на форму визуального решения. Это необходимо для бесконфликтного участия исследователя в процедуре визуального анализа. Выбор выразительных средств происходит как решение задачи соответствия между языковой формой изложения и способностью пользователя к ее адекватной интерпретации.

4. Определение типа визуальной модели на основании проведенной декомпозиции проблемы исследования и дополнения условий задачи анализа дополнительными сведениями. Формирование начальной формы визуальной модели, соответствующей постановке исходного вопроса, граничным условиям и выбору выразительных средств.

5. Создание средств управления визуальной моделью, обеспечивающих использование когнитивного потенциала исследователя при переходе к состоянию визуальной модели, интерпретация которого верифицируется пользователем в качестве ответа на вопрос исследования. Система управления выбирается как инструмент сокращения времени анализа и выбора наилучшего, с точки зрения исследователя, решения.

6. Расширение базы знаний, доступной исследователю, сведениями о результативности использованного типа визуальной модели. Полученные сведения принимают участие в определении типа визуализации при решении других задач.

Комбинирование единичных элементов позволяет получать информативные, визуально воспринимаемые объекты, которые могут быть использованы для решения

самых разных задач. В их число входят визуальное представление информации, целенаправленное использование когнитивного потенциала исследователя, активизируемое воздействием образного восприятия, исследование данных, не имеющих формального описания или объяснения. Особый интерес представляет задача построения визуальных моделей, содержащих данные целых предметных областей. Это позволяет проводить визуальный анализ данных, имеющих различное происхождение, достоверность, формат и смысл, с целью их обобщения, верификации и поиска объяснения.

В качестве практической задачи, для исследования результативности использования визуальных моделей при изучении многомерных данных, была решена задача представления общей совокупности знаний в предметной области, содержащей экспериментальные сведения. Осложняющими обстоятельствами для подобной задачи стали большое число разнородных источников информации (публикации) и различающиеся структуры сравниваемых данных. Источником исходной информации в базе исследуемых данных служили статьи, опубликованные в мировых рецензируемых изданиях.

В результате, построена *общая визуальная модель* эмпирических данных о современном состоянии изучения процессов получения карбида кремния электродуговым методом (рисунок 15). Модель представляет собой набор горизонтальных плоскостей, каждая из которых соответствует одной из характеристик исследуемых объектов. Линия, проведенная через точки, соответствующие одному объекту становится визуальным образом, характеризующим этот объект. Множество таких визуальных объектов составляет визуальную модель, позволяющую проводить сравнение объектов между собой или анализировать весь объем исследуемых данных одновременно. Одним из значимых результатов проведенного анализа стало определение стратегии развития предметной области. Кроме того, проведена оценка достоверности опубликованных данных и изменений областей концентрации внимания исследователей указанной предметной области, происходящие с течением времени.

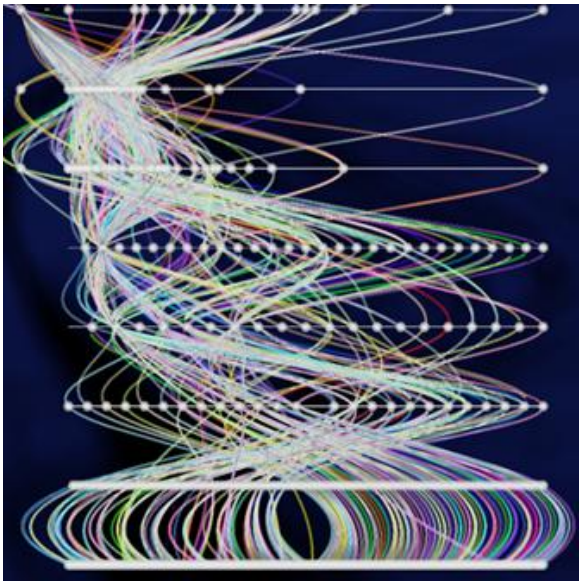


Рисунок 15 - Визуальная модель эмпирических данных

С помощью построенной визуальной модели проанализированы сведения о 260 экспериментах, посвященных получению углеродных ультрадисперсных материалов в плазме электрической дуги постоянного тока. Получены выводы о росте объема

экспериментальных данных преимущественно в отношении получения четырех наиболее популярных продуктов. При сравнении групп отдельных параметров (показатели эксперимента – получаемый продукт) обнаружено, что производительность рассматриваемых экспериментальных систем практически не оценивается за исключением обрывочных несистематизированных данных, что говорит о неготовности исследуемого метода к внедрению в промышленное производство.

В результате проведенного исследования были подтверждены основные предположения, описывающие структуру визуального анализа. Благодаря созданию системы интерактивного управления визуальной моделью, достигнута высокая результативность анализа. Это связано с сокращением времени анализа, по сравнению с традиционными способами изучения подобных данных, для которых время исследования измеряется несколькими днями, даже с учетом опыта и квалификации специалистов.

По совокупности представленных сведений в отношении анализа диаграммы энергетических характеристик получен вывод о том, что полем для дальнейших исследований метода синтеза углеродных материалов в плазме дуги постоянного тока должна стать взаимосвязь уровня напряжения и электрической мощности с другими параметрами экспериментальных установок для получения материала с заданным фазовым составом и свойствами в контексте требований к электрической мощности системы и объема потребления электроэнергии. Основаниями для подобных выводов являются точки концентрации данных на определенных величинах выбранных параметров (рисунок 16), а также наблюдаемые в модели явные области недостаточной изученности.

Предложенный подход к решению задач анализа с помощью визуальных моделей данных позволяет оперативно получать ответы на вопросы исследования, в том числе, высокой степени сложности. Построение визуальной модели, предоставляющей возможность отображать исходные данные и формулировать промежуточные вопросы с целью поиска неизвестных ранее внутренних закономерностей, дает пользователю аналитический инструмент для исследования поставленной перед ним задачи. Помимо возможности достичь цели анализа с помощью этого инструмента, пользователь получает возможность существенно сократить время, затрачиваемое им на поиск ответа или принятие решения в ситуации выбора наилучшего варианта. Сделанные утверждения проверены при анализе опубликованных данных, посвященных получению углеродных ультрадисперсных материалов.

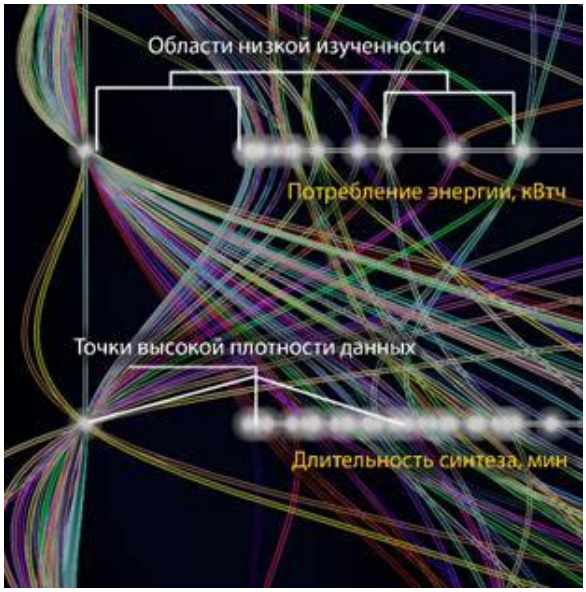


Рисунок 16 - Визуальный поиск закономерностей

Рекомендуемая литература:

1 А.А. Захарова, Е.В. Вехтер, А.В. Шкляр Решение задач анализа данных с помощью аналитических визуальных моделей http://sv-journal.org/2017-4/08/index.php?lang=ru

2 Bondarev A.E., Galaktionov V.A.: Multidimensional data analysis and visualization for time-dependent CFD problems. Programming and Computer Software 41(5), 247–252 (2015). DOI: 10.1134/S0361768815050023.

3 Michelle A. Borkin, Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, Hanspeter Pfister, What Makes a Visualization Memorable?, IEEE Transactions on Visualization and Computer Graphics, v.19 n.12, p.2306-2315, December 2013 [DOI: 10.1109/TVCG.2013.234]

4 Chen, C.: Mapping Scientific Frontiers: The Quest for Knowledge Visualization. 2nd ed. London: Springer, (2013).

5 Chen C.: Top 10 unsolved information visualization problems. IEEE Computer Graphics and Applications 25(4), 12–16 (2005).

6 Eppler, M., Burkhard, R.A.: Visual representations in knowledge management: Framework and cases. Journal of Knowledge Management, vol. 11(4), pp. 112-122. QEmerald Group Publishing Limited (2007). ISSN 1367-3270 DOI 10.1108/1367327071076275.

7 Huang, D., Tory, M., Adriel Aseniero, B., Bartram, L., Bateman, S., Carpendale, S., Tang, A., Woodbury, R.: Personal visualization and personal visual analytics. IEEE Trans. Vis. Comput. Graph. 21, 420–433 (2015).

8 Klimentov, A.; Buncic, P.; De, K.; Jha, S.; Maeno, T.; Mount, R.;Nilsson, P.; Oleynik, D.; Panitkin, S.; Petrosyan, A.; Porter, R. J.; Read,K. F.; Vaniachine, A.; Wells, J. C.; Wenaus, T.; Iop, Next GenerationWorkload Management System For Big Data on HeterogeneousDistributed Computing. Proceedings from the 16th InternationalWorkshop on Advanced Computing and Analysis Techniques in PhysicsResearch, Prague, Czech Republic, September 1−5, 2014; IopPublishing Ltd: Bristol, U.K., 2015; Vol. 608.

9 Korobkin D., Fomenkov S., Kravets A., Kolesnikov S., Dykov M.: Three-Steps Methodology for Patents Prior-Art Retrieval and Structured Physical Knowledge Extracting. In: Creativity in Intelligent Technologies and Data Science First Conference, CIT&DS 2015, pp.124-138. Volgograd, Russia (2015).

10 Sedig Kamran, Parsons Paul, Hai-Ning Liang, Jim Morey: Supporting Sensemaking of Complex Objects with

11 Visualizations. Visibility and Complementarity of Interactions, Informatics 3(4), 20 (2016). doi:10.3390/informatics3040020.

12 Shneiderman, B., Plaisant, C., Cohen, M.S., Jacobs, S.M., Elmqvist, N., Diakopoulos, N.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. 6th ed. Pearson: Upper Saddle River, NJ, USA (2016).

Тема 7 Современные программные средства анализа больших объемов информации

Цель: рассмотреть современные популярные программные средства анализа данных: Statistica, Excel, их преимущества и недостатки

План:

7.1 Обзор программного средства анализа данных: Statistica

7.1 Обзор программного средства анализа данных: Statistica

STATISTICA — это интегрированная система анализа и управления данными. STATISTICA — это инструмент разработки пользовательских приложений в бизнесе, экономике, финансах, промышленности, медицине, страховании и других областях. STATISTICA легка в освоении и использовании.

Все аналитические инструменты, имеющиеся в системе, доступны пользователю и могут быть выбраны с помощью альтернативного пользовательского интерфейса. Пользователь может всесторонне автоматизировать свою работу, начиная с применения простых макросов для автоматизации рутинных действий вплоть до углубленных проектов, включающих в том числе интеграцию системы с другими приложениями или Интернет. Технология автоматизации позволяет даже неопытному пользователю настроить систему на свой проект.

Процедуры системы STATISTICA имеют высокую скорость и точность вычислений.

Гибкая и мощная технология доступа к данным позволяет эффективно работать как с таблицами данных на локальном диске, так и с удаленными хранилищами данных.

Система обладает следующими общепризнанными достоинствами:

- содержит полный набор классических методов анализа данных: от основных методов статистики до продвинутых методов, что позволяет гибко организовать анализ;
- является средством построения приложений в конкретных областях;
- в комплект поставки входят специально подобранные примеры, позволяющие систематически осваивать методы анализа;
- отвечает всем стандартам Windows, что позволяет сделать анализ высокоинтерактивным;
- система может быть интегрирована в Интернет;
- поддерживает web-форматы: HTML, JPEG, PNG;
- легка в освоении, и как показывает опыт, пользователи из всех областей применения быстро осваивают систему;
- данные системы STATISTICA легко конвертировать в различные базы данных и электронные таблицы;
- поддерживает высококачественную графику, позволяющую эффектно визуализировать данные и проводить графический анализ;
- является открытой системой: содержит языки программирования, которые позволяют расширять систему, запускать ее из других Windows-приложений, например, из Excel.

STATISTICA состоит из набора модулей, в каждом из которых собраны тематически связные группы процедур. При переключении модулей можно либо оставлять открытым только одно окно приложения STATISTICA, либо все вызванные ранее модули, поскольку каждый из них может выполняться в отдельном окне (как самостоятельное приложение Windows).

При исполнении модулей STATISTICA как самостоятельных приложений в любой момент времени в любом модуле имеется прямой доступ к «общим» ресурсам (таблицам данных, языкам BASIC и SCL, графическим процедурам).

Командный язык STATISTICA (SCL)

STATISTICA содержит два встроенных языка программирования STATISTICA BASIC и SCL (командный язык). Оба языка предназначены для работы в среде STATISTICA и содержат встроенные операции для обращения к таблицам исходных данных, таблицам результатов и графическим функциям.

Язык STATISTICA BASIC представляет собой простой и одновременно достаточно мощный язык программирования. С его помощью можно создать широкий спектр

приложений, начиная от простых программ преобразования данных и кончая сложными пользовательскими процедурами комплексного анализа и вывода информации.

Этот язык программирования пригоден для решения больших вычислительных задач, поскольку обрабатываемые массивы данных могут иметь до 8 измерений и нет ограничений на размеры массивов. Таким образом, пользователь может использоваться всю доступную память и создавать процедуры, включающие операции с большими многомерными матрицами.

Встроенный язык STATISTICA BASIC доступен в любой момент анализа вместе с интегрированной средой, которая позволяет писать, редактировать, проверять, отлаживать (предварительно прогонять) и выполнять программы.

Язык STATISTICA BASIC, как обычный язык программирования, поддерживает циклические операции и условные переходы, функции и подпрограммы, а также работу с динамическими библиотеками (DLL). В то же время он «понимает» структуру файлов данных системы STATISTICA и позволяет организовать интерактивную обработку данных в среде самой системы с помощью пользовательских диалоговых окон. С помощью этого языка пользователь может создавать свои собственные сложные программы анализа данных, одновременно используя готовые алгоритмы расчетов и построения графиков, предусмотренные в системе STATISTICA.

Командный язык SCL (STATISTICA Command Language) предназначен для организации пакетной обработки данных и создания собственных приложений на основе процедур, содержащихся в системе STATISTICA. Для того чтобы пользователь мог при этом реализовать собственные алгоритмы расчетов, предусмотрена возможность интеграции языков STATISTICA BASIC и SCL.

Программы, написанные на встроенных языках системы STATISTICA, доступны в любом модуле системы и на любом этапе анализа данных, при этом их можно вызывать и выполнять как с помощью кнопок автозадач, так и непосредственно из окна редактирования. Пользователь также имеет возможность создавать собственные библиотеки функций и подпрограмм и таким образом значительно расширять предлагаемый набор процедур обработки данных и представления результатов.

Ввод и исполнение SCL-программ. STATISTICA может работать в «истинном» пакетном режиме как система, управляемая командами, с помощью встроенного языка управления приложениями SCL (STATISTICA Command Language), доступного в любом модуле системы из выпадающего меню Анализ. Можно ввести последовательность команд для выполнения определенных действий, а затем сколько угодно раз исполнять ее в пакетном режиме.

Возможен и другой способ действий — использование диалогового окна **Мастер команд** для быстрого выбора и ввода требуемого списка команд.

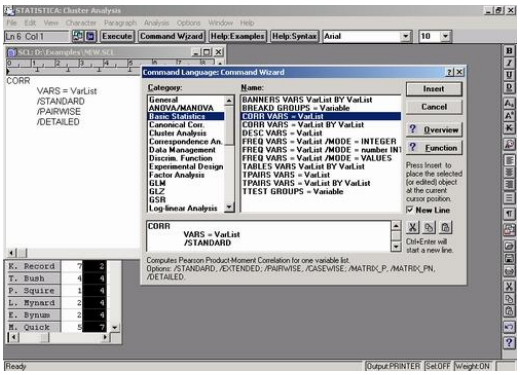


Рисунок 17 - Диалоговое окно Мастер команд

Для написания и отладки «пакетов» команд используется интегрированная среда языка SCL. Она включает текстовый редактор, совмещенный с окном **Мастер команд** (см. иллюстрацию выше — кнопка **Мастер команд** на панели инструментов **Командного языка**), систему помощи по синтаксису языка с примерами и интегрированные средства проверки правильности программ (доступны из выпадающего меню **Сервис**).

Пользовательские расширения языка SCL. Программы на языке SCL могут включать не только предопределенные параметры и команды для выполнения действий по статистической обработке, управлению и графическому выводу данных (см. кнопки **Справка: примеры** и **Справка: синтаксис** на панели инструментов), но и пользовательские «команды», определенные с помощью инструмента Назначить клавиши **(SendKeys)** (в соответствии с правилами, принятыми в MS Visual BASIC).

Написанные таким образом программы могут выполнять, например, операции с буфером обмена **(Копировать, Вставить**), менять параметры вывода, принятые по умолчанию в различных процедурах, и выполнять другие функции.

SCL -программы могут также включать в себя программы и процедуры, написанные на языке STATISTICA BASIC (языке STATISTICA, предназначенном для преобразования данных и графиков и управления ими, который доступен из любого модуля пакета). Например, определенные пользователем графические или вычислительные процедуры на языке STATISTIC A B ASIC могут выполняться как часть пакета команд SCL.

Пользовательский интерактивный интерфейс для SCL-программ. Несмотря на то что в командном языке SCL не заложен в непосредственном виде специальный пользовательский интерактивный интерфейс, тем не менее для этих целей можно использовать программы на языке STATISTICA BASIC, вызываемые из SCL-nporpaмм, например, для создания диалоговых окон, позволяющих выбирать переменные, файлы данных и т. п. в ходе выполнения программы (см. примеры в Электронном руководстве).

Исполняемый модуль STATISTICA. Командный язык содержит специальный Исполняемый модуль, позволяющий разрабатывать приложения «под ключ», которые вызываются двойным щелчком на значке соответствующего «пользовательского приложения» на рабочем столе Windows.

Эта возможность позволяет экономить время пользователя, когда многократно повторяется одна и та же процедура или последовательность процедур анализа, а также дает возможность использовать SCL-программы пользователями, которые не знакомы с соглашениями системы STATISTICA.

Чтобы создать такое приложение «под ключ», сначала нужно написать саму SCL-программу и сохранить ее обычным образом (например, в файле Program 1.sct). Затем в окне Диспетчер программ системы Windows нужно создать пиктограмму для исполняемого модуля с именем Sta_run.exe (оно находится в папке STATISTICA на диске).

В поле команд нужно задать имя SCL-программы, подлежащей исполнению (например, d:\data\program1.scl'). Теперь при щелчке мышью на этом значке будет начинаться выполнение программы (в данном случае Program1.se!). Описанным способом можно создать любое количество пользовательских приложений, а с помощью окна Диспетчер программ дать им содержательные имена, соответствующие тем задачам анализа данных, которые эти приложения выполняют.

Кнопки автозадач — это всплывающая настраиваемая панель инструментов (включить или выключить ее можно клавишами CTRL+M).

Кнопки на этой панели инструментов можно назначить/переопределить с помощью кнопки **Настройка**... (или нажатия на соответствующую кнопку при удерживаемой клавише CTRL). В диалоговом окне, которое при этом открывается, можно присвоить имена уже имеющимся и новым кнопкам.

Перейдем к более систематическому изложению.

Часто при выполнении сложной задачи возникает необходимость выполнять одну и ту же последовательность действий, например, открывать ранее сохраненные графики, данные или листинги программ. Постоянная потребность выполнять мало относящиеся к основной работе операции может отнимать время или даже раздражать. В системе STATISTICA предусмотрены возможности, которые избавляют пользователя от однообразных операций и способствует созданию комфортных условий работы.

Кнопки автозадач — это настраиваемая панель, которую в случае необходимости вы легко можете убрать с экрана или снова восстановить (восстановить или скрыть эту панель можно с помощью комбинации кнопок CTRL+M). На панели **«Кнопки автозадач»** нажмите кнопку **Настройка...** Откроется окно настройки кнопок автозадач. В центральной части окна расположен столбец кнопок, позволяющий:

- Изменить или задать кнопку. Нажав на эту кнопку, вы можете задать последовательность нажатий кнопок клавиатуры. Для организации такой последовательности достаточно нажать кнопку **Запись** в правой части диалогового окна. С этого момента система автоматически начнет запоминать и переводить на язык команд ваши действия. Нажав, например, на клавиатуре кнопку Alt, вы попадете в главное меню, по которому сможете передвигаться с помощью стрелок и клавиши Enter. Свободно перемещаться внутри диалоговых окон вам поможет клавиша Tab и т. д. Для окончания записи нажмите CTRL+F3. В нижней части окна **Настройка кнопок автозадач** будут описаны кнопки перемещений по окнам и соответствующий им синтаксис.
- Удалить кнопку. В любой момент времени вы можете удалить ставшую ненужной кнопку. Задать последовательность функций или операций на Командном языке STATISTICA (SCL).
- Использовать написанные на языке STATISTICA BASIC процедуры вычислительного характера, преобразования данных, операции по управлению данными, графические процедуры, а также процедуры, написанные на любом другом языке программирования, вызываемые из STATISTICA BASIC.
- Открывать файлы данных и любые вспомогательные файлы системы STATISTICA.
- Создавать и редактировать макрокоманды (последовательности нажатий клавиш), соответствующие часто выполняемым процедурам, заданиям или настройкам. Такие редактируемые команды можно вводить в текстовом виде или, например, как последовательности движений мышью.

В каждом из описанных выше окон предусмотрена возможность создания сочетаний «горячих клавиш». Вы можете назначить сочетание клавиши CTRL и любой буквы от A до Z или цифры от 0 до 9. После сохранения этой установки вам будет достаточно нажать определенную комбинацию клавиш, что будет равносильно нажатию на кнопку автозадачи.

Панель инструментов может быть глобальной или локальной и содержать большие библиотеки пользовательских заданий и процедур. Локальная панель инструментов связана с конкретным модулем или проектом. Имя открытой в данный момент панели высвечивается в строке заголовка диалогового окна.

Настроенную панель инструментов **Кнопки автозадач** можно затем сохранить, используя команды диалогового окна **Настройка**....

Панель инструментов **Кнопки автозадач** можно использовать как удобный интерфейс для пользовательских расширений стандартных процедур.

STATISTICA постоянно развивается, открывая новые возможности для пользователей. Если говорить кратко, то развитие системы происходит в духе развития современных Windows-технологий. Гибкая настраиваемость для задач конкретного проекта, широкий набор статистических опций, доступных пользователю из других приложений, глобальная интеграция с другими приложениями, например, с помощью VB,

C++, Java, оптимизация для Web и мультимедийных приложений — ближайшие перспективы STATISTICA. В таблицы с данными (мультимедийные электронные таблицы) можно будет встраивать различные объекты: звук, фото и т. д.

Первые шаги в системе STATISTICA

Наше знакомство с системой STATISTICA, конечно, следует начать с ввода данных. Вы увидите, как легко вводятся в STATISTICA самые разнообразные данные. Предполагается, что система STATISTICA установлена на вашем компьютере и вы последовательно повторяете описываемые действия.

В качестве конкретной области выберем медицинский пример.

Как вы уже знаете, исходные данные в системе STATISTICA организованы в виде таблиц. Если у вас имеется опыт работы с электронными таблицами (типа MS Excel), то вы быстро привыкнете к таблицам STATISTICA. Заметим, что табличная структура данных STATISTICA позволяет естественно отобразить большинство реальных данных.

Электронная таблица состоит из строк и столбцов. Столбцы таблицы STATISTICA называются **Variables — Переменные**, а строки **Cases — Наблюдения.**

Например, в медицине наблюдения — это пациенты, переменные — пол, возраст, дата поступления в больницу, дата диагноза, дата операции, перевода в другую больницу, выписки и т. д. Вы можете представить такую таблицу как страницу записной книжки врача, где строки — это, например, имена пациентов, столбцы — характеристики (переменные, описывающие течение болезни).

Для того чтобы создать таблицу с данными, проделайте следующее:

1. Запустите программу STATISTICA.

2. Откроется меню**Статистических модулей** (STATISTICA Module Switcher).

3. Выберите из меню модуль **Основные статистики и таблицы** и щелкните по нему мышью.

4. Теперь вы находитесь в модуле **Основные статистики и таблицы**, в котором можете выбрать любую статистическую процедуру, входящую в этот модуль. Но поскольку у вас другая цель, просто щелкните мышью по кнопке **Выход** (Cancel).

Итак, вы находитесь в рабочем окне модуля **Основные статистики и таблицы** системы STATISTICA. В основном рабочем окне системы подведите курсор мыши к строке меню **Файл** и щелкните левой кнопкой. В выпадающем меню выберите команду **Создать данные.** На экране компьютера сразу же появляется окно **Создание данных** (см. рисунок ниже).

В этом окне можно ввести имя файла, например, medicine 1 .sta (файл может быть назван и по-русски, однако по ряду причин целесообразнее использовать английские имена).

Теперь поместите курсор мыши в поле **File name — Имя** файла и наберите с клавиатуры нужное имя.

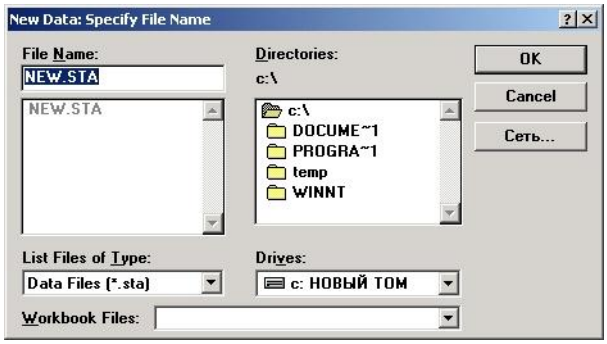
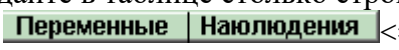


Рисунок 18 - Поле **File name**

После нажатия клавиши **Enter** на клавиатуре или кнопки **Save** программа создаст пустую таблицу, содержащую 10 строк и 10 столбцов.

Вы легко можете увеличить или уменьшить как количество строк, так и количество столбцов этой таблицы. Создайте в таблице столько строк и столбцов, сколько нужно. Для этого используйте кнопки 

<=”” img=”” style=”padding: 0px; margin: 0px; border: none;”> на панели инструментов.

Нажмите, например, кнопку **Наблюдения**. После нажатия кнопки на экране возникнет меню, предлагающее следующий выбор для наблюдений таблицы: **Добавить, Переместить, Копировать, Удалить, Ввести имена наблюдений**. Выберите, например, пункт **Добавить**, дважды щелкнув левой кнопкой мыши. Откроется окно, в котором можно задать число наблюдений, добавляемых в таблицу:

Нажмите **ОК**, и количество строк (наблюдений) в таблице увеличится на 2, то есть станет равным 12. Аналогичным образом измените число переменных в таблице. В данном случае понадобятся 11 переменных. Нажмите кнопку **Переменные** на панели инструментов. С помощью курсора мыши в выпадающем меню выберите пункт **Добавить**. На экране появится окно, где выполните установки, как показано ниже.

Нажмите еще раз кнопку **Наблюдения** и выберите пункт меню **Имена**. На экране появится диалоговое окно, в котором можно определить, сколько символов в таблице будет зарезервировано для имен наблюдений. Раздвинуть поле для имен наблюдений можно также с помощью мыши.

Итак, вы сделали первый шаг к достижению цели — создали электронную таблицу, которая имеет 11 столбцов и 12 строк, а также место для ввода имен наблюдений (см. рисунок).

Теперь необходимо ввести название таблицы (ее заголовок) и имена переменных. Вы работаете, используя мышь и клавиатуру. Запомните основной принцип: дважды щелкая мышью по полям заголовков, вы открываете диалоговые окна, позволяющие вводить заголовки, описывать переменные и т. д. Введите заголовок таблицы. Для этого дважды щелкните мышью на верхней строке таблицы, пустой строке, которая находится над переменными. В появившемся окне введите заголовок таблицы.

Наберите с клавиатуры заголовок, нажмите **ОК**. Введенный текст отобразится в заголовке таблицы. В поле**Информация о файле и примечания** можно записать дополнительную информацию, которая будет полезна при работе с файлом.

Аналогично редактируются имена переменных и наблюдений. Например, чтобы ввести имена, необходимо дважды щелкнуть мышью в поле **Имя наблюдения** и в появившемся окне ввести имена пациентов:

Для того чтобы описать переменную, необходимо дважды щелкнуть мышью по ее имени — например, после щелчка по заголовку переменной1 (VAR1) откроется окно, в котором можно задать ее имя (или переименовать ее), формат переменной, метку, связь и т. д.

Теперь заполните созданную таблицу данными. Данные вводятся непосредственно с клавиатуры. Возможности экспорта, например в MS Word, мы обсудим позднее. Если нужно ввести числовые данные, используйте клавиатуру и стрелки перемещения курсора. Поставьте курсор на нужную ячейку таблицы и введите числовые данные. Текстовые значения вводятся иначе. Подведите курсор к ячейке переменной с текстовыми значениями и дважды щелкните мышью. В ячейке появится код 9999 — это код пропущенных значений. Сотрите код, используя кнопку DEL на клавиатуре. Затем введите нужное текстовое значение. В итоге можно получить следующую таблицу:

Таким образом, вы научились создавать таблицы и вводить в них данные. Повторив несколько раз описанные действия с другими данными, вы прочно закрепите полученные навыки.

Поскольку система STATISTICA является обычным Windows-приложением, можно легко и быстро импортировать данные, полученные в системе STATISTICA, в другое Windows-приложение, например в MS Word.

Лучше всего проделать это следующим образом: нажмите одновременно кнопки ALT и F3. На экране вместо курсора мыши появится значок «прицел». Используя мышь, поместите прицел в верхний левый угол таблицы. Затем нажмите левую кнопку мыши, зафиксируйте прицел и, удерживая кнопку мыши, переместите прицел в новое место таблицы. Выделенная часть таблицы будет отмечена прямоугольной рамкой. После того как вы отпустите кнопку мыши, отмеченная часть таблицы будет помещена в буфер обмена. Если теперь открыть нужный документ Word и набрать на клавиатуре комбинацию кнопок CTRL и V, то выбранный сегмент таблицы будет скопирован в документ.

Упражнение. Проведите сортировку данных файла medicinel.sta по возрасту пациентов и по городам. Используйте модуль **Управление данными** и опцию **Сортировка наблюдений.**

Из переключателя модулей системы STATISTICA запустите модуль **Основные статистики и таблицы.** Для этого выберите в меню модуль Основные статистики и таблицы и щелкните по нему мышью. Модуль будет выбран из списка модулей. Затем подведите курсор мыши к кнопке **Переключиться в** и нажмите ее. Произойдет запуск системы STATISTICA, и на экране появится рабочее окно модуля Основные статистики и таблицы. Именно в этом модуле мы будем работать.

В модуле **Основные статистики и таблицы** создайте файл данных, как показано на рисунке.

В файле содержатся результаты опроса 10 женщин (данные являются модельными) относительно их семейного положения и состояния уровня тревожности. Первая переменная **СЕМ_ПОЛ** описывает семейное положение женщин. Эта переменная принимает два значения: **П_семья** — полная семья, **Н_семья** — неполная семья. Вторая переменная, **ТРЕВОГА**, описывает самооценку личностной тревожности женщины. Она принимает два значения: низкая, высокая. Известно, что личностная тревожность характеризуется устойчивой склонностью воспринимать жизненную ситуацию как угрожающую (содержащую в себе тайную угрозу). Вы видите, что первая опрошенная женщина — наблюдение номер 1 (первая строка в таблице) — имеет полную семью и характеризует свое душевное состояние как тревожное. Вторая опрошенная женщина — наблюдение номер 2 (вторая строка таблицы) — имеет неполную семью и оценивает уровень своей тревожности как низкий и т. д.

Назовите этот файл women1sta.

Заметьте, переменные в этом файле принимают текстовые значения, что типично для социологических опросов.

Примите совет, позволяющий эффективнее организовать ввод текстовых данных. Переменные принимают текстовые значения, и если каждый раз вводить текст в таблицу, то это займет слишком много времени. Для удобства лучше численные значения, а затем перейти в текстовый режим, нажав кнопку на панели инструментов. Удобно закодировать значения переменных. Покажем, как это делается. Начнем с переменной**СЕМ_ПОЛ.** Дважды щелкните по ее заголовку левой кнопкой мыши, и на экране отобразится окно **Диспетчер текстовых значений — СЕМ_ПОЛ**.

В этом окне в колонке **Текст** наберите в первой строке **П_семья**, а в колонке **Число** наберите 1. Это приведет к тому, что текстовому значению **П_семья** будет присвоен код 1. Во второй строке **Диспетчера текстовых значений** наберите **Н_семья**, а в колонке **Число** наберите 2 — текстовому значению **Н_семья** будет присвоен код 2. Далее нажмите кнопку **ОК**.

Теперь введите значения 1 в те ячейки переменной **СЕМ_ПОЛ**, в которых должно стоять текстовое значение**П_семья**.

Введите значения 2 в те ячейки переменной **СЕМ_ПОЛ**, в которых должно стоять текстовое значение **Н_семья**.

Теперь достаточно нажать кнопку <=”” img=”” style="padding: 0px; margin: 0px; border: none;">на панели инструментов STATISTICA, чтобы получить нужные текстовые значения.

Точно таким же образом введите текстовые значения в ячейку переменной **ТРЕВОГА**

Итак, вы создали файл women1.sta. Теперь построим, исходя из этого файла исходных данных, таблицу сопряженности. Это очень легко сделать в STATISTICA.

Шаг 1. Подведите курсор мыши к пункту **Анализ**. Щелкните по нему мышью. В появившемся меню сделайте выбор:**Стартовая панель.**

Вы увидите различные виды анализа, которые доступны в модуле. Выберите анализ: **Таблицы и заголовки** и нажмите кнопку **ОК**.

На экране появится окно **Задайте таблицы.**

Шаг 2. Сначала в строке **Анализ** выберите Т**аблицы сопряженности** (возможен вариант **Таблицы флагов и заголовков)**.

Шаг 3. Далее нажмите кнопку **Задать таблицы**. В появившемся окне выберите переменные, которые будут табулированы в таблице. Эти переменные задают разбиение исходных данных на группы, поэтому часто их называют также группирующими переменными. В данном случае нужно табулировать значения переменных **СЕМ_ПОЛ**и **ТРЕВОГА**.

Поэтому выберите их, как это показано на рисунке ниже.

Заметьте, что вообще можно выбрать до 6 списков группирующих переменных, что позволяет построить чрезвычайно сложные таблицы, содержащие гораздо большее число переменных, чем в описываемом примере. Именно такие таблицы часто возникают при массовых обследованиях, и их нужно уметь строить.

После выбора переменных нажмите кнопку **ОК**. Вы вновь вернетесь в диалоговое окно, показанное на рисунке. Обратите внимание, что окно немного изменилось: около надписи Число таблиц появилась цифра 1, потому что вы выбрали переменные и попросили систему построить одну таблицу.

Шаг 4. Нажмите **Enter** на клавиатуре или кнопку **ОК** в верхнем правом углу диалогового окна.

Система произведет вычисления и предложит посмотреть результат в окне **Результаты кросстабуляции.**

Шаг 5. В окне **Результаты кросстабуляции** нажмите кнопку **Просмотреть итоговые таблицы**. На экране появится следующая таблица сопряженности:

Вы видите, что в этой таблице табулированы переменные **СЕМ_ПОЛ** и **ТРЕВОГА**. На пересечении строк и столбцов стоят абсолютные значения, вычисленные из исходного файла данных womenl .sta.

Мы табулировали совместно значения двух переменных, **СЕМ_ПОЛ** и **ТРЕВОГА**, и такое действие часто называется кросстабуляцией (от английского cross — пересекать).

Из построенной таблицы, называемой на сленге таблицей сопряженности, видно, что три женщины имеют полную семью и низкий уровень тревоги, две женщины имеют неполную семью и низкий уровень тревоги и т. д. Если вас интересует раздельная табуляция каждой переменной, посмотрите на крайний правый столбец и нижнюю строку таблицы. Вы увидите, что всего среди опрошенных женщин пять имели полную семью и пять — неполную семью; пять женщин имели высокий уровень тревожности (см. крайний правый столбец), пять — низкий уровень тревожности (см. нижнюю строку).

Часто возникает необходимость вместе с абсолютными значениями привести в таблице проценты. Система STATISTICA позволяет выбрать те проценты, которые

требуются: например, только проценты по строке, или проценты по столбцу, или проценты от общего количества, или же и те и другие.

Проценты по столбцу — это проценты, вычисленные относительно суммарного значения частот по столбцу. Проценты по строке — это проценты, вычисленные относительно суммарного значения частот по строке. Проценты от общего числа вычисляются относительно суммы частот в таблице. Рассмотрим, как это делается.

Шаг 6. Нажмите кнопку **Далее** в верхнем левом углу таблицы (см. рисунок).

Вы вновь вернетесь в окно **Результаты кросстабуляции.**

Шаг 7. В окне **Результаты кросстабуляции** обратите внимание на опции в правой части, объединенные в группу**Таблицы**.

Выберите, например, опцию **Проценты от общего числа**. Подведите курсор мыши к соответствующему квадрату и щелкните мышью. В окне **Результаты кросстабуляции** нажмите кнопку **Просмотреть итоговые таблицы**. На экране появится следующая таблица:

Здесь рядом с абсолютными значениями появились относительные величины — проценты, вычисленные от общего числа женщин, то есть от 10.

Итак, из таблицы видно (пожалуйста, проверьте!), что:

- 30% женщин имеют полную семью и низкий уровень тревоги (первая клетка таблицы),
- 20% женщин имеют полную семью и высокий уровень тревоги (вторая клетка таблицы),
- 20% женщин имеют неполную семью и низкий уровень тревоги,
- 30% женщин имеют неполную семью и высокий уровень тревоги.

Построенную таблицу можно отредактировать, изменить ее вид, надписи и т. д.

Шаг 8. Редактирование таблицы.

Дважды щелкните, например, по полю **Всего %** в построенной таблице. В появившемся окне **Имя строки таблицы результатов** вместо **Всего %** введите **%.**

Шаг 9. Построение отдельных таблиц с процентами.

Вернитесь вновь в окно **Результаты кросстабуляции** и обратите внимание на опцию **Отображать выбранные %** в отдельных таблицах.

Сделайте следующие установки: выберите опцию**Проценты от общего числа** и опцию **Отображать выбранные %** в отдельных таблицах. Затем нажмите кнопку **Просмотреть итоговые таблицы.**

Вы увидите две таблицы, одна из которых будет содержать только абсолютные значения, а другая — проценты, вычисленные от общего количества опрошенных.

Шаг 10. Создание автоотчета.

В системе STATISTICA имеется полезное средство подготовки отчета, которое позволяет представить все полученные результаты в формате rtf; далее отчет можно вывести на принтер, отредактировать и красиво распечатать.

Проделайте следующее: войдите в меню Вид и выберите опцию **Окно текста/ вывода**. Из построенных таблиц (они находятся в рабочем окне системы) выберите ту, которую нужно сохранить для отчета. Щелкните по ней мышью. Вновь войдите в меню **Файл** и выберите опцию **Печать**. Отмеченная таблица результатов будет распечатана.

В этом окне можно, например, отредактировать таблицу и подготовить ее в том формате, какой требуется для исследовательского отчета или статьи.

Обратите внимание, что в процессе работы ни разу не использовался какой-либо язык программирования, все действия носят интерактивный характер, и это большое достоинство системы STATISTICA. Работать в ней так же просто, как, например, в текстовом редакторе MS Word. В заключение вам предлагается упражнение, которое закрепит полученные навыки.

Пример. Создайте в STATISTICA файл women2.sta. Для градации значений переменных используются более реалистичные шкалы. Шкала семейного положения женщины: одинокая, неполная семья, полная семья. Шкала тревожности женщины: низкая, умеренная, высокая.

Графический анализ таблиц сопряженности

Таблицы сопряженности позволяют компактно описывать данные. Они удобны и требуют минимум комментариев, поэтому популярны среди врачей, социологов, маркетологов. В системе STATISTICA очень легко строятся даже самые сложные таблицы сопряженности.

Здесь мы рассмотрим, как визуализировать построенные таблицы, т. е. познакомимся со средствами STATISTICA, позволяющими графически проанализировать таблицы. Визуально гораздо проще увидеть закономерности, содержащиеся в таблицах. В примерах используются данные небольшого объема, чтобы можно было отчетливо представить основные приемы работы. Представьте, в каком сложном положении вы оказались, если бы имели дело с громадными таблицами, а именно такие таблицы возникают на практике. «Делайте вслед за нами!» — по-прежнему остается нашим главным девизом.

Итак, система STATISTICA запущена на компьютере, вы работаете в модуле Основные статистики и таблицы (в английской версии STATISTICA модуль Основные статистики и таблицы, называется Basic Statistics and Tables').

Пример (продолжение)

Файл данных women1.sta, с которым вы работаете, открыт в рабочем окне. Напомним, что в этом файле приведены результаты опроса 10 женщин (данные являются модельными) относительно их семейного положения и уровня тревожности.

Первая переменная **СЕМ_ПОЛ** — семейное положение женщин. Эта переменная принимает два значения: **П_семья** — полная семья, **Н_семья** — неполная семья.

Вторая переменная **ТРЕВОГА** — самооценка личностной тревожности женщины. Она принимает два значения: низкая, высокая. Известно, что личностная тревожность характеризуется устойчивой склонностью личности воспринимать жизненную ситуацию как угрожающую. В данном упрощенном примере мы использовали две степени тревожности: низкая и высокая.

Вы видите, что первая опрошенная женщина — наблюдение номер 1 (первая строка в таблице) — имеет полную семью и характеризует свое состояние как тревожное. Вторая опрошенная женщина — наблюдение номер 2 (вторая строка таблицы) — имеет неполную семью и оценивает уровень тревожности как низкий и т. д.

Шаг 1. Подведите курсор мыши к пункту Анализ. Щелкните по нему мышью. В появившемся меню сделайте выбор:**Стартовая панель**.

Выберите анализ: **Таблицы и заголовки** и нажмите кнопку **ОК**.

С помощью опций окна задания таблицы произведите табулировку переменных **СЕМ_ПОЛ** и **ТРЕВОГА**.

Шаг 2. После того как система построит таблицу, посмотрите внимательно на окно **Результаты кросстабуляции.**

Обратите внимание на кнопки в правом нижнем углу диалогового окна **Результаты кросстабуляции.**

Шаг 3. В диалоговом окне **Результаты кросстабуляции** нажмите кнопку **Категоризованные гистограммы:**

Смысл этих гистограмм следующий: опрошенные женщины разбиты на две группы (категории): женщины из полной семьи и женщины из неполной семьи. Обычная гистограмма для этих переменных выглядит следующим образом:

Здесь ясно видно, в чем состоит отличие категоризованных гистограмм от обычных. На обычной гистограмме количество женщин с высокой и низкой тревожностью одинаково. На категоризованной гистограмме количество женщин с

высоким уровнем тревожности в неполных семьях выше, чем в полных. Уровень тревожности женщин в полных семьях ниже, чем уровень тревожности в неполных семьях.

Продолжение примера

Рассмотрим файл данных women2.sta. Для градации значений переменных мы использовали более реалистичные шкалы: одинокая женщина, неполная семья, полная семья. Шкала тревожности женщины: низкая, умеренная, высокая.

Шаг 1. Подведите курсор мыши к пункту **Анализ**. Щелкните по нему мышью. В появившемся меню сделайте выбор:**Стартовая панель.**

Выберите **Таблицы и заголовки** и нажмите кнопку **ОК**.

Шаг 2. В строке **Анализ** выберите **Таблицы** сопряженности (возможен вариант **Таблицы флагов и заголовков)**.

Далее нажмите кнопку **Задать таблицы**. В появившемся окне выберите переменные, которые будут табулированы в таблице (подробности см. выше). В данном случае необходимо табулировать значения переменных **СЕМ_ПОЛ** и **ТРЕВОГА**.

Нажмите кнопку **Коды** и выберите коды (значения) табулируемых качественных признаков. В этом примере количество значений переменных увеличилось, т. к. используется более точная шкала измерения.

Если вы хотите, чтобы табулировались все значения переменных, нажмите кнопку **Выбрать все** в правом нижнем углу.

Заметьте, что вообще можно выбрать любой набор кодов. Коды переменных можно просмотреть, нажав кнопку**Инф**.

Шаг 3. Нажмите **Enter** на клавиатуре или кнопку **ОК** в верхнем правом углу • диалогового окна.

STATISTICA произведет вычисления, табулирует данные и предложит результат в окне **Результаты кросстабуляции.**

Шаг 4. В окне **Результаты кросстабуляции** нажмите кнопку **Просмотреть итоговые таблицы.**

Шаг 5. Нажмите кнопку **Далее** в верхнем углу таблицы, и вы вернетесь в окно результатов. В диалоговом окне **Результаты кросстабуляции** нажмите кнопку **Категоризованные гистограммы**.

Смысл гистограмм заключается в следующем: женщины разбиты на 3 группы или категории: женщины из полной семьи, женщины из неполной семьи, одинокие женщины (ср. с предыдущим примером). Для каждой группы построена отдельная гистограмма, и все эти гистограммы собраны вместе на одном графике, что позволяет визуально сравнить группы.

Шаг 6. В диалоговом окне **Результаты кросстабуляции** нажмите кнопку**3М гистограммы**.

Смысл этой гистограммы следующий: составляются всевозможные комбинации значений двух переменных: семейное положение и уровень тревожности, и подсчитывается, сколько раз встречалась каждая комбинация.

Трехмерная гистограмма очень наглядно воспроизводит таблицу кросстабуляции. Вы положили таблицу на плоскость и в каждую клетку поставили по столбцу, высота которого равна количеству наблюдений в клетке таблицы.

Если вас не устраивает ракурс построенной трехмерной гистограммы, можно его изменить, воспользовавшись средствами системы. STATISTICA предлагает удивительный инструмент работы с графиками. Например, их можно повернуть.

Нажмите кнопку **Вращение**, расположенную на панели инструментов.

Для вращения графика используйте линейку прокрутки. Немного поэкспериментируйте с ней. Сначала, например, с помощью мыши сдвиньте курсор прокрутки в крайне левое положение.

Каждый раз, когда сдвигается курсор, происходит поворот графика. Выберите тот вариант, который вас устраивает. Нажмите кнопку **ОК**. Нужный график появится на экране.

Шаг 7. Построение графиков взаимодействий частот. В окне **Результаты кросстабуляции** нажмите кнопку**Трафики взаимодействий частот**.

Смысл этого графика простой: он показывает, как взаимодействуют или как связаны между собой частоты наблюдений из разных групп.

Рекомендуемая литература:

1 Краткая экскурсия по системе STATISTICA http://hr-portal.ru/statistica/gl1/gl1.php

Лабораторная работа 1

Тема: Использование электронных таблиц Excel 2010 для вычисления выборочных характеристик больших данных

Цель: с помощью электронных таблиц Excel 2010 рассмотреть вычисления выборочных характеристик больших данных

Теоретические сведения

Математическая статистика подразделяется на две основные области: описательную и аналитическую статистику. Описательная статистика охватывает методы описания статистических данных, представления их в форме таблиц, распределений.

Аналитическая статистика или теория статистических выводов ориентирована на обработку данных, полученных в ходе эксперимента, с целью формулировки выводов, имеющих прикладное значение для самых различных областей человеческой деятельности.

1.1 Характеристика пакета Excel

Пакет Excel оснащен средствами статистической обработки данных. И хотя Excel существенно уступает специализированным статистическим пакетам обработки данных, тем не менее этот раздел математики представлен в Excel наиболее полно. В него включены основные, наиболее часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы, позволяющие проводить необходимый статистический анализ экономических, психологических, педагогических и медико-биологических типов данных.

Каждая единица информации занимает свою собственную ячейку (клетку) в создаваемой рабочей таблице. В каждой рабочей таблице 256 столбцов (из которых в новой рабочей таблице на экране видны, как правило, только первые 10 или 11 (от A до J или K) и 65 536 строк (из которых обычно видны только первые 15-20). Каждая новая рабочая книга содержит три чистых листа рабочих таблиц.

Вся помещаемая в электронную таблицу информация хранится в отдельных клетках рабочей таблицы. Но ввести информацию можно только в текущую клетку. С помощью адреса в строке формул и табличного курсора Excel указывает, какая из клеток

рабочей таблицы является текущей. В основе системы адресации клеток рабочей таблицы лежит комбинация буквы (или букв) столбца и номера строки, например A2, B12.

При рассмотрении применения методов обработки статистических данных в данной лабораторной работе ограничимся только простейшими и наиболее часто описательными статистиками, реализованными в мастере функций Excel.

1.2 Использование специальных функций

В мастере функций Excel имеется ряд специальных функций, предназначенных для вычисления выборочных характеристик.

Функция СРЗНАЧ вычисляет среднее арифметическое из нескольких массивов (аргументов) чисел. Аргументы число1, число2, ... — это от 1 до 30 массивов для которых вычисляется среднее.

Функция МЕДИАНА позволяет получать медиану заданной выборки. Медиана - это элемент выборки, число элементов выборки со значениями больше которого и меньше которого равно.

Функция МОДА вычисляет наиболее часто встречающееся значение в выборке.

Функция ДИСП позволяет оценить дисперсию по выборочным данным.

Функция СТАНДОТКЛОН вычисляет стандартное отклонение.

Функция ЭКСЦЕСС вычисляет оценку эксцесса по выборочным данным.

Функция СКОС позволяет оценить асимметрию выборочного распределения.

Функция КВАРТИЛЬ вычисляет квартили распределения. Функция имеет формат КВАРТИЛЬ(массив, значение), где массив – интервал ячеек, содержащих значения СВ; значение определяет какая квартиль должна быть найдена (0 – минимальное значение, 1 – нижняя квартиль, 2 – медиана, 3 – верхняя квартиль, 4 – максимальное значение распределения).

Таблица 5 - Провести статистический анализ методом описательной статистики доходов населения в регионе 1 и регионе 2.

1	49	
1	51	
1	49	
1	51	
1	49	
1	51	
1	49	
1	51	
1	49	
491	51	
500	500	сумма
50	50	среднее
24010	1,11	дисперсия
154,95	1,05	станд. отклонение
1	49	квартили
1	51	квартили
1	50	медиана
1	49	мода
10	-2,57	эксцесс
3,16	0	скос(ассиметрия)

Задания для самостоятельной работы

1. Наблюдение посещаемости четырех внеклассных мероприятий в экспериментальном (20 человек) и контрольном (30 человек) классах дали значения (соответственно): 18, 20, 20, 18 и 15, 23, 10, 28. Требуется найти среднее значение, стандартное отклонение, медиану и квартили этих данных.

2. Найти среднее значение, медиану, стандартное отклонение и квартили результатов бега на дистанцию 100 м у группы студентов (с): 12,8; 13,2; 13,0; 12,9; 13,5; 13,1.

3. Определите верхнюю и нижнюю квартиль, выборочную асимметрию и эксцесс для данных измерений роста групп студенток: 164, 160, 157, 166, 162, 160, 161, 159, 160, 163, 170, 171.

4. Найти наиболее популярный туристический маршрут из четырех реализуемых фирмой, если за неделю последовательно были реализованы следующие маршруты: 1, 3, 3, 2, 1, 1, 4, 4, 2, 4, 1, 3, 2, 4, 1, 4, 4, 3, 1, 2, 3, 4, 1, 1, 3.

1.3 Использование инструмента Пакет анализа

В пакете Excel помимо мастера функций имеется набор более мощных инструментов для работы с несколькими выборками и углубленного анализа данных, называемый Пакет анализа, который может быть использован для решения задач статистической обработки выборочных данных.

Для установки пакета Анализ данных в Excel сделайте следующее:

- в меню Сервис выберите команду Надстройки;
- в появившемся списке установите флажок Пакет анализа.

Для использования статистического пакета анализа данных необходимо:

- указать курсором мыши на пункт меню Сервис и щелкнуть левой кнопкой мыши;
- в раскрывающемся списке выбрать команду Анализ данных (если команда Анализ данных отсутствует в меню Сервис, то необходимо установить в Excel пакет анализа данных);
- выбрать строку Описательная статистика и нажать кнопку Ok
- в появившемся диалоговом окне указать входной интервал, то есть ввести ссылки на ячейки, содержащие анализируемые данные;
- указать выходной интервал, то есть ввести ссылку на ячейку, в которую будут выведены результаты анализа;
- в разделе Группирование переключатель установить в положение по столбцам или по строкам;
- установить флажок в поле Итоговая статистика и нажать Ок.

Задание для самостоятельной работы

1. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг./м3): 12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16. Необходимо определить основные выборочные характеристики.

Лабораторная работа 2

Тема: Использование электронных таблиц Excel 2010 для построения распределений случайных величин и генерации случайных чисел

Цель: с помощью электронных таблиц Excel 2010 рассмотреть построения распределений случайных величин и генерации случайных чисел

Теоретические сведения

Распределение вероятностей – одно из центральных понятий теории вероятности и математической статистики. Определение распределения вероятности равносильно заданию вероятностей всех СВ, описывающих некоторое случайное событие. Распределение вероятностей некоторой СВ, возможные значения которой x1, x2, … xn образуют выборку, задается указанием этих значений и соответствующих им вероятностей p1, p2,… pn. (pn должны быть положительны и в сумме давать единицу).

В данной лабораторной работе будут рассмотрены и построены с помощью MS Excel наиболее распространенные распределения вероятности: биномиальное и нормальное.

2.1 Биномиальное распределение

Представляет собой распределение вероятностей числа наступлений некоторого события («удачи») в n повторных независимых испытаниях, если при каждом испытании вероятность наступления этого события равна p. При этом распределении разброс вариант (есть или нет события) является следствием влияния ряда независимых и случайных факторов.

Примером практического использования биномиального распределения может являться контроль качества партии фармакологического препарата. Здесь требуется подсчитать число изделий (упаковок), не соответствующих требованиям. Все причины, влияющие на качество препарата, принимаются одинаково вероятными и не зависящими друг от друга. Сплошная проверка качества в этой ситуации не возможна, поскольку изделие, прошедшее испытание, не подлежит дальнейшему использованию. Поэтому для контроля из партии наудачу выбирают определенное количество образцов изделий (n). Эти образцы всестороннее проверяют и регистрируют число бракованных изделий (k). Теоретически число бракованных изделий может быть любым, от 0 до n.

В Excel функция БИНОМРАСП применяется для вычисления вероятности в задачах с фиксированным числом тестов или испытаний, когда результатом любого испытания может быть только успех или неудача.

Функция использует следующие параметры:

БИНОМРАСП (число_успехов; число_испытаний; вероятностъ_успеха; интегральная), где

число_успехов — это количество успешных испытаний;

число_испытаний — это число независимых испытаний (число успехов и число испытаний должны быть целыми числами);

вероятность_ успеха — это вероятность успеха каждого испытания;

интегральный — это логическое значение, определяющее форму функции.

Если данный параметр имеет значение ИСТИНА (=1), то считается интегральная функция распределения (вероятность того, что число успешных испытаний не менее значения число_ успехов);

если этот параметр имеет значение ЛОЖЬ (=0), то вычисляется значение функции плотности распределения (вероятность того, что число успешных испытаний в точности равно значению аргумента число_ успехов).

Пример 1. Какова вероятность того, что трое из четырех новорожденных будут мальчиками?

Решение:

1. Устанавливаем табличный курсор в свободную ячейку, например в A1. Здесь должно оказаться значение искомой вероятности.

2. Для получения значения вероятности воспользуемся специальной функцией: нажимаем на панели инструментов кнопку Вставка функции (fx).

3. В появившемся диалоговом окне Мастер функций - шаг 1 из 2 слева в поле Категория указаны виды функций. Выбираем Статистическая. Справа в поле Функция выбираем функцию БИНОМРАСП и нажимаем на кнопку ОК.

Появляется диалоговое окно функции. В поле Число_s вводим с клавиатуры количество успешных испытаний (3). В поле Испытания вводим с клавиатуры общее количество испытаний (4). В рабочее поле Вероятность_s вводим с клавиатуры вероятность успеха в отдельном испытании (0,5). В поле Интегральный вводим с клавиатуры вид функции распределения — интегральная или весовая (0). Нажимаем на кнопку ОК.

В ячейке А1 появляется искомое значение вероятности p = 0,25. Ровно 3 мальчика из 4 новорожденных могут появиться с вероятностью 0,25.

Если изменить формулировку условия задачи и выяснить вероятность того, что появится не более трех мальчиков, то в этом случае в рабочее поле Интегральный вводим 1 (вид функции распределения интегральный). Вероятность этого события будет равна 0,9375.

Задания для самостоятельной работы

1. Какова вероятность того, что восемь из десяти студентов, сдающих зачет, получат «незачет». (0,04)

2.2 Нормальное распределение

Нормальное распределение - это совокупность объектов, в которой крайние значения некоторого признака — наименьшее и наибольшее — появляются редко; чем ближе значение признака к математическому ожиданию, тем чаще оно встречается. Например, распределение студентов по их весу приближается к нормальному распределению. Это распределение имеет очень широкий круг приложений в статистике, включая проверку гипотез.

Диаграмма нормального распределения симметрична относительно точки a (математического ожидания). Медиана нормального распределения равна тоже a. При этом в точке a функция f(x) достигает своего максимума, который равен $\frac{1}{\sigma\sqrt{2\pi}}$.

В Excel для вычисления значений нормального распределения используются функция НОРМРАСП, которая вычисляет значения вероятности нормальной функции распределения для указанного среднего и стандартного отклонения.

Функция имеет параметры:

НОРМРАСП (x; среднее; стандартное_откл; интегральная), где:

x — значения выборки, для которых строится распределение;

среднее — среднее арифметическое выборки;

стандартное_откл — стандартное отклонение распределения;

интегральный — логическое значение, определяющее форму функции. Если интегральная имеет значение ИСТИНА(1), то функция НОРМРАСП возвращает интегральную функцию распределения; если это аргумент имеет значение ЛОЖЬ (0), то вычисляет значение функция плотности распределения.

Если среднее = 0 и стандартное_откл = 1, то функция НОРМРАСП возвращает стандартное нормальное распределение.

Пример 2. Построить график нормальной функции распределения f(x) при x, меняющемся от 19,8 до 28,8 с шагом 0,5, a=24,3 и σ=1,5.

Решение

1. В ячейку A1 вводим символ случайной величины x, а в ячейку B1 — символ функции плотности вероятности — f(x).

2. Вводим в диапазон A2:A21 значения x от 19,8 до 28,8 с шагом 0,5. Для этого воспользуемся маркером автозаполнения: в ячейку A2 вводим левую границу диапазона (19,8), в ячейку A3 левую границу плюс шаг (20,3). Выделяем блок A2:A3. Затем за правый нижний угол протягиваем мышью до ячейки A21 (при нажатой левой кнопке мыши).

3. Устанавливаем табличный курсор в ячейку B2 и для получения значения вероятности воспользуемся специальной функцией — нажимаем на панели инструментов кнопку Вставка функции (fx). В появившемся диалоговом окне Мастер функций - шаг 1 из 2 слева в поле Категория указаны виды функций. Выбираем Статистическая. Справа в поле Функция выбираем функцию НОРМРАСП. Нажимаем на кнопку ОК.

4. Появляется диалоговое окно НОРМРАСП. В рабочее поле X вводим адрес ячейки A2 щелчком мыши на этой ячейке. В рабочее поле Среднее вводим с клавиатуры значение математического ожидания (24,3). В рабочее поле Стандартное_откл вводим с клавиатуры значение среднеквадратического отклонения (1,5). В рабочее поле Интегральная вводим с клавиатуры вид функции распределения (0). Нажимаем на кнопку ОК.

5. В ячейке B2 появляется вероятность p = 0,002955. Указателем мыши за правый нижний угол табличного курсора протягиванием (при нажатой левой кнопке мыши) из ячейки B2 до B21 копируем функцию НОРМРАСП в диапазон B3:B21.

6. По полученным данным строим искомую диаграмму нормальной функции распределения. Щелчком указателя мыши на кнопке на панели инструментов вызываем Мастер диаграмм. В появившемся диалоговом окне выбираем тип диаграммы График, вид — левый верхний. После нажатия кнопки Далее указываем диапазон данных — B1:B21 (с помощью мыши). Проверяем, положение переключателя Ряды в: столбцах. Выбираем закладку Ряд и с помощью мыши вводим диапазон подписей оси X: A2:A21. Нажав на кнопку Далее, вводим названия осей X и У и нажимаем на кнопку Готово.

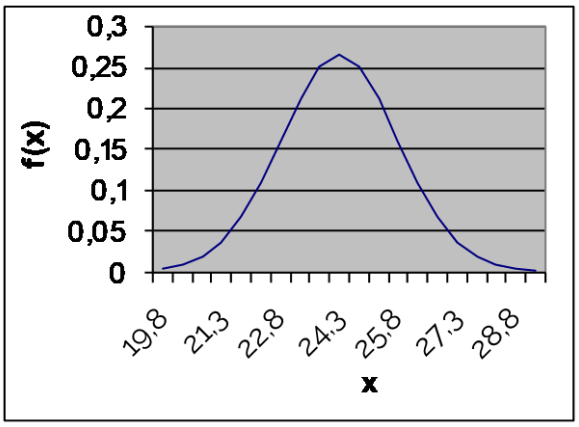


Рисунок 19 - График нормальной функции распределения

Получен приближенный график нормальной функции плотности распределения (рисунок 1).

Задания для самостоятельной работы

1. Построить график нормальной функции плотности распределения f(x) при x, меняющемся от 20 до 40 с шагом 1 при σ=3.

2.3 Генерация случайных величин

Еще одним аспектом использования законов распределения вероятностей является генерация случайных величин. Бывают ситуации, когда необходимо получить последовательность случайных чисел. Это, в частности, требуется для моделирования объектов, имеющих случайную природу, по известному распределению вероятностей.

Процедура генерации случайных величин используется для заполнения диапазона ячеек случайными числами, извлеченными из одного или нескольких распределений.

В MS Excel для генерации СВ используются функции из категории Математические:

СЛЧИС () – выводит на экран равномерно распределенные случайные числа больше или равные 0 и меньшие 1;

СЛУЧМЕЖДУ (ниж_граница; верх_граница) – выводит на экран случайное число, лежащее между произвольными заданными значениями.

В случае использования процедуры Генерация случайных чисел из пакета Анализа необходимо заполнить следующие поля:

- число переменных вводится число столбцов значений, которые необходимо разместить в выходном диапазоне. Если это число не введено, то все столбцы в выходном диапазоне будут заполнены;

- число случайных чисел вводится число случайных значений, которое необходимо вывести для каждой переменной, если число случайных чисел не будет введено, то все строки выходного диапазона будут заполнены;

- в поле распределение необходимо выбрать тип распределения, которое следует использовать для генерации случайных переменных:

1. равномерное - характеризуется верхней и нижней границами. Переменные извлекаются с одной и той же вероятностью для всех значений интервала.

2. нормальное — характеризуется средним значением и стандартным отклонением. Обычно для этого распределения используют среднее значение 0 и стандартное отклонение 1.

3. биномиальное — характеризуется вероятностью успеха (величина p) для некоторого числа попыток. Например, можно сгенерировать случайные двухальтернативные переменные по числу попыток, сумма которых будет биномиальной случайной переменной;

4. дискретное — характеризуется значением СВ и соответствующим ему интервалом вероятности, диапазон должен состоять из двух столбцов: левого, содержащего значения, и правого, содержащего вероятности, связанные со значением в данной строке. Сумма вероятностей должна быть равна 1;

5. распределения Бернулли, Пуассона и Модельное.

- в поле случайное рассеивание вводится произвольное значение, для которого необходимо генерировать случайные числа. Впоследствии можно снова использовать это значение для получения тех же самых случайных чисел.

- выходной диапазон вводится ссылка на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные.

Рассмотрим пример.

Пример 3. Повар столовой может готовить 4 различных первых блюда (уха, щи, борщ, грибной суп). Необходимо составить меню на месяц, так чтобы первые блюда чередовались в случайном порядке.

Решение

Пронумеруем первые блюда по порядку: 1 — уха, 2 — щи, 3 — борщ, 4 — грибной суп. Введем числа 1-4 в диапазон A2:A5 рабочей таблицы.

Укажем желаемую вероятность появления каждого первого блюда. Пусть все блюда будут равновероятны (p=1/4). Вводим число 0,25 в диапазон B2:B5.

В меню Сервис выбираем пункт Анализ данных и далее указываем строку Генерация случайных чисел. В появившемся диалоговом окне указываем Число переменных — 1, Число случайных чисел — 30 (количество дней в месяце). В поле Распределение указываем Дискретное (только натуральные числа). В поле Входной интервал значений и вероятностей вводим (мышью) диапазон, содержащий номера супов и их вероятности. – A2:B5.

Указываем выходной диапазон и нажимаем ОК. В столбце С появляются случайные числа: 1, 2, 3, 4.

Задание для самостоятельной работы:

Сформировать выборку из 10 случайных чисел, лежащих в диапазоне от 0 до 1.

Сформировать выборку из 20 случайных чисел, лежащих в диапазоне от 5 до 20.

Пусть спортсмену необходимо составить график тренировок на 10 дней, так чтобы дистанция, пробегаемая каждый день, случайным образом менялась от 5 до 10 км.

Составить расписание внеклассных мероприятий на неделю для случайного проведения: семинаров, интеллектуальных игр, КВН и спец. курса.

Составить расписание на месяц для случайной демонстрации на телевидении одного из четырех рекламных роликов турфирмы. Причем вероятность появления рекламного ролика №1 должна быть в два раза выше, чем остальных рекламных роликов.

Лабораторная работа 3

Тема: Использование электронных таблиц Excel 2010 для построения выборочных функций распределения

Цель: с помощью электронных таблиц Excel 2010рассмотреть построения выборочных функций распределения

Теоретические сведения

Рассмотренные в лабораторной работе 2 распределения вероятностей СВ опираются на знание закона распределения СВ. Для практических задач такое знание – редкость. Здесь закон распределения обычно неизвестен, или известен с точностью до некоторых неизвестных параметров. В частности, невозможно рассчитать точное значение соответствующих вероятностей, так как нельзя определить количество общих и благоприятных исходов. Поэтому вводится статистическое определение вероятности. По этому определению вероятность равна отношению числа испытаний, в которых событие произошло, к общему числу произведенных испытаний. Такая вероятность называется статистической частотой.

Связь между эмпирической функцией распределения и функцией распределения (теоретической функцией распределения) такая же, как связь между частотой события и его вероятностью.

Для построения выборочной функции распределения весь диапазон изменения случайной величины X (выборки) разбивают на ряд интервалов (карманов) одинаковой ширины. Число интервалов обычно выбирают не менее 3 и не более 15. Затем определяют число значений случайной величины X, попавших в каждый интервал (абсолютная частота, частота интервалов).

Частота интервалов – число, показывающее сколько раз значения, относящиеся к каждому интервалу группировки, встречаются в выборке. Поделив эти числа на общее количество наблюдений (n), находят относительную частоту (частость) попадания случайной величины X в заданные интервалы.

По найденным относительным частотам строят гистограммы выборочных функций распределения. Гистограмма распределения частот – это графическое представление выборки, где по оси абсцисс (OX) отложены величины интервалов, а по оси ординат (OУ) – величины частот, попадающих в данный классовый интервал. При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график плотности распределения.

Накопленная частота интервалов – это число, полученное последовательным суммированием частот в направлении от первого интервала к последнему, до того интервала включительно, для которого определяется накопленная частота.

В Excel для построения выборочных функций распределения используются специальная функция ЧАСТОТА и процедура Гистограмма из пакета анализа.

Функция ЧАСТОТА (массив_данных, двоичный_массив) вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив цифр, где

- массив_данных — это массив или ссылка на множество данных, для которых вычисляются частоты;

- двоичный_массив — это массив интервалов, по которым группируются значения выборки.

Процедура Гистограмма из Пакета анализа выводит результаты выборочного распределения в виде таблицы и графика. Параметры диалогового окна Гистограмма:

- Входной диапазон - диапазон исследуемых данных (выборка);

- Интервал карманов - диапазон ячеек или набор граничных значений, определяющих выбранные интервалы (карманы). Эти значения должны быть введены в возрастающем порядке. Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

- выходной диапазон предназначен для ввода ссылки на левую верхнюю ячейку выходного диапазона.

- переключатель Интегральный процент позволяет установить режим включения в гистограмму графика интегральных процентов.

- переключатель Вывод графика позволяет установить режим автоматического создания встроенной диаграммы на листе, содержащем выходной диапазон.

Пример 1. Построить эмпирическое распределение веса студентов в килограммах для следующей выборки: 64, 57, 63, 62, 58, 61, 63, 70, 60, 61, 65, 62, 62, 40, 64, 61, 59, 59, 63, 61.

Решение

В ячейку A1 введите слово Наблюдения, а в диапазон A2:A21 — значения веса студентов (см. рис. 1).

В ячейку B1 введите названия интервалов Вес, кг. В диапазон B2:B8 введите граничные значения интервалов (40, 45, 50, 55, 60, 65, 70).

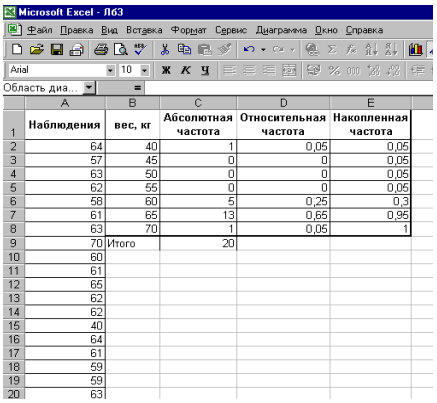
Введите заголовки создаваемой таблицы: в ячейки C1 — Абсолютные частоты, в ячейки D1 — Относительные частоты, в ячейки E1 — Накопленные частоты.(см. рис. 1).

С помощью функции Частота заполните столбец абсолютных частот, для этого выделите блок ячеек C2:C8. С панели инструментов Стандартная вызовите Мастер функций (кнопка fx). В появившемся диалоговом окне выберите категорию Статистические и функцию ЧАСТОТА, после чего нажмите кнопку ОК. Указателем мыши в рабочее поле Массив_данных введите диапазон данных наблюдений (A2:A8). В рабочее поле Двоичный_массив мышью введите диапазон интервалов (B2:B8). Слева на клавиатуре последовательно нажмите комбинацию клавиш Ctrl+Shift+Enter. В столбце C должен появиться массив абсолютных частот (рисунок 20).

В ячейке C9 найдите общее количество наблюдений. Активизируйте ячейку C9, на панели инструментов Стандартная нажмите кнопку Автосумма. Убедитесь, что диапазон суммирования указан правильно и нажмите клавишу Enter.

Заполните столбец относительных частот. В ячейку введите формулу для вычисления относительной частоты: =C2/C9. Нажмите клавишу Enter. Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон и получите массив относительных частот.

Заполните столбец накопленных частот. В ячейку D2 скопируйте значение относительной частоты из ячейки E2. В ячейку D3 введите формулу: =E2+D3. Нажмите клавишу Enter. Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон D3:D8. Получим массив накопленных частот.



	A	B	C	D	E
1	Наблюдения	вес, кг	Абсолютная частота	Относительная частота	Накопленная частота
2	64	40	1	0,05	0,05
3	57	45	0	0	0,05
4	63	50	0	0	0,05
5	62	55	0	0	0,05
6	58	60	5	0,25	0,3
7	61	65	13	0,65	0,95
8	63	70	1	0,05	1
9	70	Итого	20		
10	60				
11	61				
12	65				
13	62				
14	62				
15	40				
16	64				
17	61				
18	59				
19	59				
20	63				

Рисунок 20 - Результат вычислений из примера 1

Постройте диаграмму относительных и накопленных частот. Щелчком указателя мыши по кнопке на панели инструментов вызовите Мастер диаграмм. В появившемся диалоговом окне выберите закладку Нестандартные и тип диаграммы График/гистограмма. После редактирования диаграмма будет иметь такой вид, как на рисунок 21.

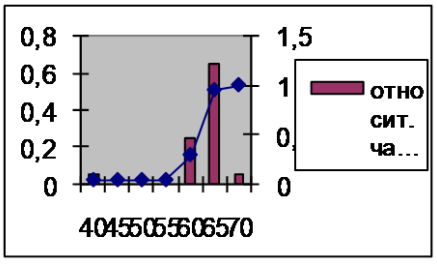


Рисунок 21 - Диаграмма относительных и накопленных частот из примера

Задания для самостоятельной работы:

1. Для данных из примера 1 построить выборочные функции распределения, воспользовавшись процедурой Гистограмма из пакета Анализа.

2. Построить выборочные функции распределения (относительные и накопленные частоты) для роста в см. 20 студентов: 181, 169, 178, 178, 171, 179, 172, 181, 179, 168, 174, 167, 169, 171, 179, 181, 181, 183, 172, 176.

3. Найдите распределение по абсолютным частотам для следующих результатов тестирования в баллах: 79, 85, 78, 85, 83, 81, 95, 88, 97, 85 (используйте границы интервалов 70, 80, 90).

4. Рассмотрим любой из критериев оценки качеств педагога-профессионала, например, «успешное решение задач обучения и воспитания». Ответ на этот вопрос анкеты типа «да», «нет» достаточно груб. Чтобы уменьшить относительную ошибку такого измерения, необходимо увеличить число возможных ответов на конкретный критериальный вопрос. В табл. 1 представлены возможные варианты ответов.

Обозначим этот параметр через x. Тогда в процессе ответа на вопрос величина x примет дискретное значение x, принадлежащее определенному интервалу значений. Поставим в соответствие каждому из ответов определенное числовое значение параметра x (таблица 6).

Таблица 6 - Критериальный вопрос: успешное решение задач обучения и воспитания

№ п/п	Варианты ответов	X
1	Абсолютно неуспешно	0,1
2	Неуспешно	0,2
3	Успешно в очень малой степени	0,3
4	В определенной степени успешно, но еще много недостатков	0,4
5	В среднем успешно, но недостатки имеются	0,5
6	Успешно с некоторыми оговорками	0,6
7	Успешно, но хотелось бы улучшить результат	0,7
8	Достаточно успешно	0,8
9	Очень успешно	0,9
10	Абсолютно успешно	1

При проведении анкетирования в каждой отдельной анкете параметр x принимает случайное значение, но только в пределах числового интервала от 0,1 до 1.

Тогда в результате измерений мы получаем неранжированный ряд случайных значений (таблица 7).

Таблица 7 - Результаты опроса ста учителей

0,6	0,7	1	0,6	0,2	0,8	0,3	0,5	0,9	0,3
0,5	0,1	0,4	0,5	0,5	0,4	0,4	0,6	0,5	0,4
0,6	0,9	0,7	0,9	0,8	0,5	0,5	0,6	0,8	0,4
0,4	0,4	0,8	0,7	0,6	0,6	0,7	0,8	0,5	0,6
0,7	0,6	0,7	0,3	0,2	0,7	0,5	0,3	0,4	0,5
0,9	0,7	0,6	0,5	0,7	0,6	0,2	0,8	0,8	0,3
0,7	0,5	0,7	0,6	0,2	0,5	0,8	0,3	0,7	0,8
0,7	0,6	0,6	0,8	0,4	0,6	0,6	0,6	0,9	0,7
0,7	0,5	0,7	0,6	0,9	0,4	0,8	0,7	0,5	0,8
0,8	0,9	0,4	0,3	0,4	0,6	0,4	0,5	0,3	0,5

Сгруппируйте полученную выборку, рассчитайте среднее значение выборки, стандартное отклонение, абсолютную и относительную частоту появления параметра, а также постройте график плотности вероятности $f(x)=\frac{W(x)}{\sigma\sqrt{2\pi}}$, где

W(x) – относительная частота наступления события;

σ - стандартное отклонение;

π =3,14.

Постройте график функции f(x) и сравните его с нормальным распределением Гаусса.

Лабораторная работа 4

Тема: Использование электронных таблиц Excel 2010 для обработки данных

Цель: с помощью электронных таблиц Excel 2010 рассмотреть обработку данных

Теоретические сведения

В процессе обучения постоянно ощущается потребность в хорошо разработанных методах измерения уровня обученности в самых различных областях знаний. Известно, что профессиональное тестирование было начато еще в 2200 году до нашей эры, когда служащие Китайского императора тестировались, чтобы определить их пригодность для императорской службы. По некоторым оценкам в 1986 году по крайней мере 800 профессий лицензировались в Соединенных Штатах на основании тестирования (А.А. Захаров, А.В. Колпаков Современные математические методы объективных педагогических измерений)

Почти каждый педагог разрабатывает тестовые задания по своей дисциплине, но не каждый может грамотно обработать и интерпретировать результаты теста. Напротив, грамотное конструирование теста на основе знания теории тестирования позволит педагогу-исследователю создать инструмент, позволяющий провести объективное измерение знаний, умений и навыков по данному курсу с необходимой точностью.

В настоящее время существуют два теоретических подхода к созданию тестов: классическая теория и современная теория IRT (Item Response Theory). Оба подхода базируются на последующей статистической обработке так называемого сырого балла (raw score), то есть балла, набранного в результате тестирования. Только после проведения многократных статистических обработок можно говорить о создании теста с устойчивыми параметрами качества (надежностью и валидностью).

Для обработки данных, полученных на этапе тестирования, воспользуемся пакетом MS Office 2010 и электронными таблицами MS Excel.

После сбора эмпирических данных необходимо провести статистическую обработку, которую будем проводить на ЭВМ. Этап математико–статистической обработки разобьем на ряд шагов.

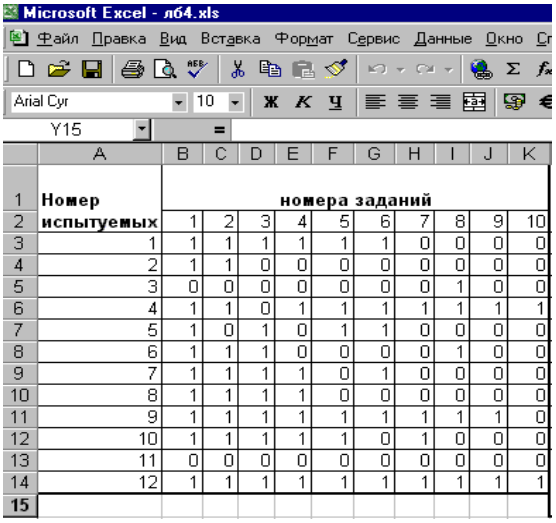
Шаг 1. Формирование матрицы тестовых результатов.

Результаты ответов учеников на задания тестов оцениваются в дихотомической шкале: за каждый правильный ответ учащийся получает один балл, а за неправильный ответ или за пропуск задания – нуль баллов (рисунок 23).

Шаг 2. Преобразование матрицы тестовых результатов.

На втором шаге из матрицы тестовых результатов устраняются строки и столбцы, состоящие только из нулей или только из единиц. В приведенном выше примере таких

столбцов нет, а строк только две. Одна из них, нулевая строка соответствует ответам одиннадцатого испытуемого, который не смог выполнить правильно ни одного задания в тесте.



	A	B	C	D	E	F	G	H	I	J	K
1	Номер	номера заданий									
2	испытуемых	1	2	3	4	5	6	7	8	9	10
3	1	1	1	1	1	1	1	0	0	0	0
4	2	1	1	0	0	0	0	0	0	0	0
5	3	0	0	0	0	0	0	0	1	0	0
6	4	1	1	0	1	1	1	1	1	1	1
7	5	1	0	1	0	1	1	0	0	0	0
8	6	1	1	1	0	0	0	0	1	0	0
9	7	1	1	1	1	0	1	0	0	0	0
10	8	1	1	1	1	0	0	0	0	0	0
11	9	1	1	1	1	1	1	1	1	1	0
12	10	1	1	1	1	1	0	1	0	0	0
13	11	0	0	0	0	0	0	0	0	0	0
14	12	1	1	1	1	1	1	1	1	1	1
15											

Рисунок 23 - Матрица результатов тестирования

В этом случае вывод довольно однозначен: тест непригоден для оценки знаний такого ученика. Для выявления его уровня знаний тест необходимо облегчить, добавив несколько более легких заданий, которые, скорее всего, выполнит правильно большинство остальных испытуемых группы.

Столь же непригоден, но уже по другой причине, тест для оценки знаний двенадцатого ученика, который выполнил правильно все без исключения задания теста. Причина непригодности теста заключается в его излишней легкости, не позволяющий выявить истинный уровень подготовки двенадцатого ученика. Возможно, двенадцатый ученик знает много чего другого и в состоянии выполнить по контролируемым разделам содержания гораздо более трудные задания, которые просто не были включены в тест.

Таким образом, на данном шаге необходимо удалить из матрицы данных 11 и 12 строки.

Шаг 3. Подсчет индивидуальных баллов испытуемых и количество правильных ответов на каждое задание теста.

Индивидуальный балл испытуемого получается суммированием всех единиц, полученных им за правильное выполнение задания теста. В Excel для суммирования данных по строке можно воспользоваться кнопкой Автосумма Σ на панели инструментов Стандартная. Для удобства полученные индивидуальные баллы (Xi) приводятся в последнем столбце матрицы результатов (рисунок 24).

Число правильных ответов на задания теста (Yi) также получается суммированием единиц, но уже расположенным по столбцам .(рис. 2)

Шаг 4. Упорядочение матрицы результатов.

Значения индивидуальных баллов необходимо отсортировать по возрастанию, для этого в MS Excel:

- выделим блок ячеек, содержащих номера испытуемых, матрицу результатов и индивидуальные баллы. Начинать выделение необходимо со столбца X (индивидуальные баллы).

- на панели инструментов Стандартная нажимаем на кнопку Сортировка по возрастанию . Матрица результатов примет вид, изображенный на рисунке 25.

M	N	O	P	Q	R	S	T	U	V	W	X
Номер испытуемых	номера заданий										Индивидуальные баллы (X)
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	1	0	0	0	0	6
2	1	1	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	1	0	0	1
4	1	1	0	1	1	1	1	1	1	1	9
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	0	1	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
8	1	1	1	1	0	0	0	0	0	0	4
9	1	1	1	1	1	1	1	1	1	0	9
10	1	1	1	1	1	0	1	0	0	0	6
число правильных ответов (Y)	9	8	7	6	5	5	3	4	2	1	50

Рисунок 24 -
Матрица с подсчетом итоговых сумм

M	N	O	P	Q	R	S	T	U	V	W	X
Номер испытуемых	номера заданий										Индивидуальные баллы (X)
	1	2	3	4	5	6	7	8	9	10	
3	0	0	0	0	0	0	0	1	0	0	1
2	1	1	0	0	0	0	0	0	0	0	2
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	0	1	0	0	4
8	1	1	1	1	0	0	0	0	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
1	1	1	1	1	1	1	0	0	0	0	6
10	1	1	1	1	1	0	1	0	0	0	6
4	1	1	0	1	1	1	1	1	1	1	9
9	1	1	1	1	1	1	1	1	1	0	9
число правильных ответов (Y)	9	8	7	6	5	5	3	4	2	1	50

Рисунок 25 -
Упорядоченная матрица результатов

Шаг 5. Графическое представление данных.

Эмпирические результаты тестирования можно представить в виде полигона частот, гистограммы, сглаженной кривой или графика.

Для построения кривых упорядочим результаты эксперимента и подсчитаем частоту получения баллов (рисунок 26 - 28).

	A	B
	Номер	Балл
2	1	6
3	2	2
4	3	1
5	4	9
6	5	4
7	6	4
8	7	5
9	8	4
10	9	9
11	10	6

Рисунок 26 -
Несгруппированный ряд

	A	B	C
1	Номер	Балл	Ранг
2	3	1	1
3	2	2	2
4	5	4	3
5	6	4	3
6	8	4	3
7	7	5	6
8	1	6	7
9	10	6	7
10	4	9	9
11	9	9	9
12			

Рисунок 27 -
Ранжированный ряд

Балл	Частота
1	1
2	1
4	3
5	1
6	2
9	2

Рисунок 28 -
Частотное распределение

Для расчета рейтинга (ранга) каждого учащегося по индивидуальным балам необходимо применить функцию РАНГ, которая возвращает ранг числа в списке чисел. Ранг числа – это его величина относительно других значений в списке.

В MS Excel 2000 для вычисления ранга используется функция

РАНГ (число; ссылка; порядок), где

Число – адрес на ячейку, для которой определяется ранг.

Ссылка - ссылка на массив индивидуальных баллов (выборка).

Порядок – число, определяющее способ упорядочения. Если порядок равен 0 (нулю), или опущен, то Excel определяет ранг числа так, как если бы ссылка была списком, отсортированным в порядке убывания. Если порядок – любое ненулевое число, то Excel определяет ранг числа так, как если бы ссылка была списком, отсортированным в порядке возрастания.

Примечание. Функция РАНГ присваивает повторяющимся числам одинаковый ранг. При этом наличие повторяющихся чисел влияет на ранг последующих чисел. Например, если в списке целых чисел дважды встречается число 10, имеющее ранг 5, число 11 будет иметь ранг 7 (ни одно из чисел не будет иметь ранг 6).

По частотному распределению можно построить гистограмму (рисунок 29).

Гистограмму можно построить и по индивидуальным баллам (рисунок 30).

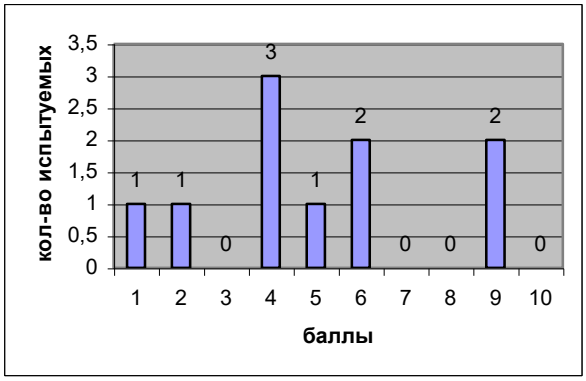


Рисунок 29 -
Столбиковая гистограмма

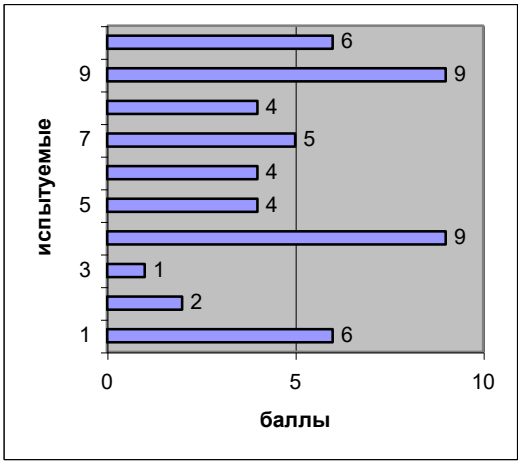


Рисунок 30
Гистограмма распределения инд. баллов

При разработке тестов необходимо помнить о том, что кривая распределения индивидуальных баллов, получаемых по репрезентативной выборке, является следствием кривой распределения трудности заданий теста. Этот факт удачно иллюстрируется на рисунок 31.

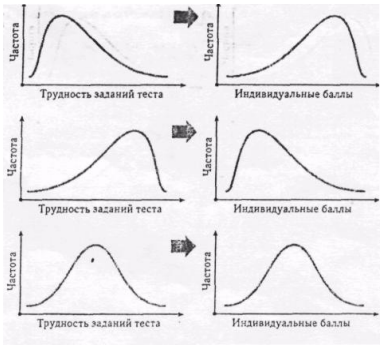


Рисунок 31 - Связь распределения индивидуальных баллов и трудности заданий теста

Для первого распределения слева характерно явное смещение в тесте в сторону легких заданий, что, несомненно, приведет к появлению большого числа завышенных баллов у репрезентативной выборки учеников. Большая часть учеников выполнит почти все задания теста.

Второй случай (слева) отражает существенное смещение в сторону трудных заданий при разработке теста, что не может не сказаться на снижении результатов учеников, поэтому распределение индивидуальных баллов имеет явно выраженный всплеск вблизи начала горизонтальной оси. Основная часть учеников выполнит незначительное число наиболее легких заданий теста.

В третьем случае задания теста обладают оптимальной трудностью, поскольку распределение имеет вид нормальной кривой. Отсюда автоматически возникает нормальность распределения индивидуальных баллов репрезентативной выборки учеников, что в свою очередь позволяет считать полученное распределение устойчивым по отношению к генеральной совокупности.

В профессионально разработанных нормативно-ориентированных тестах типичным является результат, когда приблизительно 70% учеников выполняют правильно от 30 до 70% заданий теста. а наиболее часто встречается результат в 50%.

Шаг 6. Определение выборочных характеристик результатов.

На данном этапе необходимо вычислить среднее значение, моду, медиану, дисперсию, стандартное отклонение выборки, ассиметрию и эксцесс (рисунок 32).

Степень отклонения распределения наблюдаемых частот выборки от симметричного распределения, характерного для нормальной кривой, оценивается с помощью асимметрии. Наличие асимметрии легко установить визуально, анализируя полигон частот или гистограмму Более тщательный анализ можно провести с помощью обобщенных статистических характеристик, предназначенных для оценки величины асимметрии в распределении.

Функция СКОС MS Excel возвращает ассиметрию распределения.

СКОС (число 1; число 2), где число1 – ссылка на массив данных, содержащих индивидуальные баллы учеников.

При интерпретации полученного значения асимметриии 0,277 необходимо обратить внимание на то, что величина ассиметрии получилась положительной и небольшой (рисунок 32,33).

	A	L
1	**Номер**	**Индивиду-**
2	**испытуемых**	**альные баллы**
3	1	6
4	2	2
5	3	1
6	4	9
7	5	4
8	6	4
9	7	5
10	8	4
11	9	9
12	10	6
13	**среднее**	**5**
14	**мода**	**4**
15	**медиана**	**4.5**
16	**дисперсия**	**6.889**
17	**ст. отклонение**	**2.625**
18	**ассиметрия**	**0.277**
19	**эксцесс**	**-0.4117**

Рисунок 32 – Описательные характеристики выборки

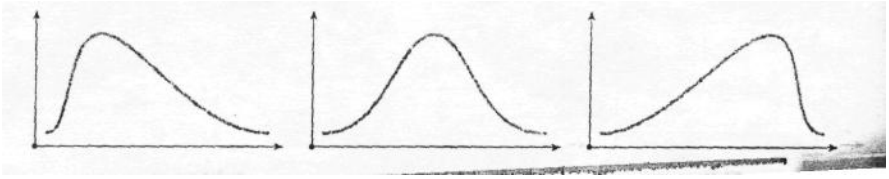


Рисунок 33- Кривые распределения с отрицательной, нулевой и положительной ассиметрией (слева направо) соответственно.

Асимметрия распределения положительна, если основная часть значений индивидуальных баллов лежит справа от среднего значения, что обычно характерно для излишне легких тестов.

Асимметрия распределения баллов отрицательна, если большинство учеников получили оценки ниже среднего балла. Эффект отрицательной асимметрии встречается в излишне трудных тестах, не сбалансированных правильно по трудности при отборе заданий

В хорошо сбалансированном по трудности тесте, как уже отмечалось ранее, распределение баллов имеет вид нормальной кривой. Для нормального распределения характерна нулевая асимметрия, что вполне естественно, так как при полной симметрии каждое значение балла, меньшее среднего значения, уравновешивается другим симметричным, большим чем среднее.

С помощью эксцесса можно получить представление о том, является ли функция распределения частот островершинной, средневершинной или плоской.

Для расчета данного параметра применим функцию ЭКСЦЕСС (число1; число2; …), где число1 – ссылка на массив данных, содержащих индивидуальные баллы учеников.

В том случае, когда распределение данных бимодально (имеет две моды), необходимо говорить об эксцессе в окрестности каждой моды. Бимодальная конфигурация указывает на то, что по результатам выполнения теста выборка учеников разделилась на две группы. Одна группа справилась с большинством легких, а другая с большинством трудных заданий теста.

Лабораторная работа 5

Тема: Практическая работа в статистическом пакете STADIA версия 6.02

Цель: рассмотреть принцип работы приложения STADIA версия 6.03

Теоретические сведения

1. Ввод данных

Электронная таблица пакета Stadia представляет собой матрицу данных, в которой столбцы отвечают переменным, а строки – измерениям значений переменных. Элементы таблицы могут содержать как числовые, так и символьные значения, однако последние используются только лишь в информационных целях.

Чтобы ввести или изменить значение в ячейке необходимо:

сделать эту ячейку активной – щелкнув по ней указателем мыши или используя клавиши управления курсором;

набрать новое значение с клавиатуры и нажать клавишу Enter для перехода к следующей позиции.

Для изменения наименований переменной выделите ее имя щелчком мыши и произведите изменение имени в поле редактирования.

Переход к следующей позиции таблицы осуществляется клавишами перемещения курсора, а смена страниц – клавишами PageDown и PageUp. Для быстрого перемещения по таблице используются линейки прокрутки внизу и справа экрана, управляемые мышью.

Удаление числа в текущей позиции производится клавишей Del.

В таблице также можно выделять отдельные фрагменты данных. Для этого подведите мышь к верхнему левому фрагменту данных, нажмите левую клавишу и, не отпуская ее, ведите мышь к правому нижнему краю фрагмента. Далее выделенный фрагмент можно удалить, забрать в буфер или скопировать.

Для перемещения фрагмента нажмите правую кнопку мыши и, не отпуская ее, ведите указатель (он изменит свою привычную форму на стрелку с листком) до нужной ячейки. Далее отпустите кнопку мыши и переменная будет вставлена в указанное место.

Максимально возможное количество элементов (чисел) в таблице определяется поставленной версией пакета Stadia и может доходить до 20000, а число столбцов – до 500.

Операция записи содержимого страницы осуществляется нажатием функциональной клавиши F4. Чтобы очистить содержимое страницы нажмите F5.

2. Преобразование данных

Блок преобразования данных содержит обширный набор алгебраических, тригонометрических, матричных и других операций, необходимых для преобразования исходных данных в электронной таблице к нужному виду (рис. 1).

Для вызова меню выбора преобразования нужно нажать клавишу F8 или выполнить пункт «Преобраз» в верхней экранной линейке команд.

Операции преобразования разбиты на три группы в зависимости от того, изменяются ли при этом значения одной переменной, нескольких или всех переменных (матричные операции).

Операции над одной переменной

Операции над одной переменной производятся над данными в текущей ячейке (там, где находится указатель мыши). Результат преобразования записывается в ту же самую переменную.

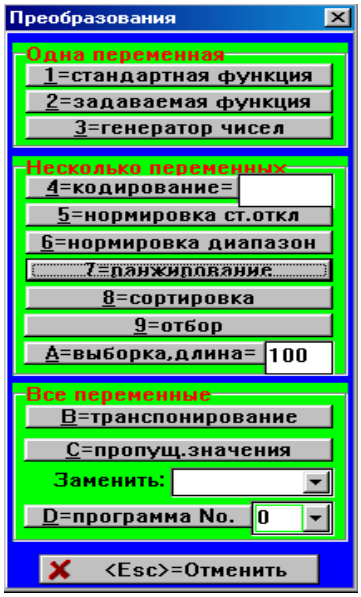


Рисунок 34 - Меню выбора преобразования

Выполнение данного пункта приводит к появлению меню выбора стандартных операций. Данное меню содержит также поля для ввода значений двух параметров a и b.

Задание 1. Найти логарифм 120. Ответ: 2,0791812

Задание 2. Возвести 82 в степень 3. Ответ: 551368

функции, задаваемые по вводимой формуле

Пункт «Задаваемая функция» приводит к появлению типового бланка формул (рисунок 35), в который необходимо ввести новую формулу преобразований или же выбрать одну из имеющихся и нажать Enter.

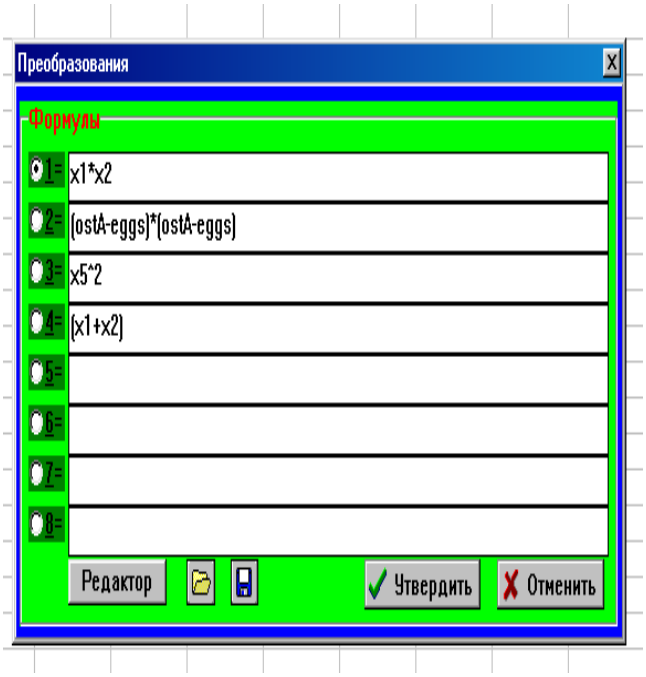


Рисунок 35 - Меню задаваемых функций

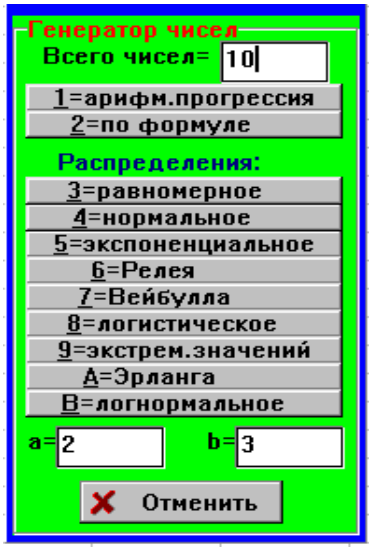


Рисунок 36 - Меню генератора чисел

Задание 3. Найти произведение 64 и 15. Ответ: 960

генератор чисел

Выполнение данного пункта приводит к появлению меню выбора типа генератора (рисунок 36).

В этом меню предоставляется выбор из следующих возможностей:

1. генерация чисел по закону арифметической прогрессии (возрастающей или убывающей) a+b*i, i=0,n-1;

2. генерация чисел по задаваемой формуле от аргумента X, где X изменяет свои значения по арифметической прогрессии a+b*i, i=0,n-l,

3. генераторы случайных чисел, распределенных по следующим законам:

- по равномерному закону в диапазоне от a до b;
- по нормальному закону со средним значением a и стандартным отклонением b;
- по другим законам распределения.

В верхнем поле ввода бланка необходимо предварительно указать количество n генерируемых чисел, а в двух нижних полях ввести значения параметров a, b, если они нужны для выбираемого генератора.

Задание 4. Сгенерировать цепочку из 10 чисел от 0 до 1 по арифметической прогрессии, по равномерному и нормальному закону распределения сгенерировать цепочку чисел из 20 элементов от 1 до 25.

Операции над несколькими переменными

Кодирование значений

Операция кодирования означает замену значений выбранных переменных некоторым кодом, которым может быть как число, так и символьное обозначение (слово, текст). Такая замена производится только для тех значений переменных, которые удовлетворяют вводимому логическому условию.

Перед выполнением этой операции в меню преобразований, расположенном справа от кнопки кодирования, необходимо ввести код, а после нажатия на кнопку «Кодирование» – указать подлежащие выборке переменные. После этого появляется типовой бланк формул, в который надо ввести логическое условие или же выбрать одно из имеющихся и нажать Enter. В качестве переменных в формулах условий можно использовать не только наименования переменных из электронной таблицы, но и обозначения x[i].

Задание 5. В сгенерированной цепочке из 10 значений по арифметической прогрессии заменить все значения большие 6 на 1.

2. Нормировка значений выбранных переменных в электронной таблице может производиться двумя способами:

- по диапазону значений. Из каждого значения переменной вычитается минимальное значение и результат делится на диапазон значений (разность между максимальным и минимальным значениями). При этом все значения становятся положительными и меньше единицы;
- по стандартному отклонению (нормализация или стандартизация). Из каждого значения переменной вычитается среднее значение и делится на стандартное отклонение данной переменной (полученные значения центрированы нулем).

После выбора операции нормировки в окне необходимо указать подлежащие нормировке переменные.

3. Ранжирование осуществляет замену числовых значений выбранных переменных на их ранги, то есть на целые числа, являющиеся порядковыми номерами этих значений в соответствии с их величиной по возрастанию. Совпадающие значения заменяются средними рангами, которые могут принимать дробные значения.

4. Сортировка позволяет переупорядочить выборку из электронной таблицы по возрастанию значений сортирующих переменных.

При выполнении этой операции появляется специальный экранный бланк, который включает следующие элементы: список переменных таблицы; список сортирующих переменных; список сортируемых переменных; кнопки переноса переменных из одного

списка в другой и обратно; кнопку выбора всех переменных в качестве сортируемых; фонарики режима сортировки: по возрастанию значений сортирующих переменных или по убыванию этих значений.

Задание 6. Отранжируйте цепочку из 20 элементов от 1 до 25, построенную по равномерному распределению и отсортируйте ее по убыванию значений.

5. Отбор позволяет оставлять в электронной таблице только те измерения указанных переменных, которые удовлетворяют вводимому логическому условию.

Задание 7. Отберите в цепочке из 20 элементов, построенной с помощью генератора чисел по нормальному закону распределения, все значения >5.

6. Выборка. При выполнении статистического анализа часто возникает следующая задача: имеется достаточно объемная выборка, то есть измеренные значения некоторой переменной. Однако исследователь не уверен, что эти значения получены достаточно случайным образом в ходе корректно организованных измерений. В таких случаях для повышения статистической точности полезно для анализа из такой выборки отобрать случайным образом некоторое подмножество значений (получить подвыборку).

Перед выполнением этой операции в меню преобразований необходимо в расположенное справа поле ввести число отбираемых случайным образом измерений (размер подвыборки), а после нажатия на кнопку «Выборка» - указать подлежащие выборке переменные.

Матричные операции (операции над всеми переменными)

Эти операции производятся над всем содержимым электронной таблицы.

1. Транспонирование. В результате транспонирования матрицы данных в электронной таблице строки становятся столбцами, а столбцы — строками.

Задание 8. Транспонировать следующую матрицу:

x1	x2	x3
2	3	
1	2	4
1	2	5

Анализ пропущенных значений.

Даже в прекрасно организованных и проведенных экспериментах некоторые наблюдения могут быть зарегистрированы неверно или не зарегистрированы совсем. Например, экспериментальное животное может умереть, пациент — не придти на назначенный прием, очередной препарат - оказаться испорченным, а регистрирующий прибор — отказать.

Замена пропущенных значений может быть произведена двумя возможными методами: замена каждого пропущенного значения средним значением, вычисленным для соответствующей переменной или замена регрессионными значениями.

Второй метод обеспечивает более корректную и дифференцированную замену. В соответствии с этим методом для каждой анализируемой переменной, содержащей пропущенные значения выбирается парная переменная по условию максимума коэффициента корреляции. Затем по парной переменной вычисляется линейная регрессия. Все пропущенные значения заменяются регрессионными. Однако если парная переменная также содержит пропущенное значение, то замена производится по методу средних.

Ввод пропущенных значений в электронную таблицу производится посредством набора любого нечислового значения.

При выполнении данной операции в экранную страницу результатов [Rez] выдается таблица пропущенных значений. По этой таблице можно визуально оценить характер распределения пропущенных элементов в матрице данных.

Задание 9. Ввести следующую матрицу:

x1	x2	x3	x4
1		44	-2 3
m	48	m	9

2 51 0 7
8 m 44 -5
m 35 m 2
5 33 2 m
5 32 1 -3
8 -1 0 1

Произвести замену пропущенных значений (m) по методу средних и по методу регрессии, сравнить полученные результаты.

Ответ:

Метод средних

1 44 -2 3
4.8333333 48 7.5 9
2 51 0 7
8 34.571429 44 -5
4.8333333 35 7.5 2
5 33 2 2
5 32 1 -3
8 -1 0 1

Метод регрессии

1 44 -2 3
2.1196388 48 -0.0214695 9
2 51 0 7
8 6.53 44 -5
3.7890645 35 0.1562574 2
5 33 2 1.6703157
5 32 1 -3
80 1

3. Визуализация данных (графические возможности)

Пакет STADIA имеет все современные и необходимые возможности доступного отображения графической информации.

Использование средств графического представления в системе STADIA возможно для:

а) исходных данных (рисунок 37);

б) результатов анализа.

Построение графиков данных производится по нажатию клавиши F6 (или соответствующего пункта из верхней командной строки), а построение графика результатов анализа — при выполнении конкретного статистического метода. В обоих случаях график выдается в отдельную экранную страницу.

При активизации экранной страницы с графиком в третьей инструментальной линейке экрана появляется ряд дополнительных кнопок общего назначения, посредством которых можно модифицировать уже построенный график (рисунок 38):

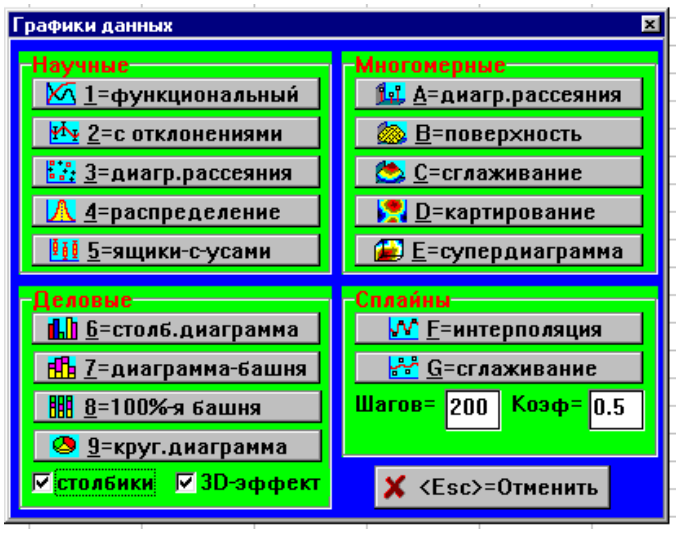


Рисунок 37 - Меню выбора типа графика данных

Рисунок 38 - Дополнительное меню для графиков

1. кнопка сохранения данных с графика в электронную таблицу

2. кнопка изменения толщины линий на графике (действует не на все типы графиков);

3. кнопка переключения: цветное/черно-белое изображение (полезно использовать перед выводом графика на принтер);

4. кнопка добавления подрисуночных надписей и легенд;

5. кнопка добавления/снятия координатной сетки;

Наряду с этим, отдельно для категорий научной графики и деловой графики контекстно появляются еще и дополнительные инструментальные кнопки:

6. кнопка изменения формы графика имеет четыре состояния, циклически изменяемые при каждом нажатии:

— график в виде линий;

— график в виде линий с маркерами точек;

— график в виде маркированных точек;

— график в виде столбиков;

7. кнопка номера зависимости Y=f(X) устанавливает одну из нескольких экспонируемых на графике зависимостей в качестве активной.

Доступные в пакете графические формы разбиты на 4 группы: научная графика, деловая графика, многомерная графика и сплайны.

Научная графика

включает преимущественно наиболее употребительные в научных и инженерных исследованиях формы двухмерных графиков:

1) функциональный график (рисунок 38) отображает одну или несколько экспериментальных или теоретических зависимостей вида Y=f(X), представленных множеством пар значений переменных X,Y. В бланке выбора переменной можно указать и одну переменную, тогда по оси абсцисс будут расположены порядковые номера.

2) график с отклонениями, аналогичен функциональному, но в нем каждое значение Y интерпретируется, как некоторое среднее значение, и для него еще

указывается третья переменная dY, представляющая собой интервал ошибки или стандартное отклонение значений;

3) в диаграмме рассеяния данные интерпретируются как множество пар значений X,Y, каждая из которых представляет координаты некоторой точки в двумерном пространстве;

4) график распределения выборочных значений представляет изображение одной или нескольких выборок (переменных) в порядке возрастания их значений;

5) ящики с усами являются модификацией получившего распространение способа компактного совместного изображения многих выборок с указанием их средних значений и стандартных отклонений.

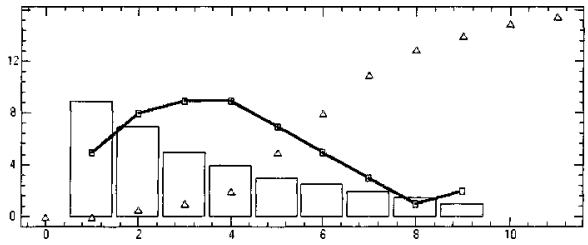


Рисунок 39 - Функциональный график в форме изображения: линиями, точками (диаграмма рассеяния) и столбиками (столбиковая диаграмма)

Деловая графика

Данный раздел объединяет наиболее употребительные формы изображения данных гуманитарного и экономического характера, представленных в виде матрицы со значениями нескольких переменных (столбцы), измеренных у ряда объектов (строки). Из меню деловая графика можно выбрать следующие типы диаграмм:

1. столбиковая диаграмма обеспечивает последовательное линейное расположение значений переменных в виде столбиков;

2. диаграмма-башня изображает значения переменных для каждого объекта одно над другим в виде башни;

3. 100% -я башня представляет вариант диаграммы—башни, у которой каждая вертикальная колонка (значения переменных объекта) нормирована на 100%; такая диаграмма позволяет увидеть процентные сопоставимости каждой переменной у разных объектов;

4. круговая диаграмма изображает значения некоторой переменной у разных объектов в виде секторов круга.

Многомерная графика

объединяет следующие формы представления многомерных данных:

1. диаграмма рассеяния изображает множество триад значений X,Y,Z, в виде точек в трехмерном пространстве;

2. поверхность представляет изображение поверхности в трехмерном| пространстве, заданной алгебраической формулой.

3. поверхность сглаживания представляет изображение поверхности в трехмерном пространстве, сглаживающую множество точек, заданных координатами X,Y,Z;

4. картирование аналогично сглаживанию, но результат представляется в виде двумерной карты, на которой высоты сглаживающей поверхности представлены в цветной или черно-белой тональной шкале, приведенной справа от карты;

5. супердиаграмма предназначена для визуализации многомерных данных (более трех измерений). Супердиаграмма – это динамичное четырехмерное изображение (суперкуб), представляющий срез многомерного пространства.

Задание 10. Произвести сглаживание следующего распределения точек:

X	Y	Z
1	1	20
1	3	100
2	1	100
2	2	200
2	4	20
3	3	150
3	4	50
4	2	20

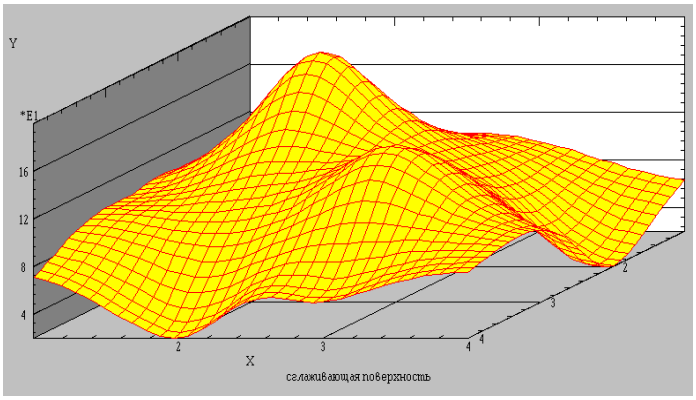


Рисунок 40 – Сглаживание точек

Сплайны можно отнести к специальному разделу научной графики, где промежутки между экспериментальными точками сглаживаются/интерполируются посредством специальных кривых — сплайнов:

1. сплайн-интерполяция обеспечивает прохождение сплайнов непосредственно через заданные точки;

2. сплайн-сглаживание обеспечивает прохождение сплайнов на некотором удалении от заданных точек с меньшими колебаниями.

Для графика сплайн-сглаживания в соседнем поле ввода дополнительно необходимо указать значение коэффициента сглаживания. Этот коэффициент указывает среднее расстояние, на которое сглаживающий сплайн будет отстоять от заданных точек. Чем ближе этот коэффициент к нулю, тем ближе к экспериментальным точкам будет проходить сплайн.

После нажатия клавиши сплайн-графика в бланке переменных необходимо выбрать две переменные из электронной таблицы в качестве X и Y. В бланке выбора можно указать и одну переменную Y, тогда по оси абсцисс будут расположены порядковые числа.

4. Статистический анализ

Блок статистического анализа в пакете Stadia 6.0 содержит набор процедур, реализующих широко применяемые методы анализа данных и представления результатов.

Чтобы провести статистический анализ необходимо выполнить ряд последовательных шагов:

1. Ввести данные в электронную таблицу (пункт 1. Ввод данных). Обрабатываемые данные должны соответствовать выбранному методу анализа.

2. Вызвать меню статистических методов (рисунок 41) нажатием клавиши F9. В этом меню нажмите на кнопку нужного метода.

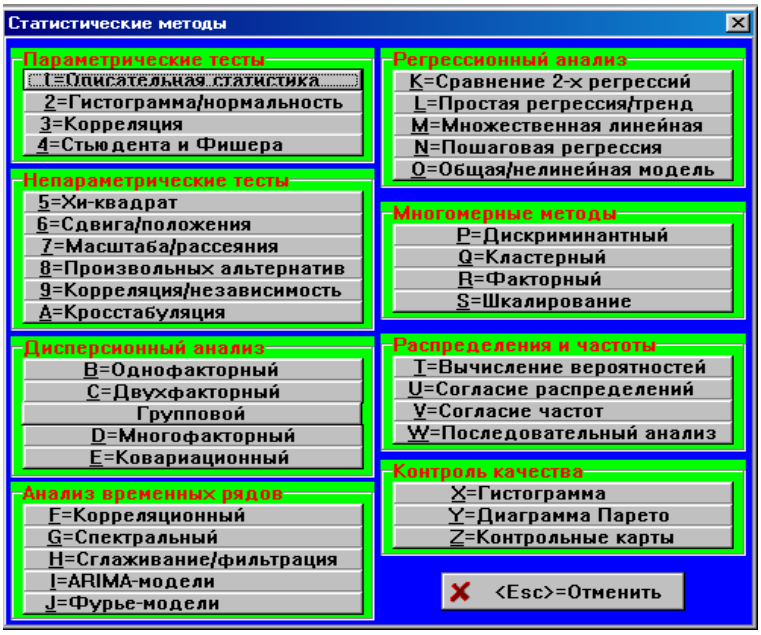


Рисунок 41 - Меню выбора статистического метода

3. Далее появляется блок выбора переменных, в котором надо определить подлежащие анализу переменные. Далее протекает диалог, характерный для выбранного Вами метода.

4. Выдача числовых результатов и их интерпретация их происходит в экранной странице [Rez]. Результаты анализа можно перенести в электронную таблицу через буфер обмена.

При проверке статистических гипотез STADIA выводит на экран вычисленное значение уровня значимости P и сообщение-подсказку о возможности принятия или непринятия нулевой гипотезы по условию P>0.05. (критический уровень 0.05 может быть изменен при настройке).

STADIA выводит не только вычисленное значение уровня значимости P, но и значение статистики критерия T, а также значения специальных параметров распределения (обычно называемых числом степеней свободы).

5. Результаты статистического анализа в виде графика появляются на экранной странице [Gri], i=1,8.

1 Параметрические критерии

В группу параметрических процедур входят методы для вычисления описательных статистик, построения графиков на нормальность распределения, проверка гипотез о принадлежности двух выборок одной совокупности. Эти методы основываются на предположении о том, что распределение выборок подчиняется нормальному (гауссовому) закону распределения.

Описательная статистика

Данная процедура вычисляет общеупотребительные выборочные характеристики распределения. Размер выборки для данного метода должен быть больше 4 и меньше 100 (Demo-версия). После запуска процедуры в типовом бланке необходимо выбрать для анализа одну, несколько или все переменные из электронной таблицы. Результатом выполнения будет выдача на экран следующих основных характеристик (рисунок 42):

Далее по подтверждению может быть выдана дополнительная статистика (рисунок 42):

медиана, квартили, размах доверительного интервала, границы доверительного интервала для дисперсии, ошибка стандартного отклонения

коэффициенты асимметрии и эксцесса с уровнями значимости P. Нулевая гипотеза определена как отсутствие различий данного распределения от нормального распределения по каждому из коэффициентов. Если P>0,05 – нулевая гипотеза может быть принята.

Результаты

ОПИСАТЕЛЬНАЯ СТАТИСТИКА. Файл:

Переменная	Размер	<---Диапазон--->		Среднее	---Ошибка	Дисперс	Ст.откл	Сумма
x1	10	1	491	50	49	2.401E4	155	500
x2	10	49	51	50	0.3333	1.111	1.054	500

Переменная	Медиана	<--Квартили-->		ДовИнтСр.	<-ДовИнтДисп->		Ош.СтОткл
x1	1	1	1	109.7	1.136E4	8.004E4	74.71
x2	50	49	51	0.7465	0.5259	3.704	0.5083

Переменная	Асимметр.	Значим	Эксцесс	Значим
x1	2.667	0	8.111	0
x2	0	0.4999	1	0.0269

Рисунок 42 - Результат вычисления показателей описательной статистики

Задание 11. Вычислить показатели описательной статистики и сделать выводы о нормальности данных распределений.

x1	1	1	1	1	1	1	1	1	1	491
x2	49	51	49	51	49	51	49	51	49	51

Гистограмма и проверка на нормальность

Задание 12. Вычислить гистограмму и проверить выборку на нормальность для переменной x1 с построением результирующего графика

51	73	55	40	58	48	58	69	61	33

Для выполнения данного задания необходимо ввести данные в таблицу и запустить процедуру 2=Гистограмма/нормальность в окне выбора Статистических методов. Далее выбрать переменную для анализа данных (x1) и нажать кнопку Утвердить.

Затем в бланке Гистограмма указать число интервалов и область определения гистограммы. В качестве числа интервалов программа Stadia показывает значение, вычисленное по формуле $\mathrm{int}(1.5+3.3\cdot\log_{10}(N))$. Область определения (левая граница и правая граница) равна диапазону значений. Если возникнет необходимость можно вручную с клавиатуры изменить число интервалов и область определения.

После нажатия на кнопку Утвердить для каждого интервала на экран выводятся следующие значения (рисунок 43): левая граница интервала в исходных единицах и в единицах стандартного отклонения; число выборочных значений, попавших в интервал (в числовом и процентном выражении); накопленное число выборочных значений до текущего интервала включительно (в числовом и процентном выражении).

Затем проводится проверка нулевой гипотезы об отсутствии различий между выборочным и нормальным распределениями и выдача трех различных статистик:

Колмогорова с уровнем значимости P;

Омега-квадрат с уровнем значимости P;

Хи-квадрат с уровнем значимости P.

При P>0,05 нулевая гипотеза может быть принята.

```
ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ.  Файл:

    X-лев. X-станд  Частота        %   Накопл.       %
        33   -1.766        2       20        2       20
        43  -0.9484        2       20        4       40
        53  -0.1308        4       40        8       80
        63   0.6868        2       20       10      100
        73    1.504
Колмогоров=0.1131, Значимость=2.322, степ.своб = 10
   Гипотеза 0: <Распределение не отличается от нормального>
Омега-квадрат=0.01733, Значимость=1.448, степ.своб = 10
   Гипотеза 0: <Распределение не отличается от нормального>
Хи-квадрат=1.198, Значимость=0.2737, степ.своб = 1
   Гипотеза 0: <Распределение не отличается от нормального>
```

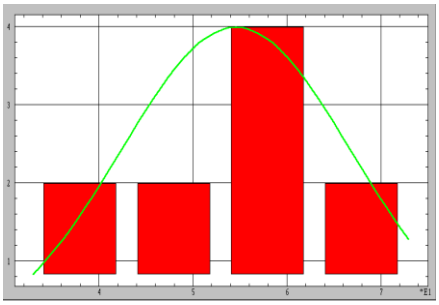


Рисунок 43 - Гистограмма распределения и кривая нормального распределения

9. Результаты выполнения задания 12

Линейная корреляция

Коэффициент корреляции определяет степень, тесноту линейной связи между величинами и может принимать значения от –1 до +1 — в зависимости от тесноты и характера связи между данными СВ.

Задание 13. Вычислить значение коэффициента корреляции.

x1	51	73	55	40	58	48	58	69	61	33
x2	66	69	67	58	87	54	91	95	88	55

После запуска процедуры 3=Корреляция в окне выбора Статистических методов нужно выбрать для анализа несколько переменных из электронной таблицы и нажать кнопку Утвердить.

Процедура вычисляет (рисунок 44):

коэффициент корреляции Пирсона

статистику Стьюдента с n-2 степенями свободы

уровень значимости P нулевой гипотезы;

число степеней свободы.

```
ПАРАМЕТРИЧЕССКАЯ КОРРЕЛЯЦИЯ.  Файл:
                          Переменные: x1, x2
Коэфф.корреляции=0.6933 T:=2.721, Значимость=0.0253, степ.своб = 8
   Гипотеза 1: <Коэффициент корреляции отличен от нуля>
```

Рисунок 44 - Выдача результатов по заданию 13

Задание 14. Вычислить значение коэффициента корреляции для данных тестирования из лабораторной работы №4:

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1	1	1	1	1	0	0	0	0
1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0
1	1	0	1	1	1	1	1	1	1
1	0	1	0	1	1	0	0	0	0
1	1	1	0	0	0	0	1	0	0
1	1	1	1	0	1	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	0

Рисунок 45 – Результаты тестирования

После запуска процедуры 3=Корреляция в окне выбора Статистических методов нужно выбрать для анализа ВСЕ переменные из электронной таблицы и нажать кнопку Утвердить. В окне электронный таблицы появятся результаты подсчета значений коэффициента корреляции между результатами по отдельным заданиям теста. (рисунок 46)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	0,66666667	0,50917508	0,40824829	0,33333333	0,33333333	0,21821789	-0,40824829	0,16666667	0,111111
0,66666667	1	0,21821789	0,61237244	0	0	0,32732684	-0,10206207	0,25	0,166666
0,50917508	0,21821789	1	0,35634832	0,21821789	0,21821789	-0,047619048	-0,35634832	-0,21821789	-0,50917
0,40824829	0,61237244	0,35634832	1	0,40824829	0,40824829	0,53452248	-0,16666667	0,40824829	0,272165
0,33333333	0	0,21821789	0,40824829	1	0,6	0,65465367	0	0,5	0,333333
0,33333333	0	0,21821789	0,40824829	0,6	1	0,21821789	0	0,5	0,333333
0,21821789	0,32732684	-0,047619048	0,53452248	0,65465367	0,21821789	1	0,35634832	0,76376262	0,509175
-0,40824829	-0,10206207	-0,35634832	-0,16666667	0	0	0,35634832	1	0,61237244	0,408248
0,16666667	0,25	-0,21821789	0,40824829	0,5	0,5	0,76376262	0,61237244	1	0,666666
0,11111111	0,16666667	-0,50917508	0,27216553	0,33333333	0,33333333	0,50917508	0,40824829	0,66666667	1

Рисунок 46 - Матрица коэффициентов корреляции задания 14

Анализ значений коэффициента корреляции позволяет выделить третье и восьмое задание теста, так как они отрицательно коррелируют с другими заданиями. Отрицательные значения коэффициента указывают на определенный просчет разработчиков в содержании этих заданий теста.

Критерий Стьюдента и Фишера

Критерий Фишера для двух выборок проверяет нулевую гипотезу о равенстве дисперсий двух выборок, а критерий Стьюдента – гипотезу о равенстве выборочных средних.

Задание 15. В двух группах учащихся — экспериментальной и контрольной — получены следующие результаты по учебному предмету.

Требуется выявить различия этих двух методов по критерию Стьюдента и Фишера.

После ввода исходных данных и запуска процедуры анализа в типовом бланке 4=Стьюдента и Фишера в окне выбора Статистических методов нужно выбрать для анализа две переменные (x1, x2). Результатом выполнения процедуры будет выдача на экран в окне Rez значения следующих статистик (рисунок 47):

статистика Фишера F;

статистика Стьюдента T (в зависимости от результатов сравнения дисперсий применяются различные формулы вычисления статистики);

в случае равенства размеров выборок выдается также статистика Стьюдента, применимая для парных переменных.

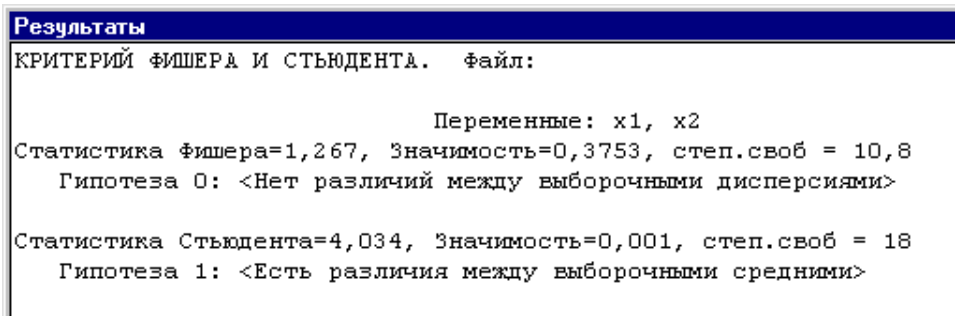


Рисунок 47 - Выдача результатов по заданию 15

Задание 16. Проверить эффективность проведенной работы по формированию ориентации на художественно-эстетические ценности до начала эксперимента и после:

x1	x2
14	18
20	19
15	22
11	17
16	24
13	21
16	25
19	26
15	24
9	15

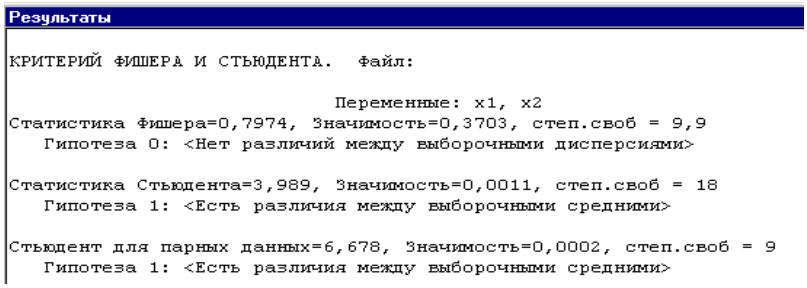


Рисунок 48 - Выдача результатов по заданию 16

Лабораторная работа 6

Тема: Использование электронных таблиц Excel и статистического пакета Stadia для проведения корреляционного анализа

Цель: с помощью электронных таблиц Excel и статистического пакета Stadia рассмотреть проведения корреляционного анализа

Теоретические сведения

Корреляционный анализ

Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между выборками. Обычно связь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет

строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей различают корреляцию и регрессию.

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами X и Y. В качестве меры такой связи используется коэффициент корреляции. Коэффициент корреляции оценивается по выборке объема n связанных пар наблюдений (xi, yi) из совместной генеральной совокупности X и Y. Существует несколько типов коэффициентов корреляции, применение которых зависит от измерения (способа шкалирования) величин X и Y.

Для оценки степени взаимосвязи величин X и Y, измеренных в количественных шкалах, используется коэффициент линейной корреляции (коэффициент Пирсона), предполагающий, что выборки X и Y распределены по нормальному закону.

Коэффициент корреляции — параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле:

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

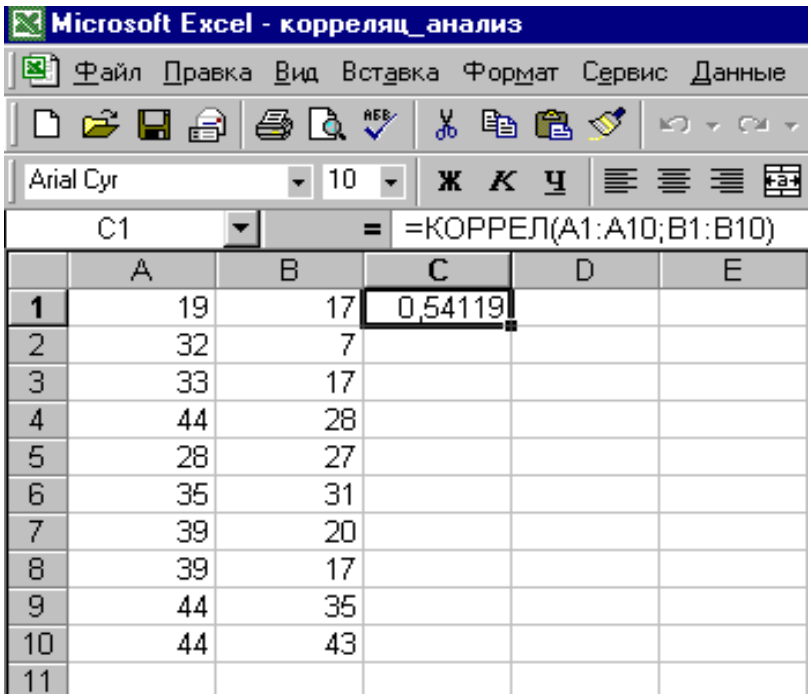
Коэффициент корреляции изменяется от -1 (строгая обратная линейная зависимость) до 1 (строгая прямая пропорциональная зависимость). При значении 0 линейной зависимости между двумя выборками нет.

В MS Excel для вычисления парных коэффициентов линейной корреляции используется специальная функция КОРРЕЛ (массив1; массив2),

где массив1 – ссылка на диапазон ячеек первой выборки (X);

массив2 – ссылка на диапазон ячеек второй выборки (Y).

Пример 1. 10 школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Исследователя интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная X — обозначает среднее время решения наглядно-образных, а переменная Y— среднее время решения вербальных заданий тесто.



	A	B	C	D	E
1	19	17	0,54119		
2	32	7			
3	33	17			
4	44	28			
5	28	27			
6	35	31			
7	39	20			
8	39	17			
9	44	35			
10	44	43			
11					

Рисунок 49 - Результаты вычисления коэффициента корреляции

Таблица 8 - Данные для выявления степени взаимосвязи

№ испытуемых	X	Y
1	19	17
2	32	7
3	33	17
4	44	28
5	28	27
6	35	31
7	39	20
8	39	17
9	44	35
10	44	43

Решение: Для выявления степени взаимосвязи, прежде всего, необходимо ввести данные в таблицу MS Excel (табллица 8, рисунок 49). Затем вычисляется значение коэффициента корреляции. Для этого курсор установите в ячейку С1. На панели инструментов нажмите кнопку Вставка функции (fx). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию КОРРЕЛ, после чего нажмите кнопку ОК. Указателем мыши введите диапазон данных выборки X в поле массив1 (A1:A10). В поле массив2 введите диапазон данных выборки У (B1:B10). Нажмите кнопку ОК. В ячейке С1 появится значение коэффициента корреляции — 0,54119. Далее необходимо по статистическим таблицам определить критические значения для полученного коэффициента корреляции (см. лекцию 7 Приложение 3). При нахождении критических значений для вычисленного коэффициента линейной корреляции Пирсона число степеней свободы рассчитывается как k = n – 2 = 8.

ккрит=0,63 > 0,54 , следовательно, гипотеза Н1 отвергается и принимается гипотеза Н0, иными словами, связь между временем решения наглядно-образных и вербальных заданий теста не доказана.

Задание для самостоятельной работы:

1. Определите, имеется ли взаимосвязь между рождаемостью и смертностью (количество на 1000 человек) в Санкт-Петербурге:

Таблица 9 - Взаимосвязь между рождаемостью и смертностью

Годы	Рождаемость	Смертность
1991	9,3	12,5
1992	7,4	13,5
1993	6,6	17,4
1994	7,1	17,2
1995	7,0	15,9
1996	6,6	14,2
1997	7,1	16
1998	8,2	13,4

Ответ: коэффициент корреляции равен –0,726

2. Рассчитайте коэффициент корреляции Пирсона из примера 1 и задания 1 в статистическом пакете Stadia (лаб. 5). Для этого выбираем процедуру 3=Корреляция в

окне Статистические методы – Параметрические тесты. Совпадают ли полученные значения.

Множественная корреляция

При большом числе наблюдений, когда коэффициенты корреляции необходимо последовательно вычислять для нескольких выборок, для удобства получаемые коэффициенты сводят в таблицы, называемые корреляционными матрицами.

Корреляционная матрица — это квадратная таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

В MS Excel для вычисления корреляционных матриц используется процедура Корреляция из пакета Анализ данных. Процедура позволяет получить корреляционную матрицу, содержащую коэффициенты корреляции между различными параметрами.

Для реализации процедуры необходимо:

1. выполнить команду Сервис - Анализ данных;

2. в появившемся списке Инструменты анализа выбрать строку Корреляция и нажать кнопку ОК;

3. в появившемся диалоговом окне указать Входной интервал, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Входной интервал должен содержать не менее двух столбцов.

4. в разделе Группировка переключатель установить в соответствии с введенными данными (по столбцам или по строкам);

5. указать выходной интервал, то есть ввести ссылку на ячейку, с которой будут показаны результаты анализа. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные. Нажать кнопку ОК.

В выходной диапазон будет выведена корреляционная матрица, в которой на пересечении каждых строки и столбца находится коэффициент корреляции между соответствующими параметрами. Ячейки выходного диапазона, имеющие совпадающие координаты строк и столбцов, содержат значение 1, так как каждый столбец во входном диапазоне полностью коррелирует сам с собой

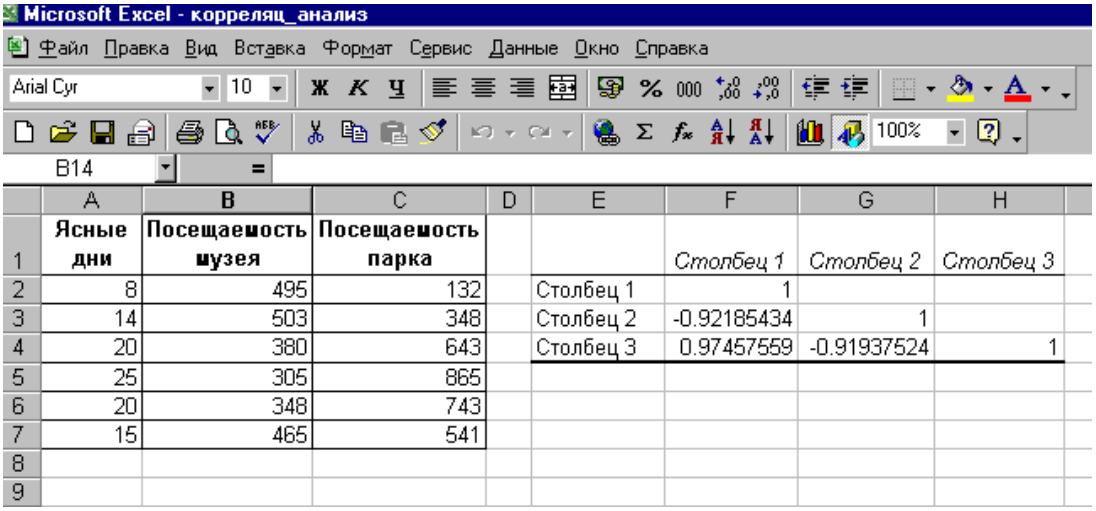
Рассматривается отдельно каждый коэффициент корреляции между соответствующими параметрами. Отметим, что хотя в результате будет получена треугольная матрица, корреляционная матрица симметрична. Подразумевается, что в пустых клетках в правой верхней половине таблицы находятся те же коэффициенты корреляции, что и в нижней левой (симметрично расположенные относительно диагонали).

Пример 2. Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью музеев и парков (таблица 10). Необходимо определить, существует ли взаимосвязь между состоянием погоды и посещаемостью музеев и парков.

Таблица 10 - Данные наблюдений

Число ясных дней	Количество посетителей музея	Количество посетителей парка
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

Решение. Для выполнения корреляционного анализа введите в диапазон A1:G3 исходные данные (рисунок 50). Затем в меню Сервис выберите пункт Анализ данных и далее укажите строку Корреляция. В появившемся диалоговом окне укажите Входной интервал (A2:C7). Укажите, что данные рассматриваются по столбцам. Укажите выходной диапазон (E1) и нажмите кнопку ОК.



	A	B	C	D	E	F	G	H
1	**Ясные дни**	**Посещаемость музея**	**Посещаемость парка**			*Столбец 1*	*Столбец 2*	*Столбец 3*
2	8	495	132		Столбец 1	1		
3	14	503	348		Столбец 2	-0.92185434	1	
4	20	380	643		Столбец 3	0.97457559	-0.91937524	1
5	25	305	865					
6	20	348	743					
7	15	465	541					
8								
9								

Рисунок 50 - Результаты вычисления корреляционной матрицы из примера 2

На рис. 2 видно, что корреляция между состоянием погоды и посещаемостью музея равна -0,92, а между состоянием погоды и посещаемостью парка — 0,97, между посещаемостью парка и музея —

-0,92.

Таким образом, в результате анализа выявлены зависимости: сильная степень обратной линейной взаимосвязи между посещаемостью музея и количеством солнечных дней и практически линейная (очень сильная прямая) связь между посещаемостью парка и состоянием погоды. Между посещаемостью музея и парка имеется сильная обратная взаимосвязь.

Задание для самостоятельной работы

1. 10 менеджеров оценивались по методике экспертных оценок психологических характеристик личности руководителя. 15 экспертов производили оценку каждой психологической характеристики по пятибальной системе (таблица 11). Психолога интересует вопрос, в какой взаимосвязи находятся эти характеристики руководителя между собой.

Таблица 11 - Психологические характеристики

Испытуемые п/п	тактичность	требовательность	критичность
1	70	18	36
2	60	17	29
3	70	22	40
4	46	10	12
5	58	16	31
6	69	18	32
7	32	9	13

8	62	18	35
9	46	15	30
10	62	22	36

Ответ: все три оцениваемые качества оказывают существенное влияние друг на друга, иными словами, такие качества личности менеджера, как критичность, тактичность и требовательность, выступают единым комплексом и в очень большой степени необходимы для успешности его профессиональной работы (рисунок 51).

	A	B	C	D	E	F	G	H
1	70	18	36			*Столбец 1*	*Столбец 2*	*Столбец 3*
2	60	17	29		Столбец 1	1		
3	70	22	40		Столбец 2	0.86376615	1	
4	46	10	12		Столбец 3	0.85224321	0.94868538	1
5	58	16	31					
6	69	18	32					
7	32	9	13					
8	62	18	35					
9	46	15	30					
10	62	22	36					
11								

Рисунок 51 - Результаты вычисления корреляционной матрицы из задания 1

2. Постройте корреляционную матрицу из примера 2 и задания 1 в статистическом пакете Stadia (лаб. 5). Для этого выбираем процедуру 3=Корреляция в окне Статистические методы – Параметрические тесты. Совпадают ли полученные значения.

Лабораторная работа 7

Тема: Использование электронных таблиц Excel и статистического пакета Stadia для проведения дисперсионного анализа

Цель: с помощью электронных таблиц Excel и статистического пакета Stadia рассмотреть проведения дисперсионного анализа

Теоретические сведения

Дисперсионный анализ предназначен для исследования задачи о действии на измеряемую случайную величину (отклик) одного или нескольких независимых факторов.

В MS Excel для проведения однофакторного дисперсионного анализа используется процедура Однофакторный дисперсионный анализ.

Для проведения дисперсионного анализа необходимо:

1. ввести данные в таблицу, так чтобы в каждом столбце оказались данные, соответствующие одному значению исследуемого фактора, а столбцы располагались в порядке возрастания (убывания) величины исследуемого фактора;

2. выполнить команду Сервис > Анализ данных;

3. в появившемся диалоговом окне Анализ данных в списке Инструменты анализа выбрать процедуру Однофакторный дисперсионный анализ, затем нажать кнопку ОК;

4. в появившемся диалоговом окне задать Входной интервал, то есть ввести ссылку на диапазон анализируемых данных, содержащий все столбцы данных.

5. в разделе Группировка переключатель установить в положение по столбцам;

6. указать выходной интервал, то есть ввести ссылку на ячейку, с которой будут показаны результаты анализа. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные. Нажать кнопку ОК.

Выходной диапазон будет включать в себя результаты дисперсионного анализа: средние, дисперсии, критерий Фишера и другие показатели. Влияние исследуемого фактора определяется по величине значимости критерия Фишера, которая находится в таблице Дисперсионный анализ на пересечении строки Между группами и столбца Р-Значение. В случаях, когда Р-Значение < 0,05, критерий Фишера значим и влияние исследуемого фактора можно считать доказанным.

Кроме рассмотренной процедуры однофакторного дисперсионного анализа, для проведения двухфакторного дисперсионного анализа в пакете анализа реализованы процедуры Двухфакторный дисперсионный анализ с повторениями и Двухфакторный дисперсионный анализ без повторений.

Пример 1. Необходимо выявить, влияет ли расстояние от центра города на степень заполняемости гостиниц. Пусть введены 3 уровня расстояний от центра города: 1) до 3 км, 2) от 3 до 5 км и 3) свыше 5 км. Данные заполняемости представлены в таблице 12.

Решение

Таблица 12 - Данные заполняемости

Расстояние	Заполняемость
До 3 км	92 98 89 97 90 94
От 3 до 5 км	90 86 84 91 83 82
Свыше 5 км	87 79 74 85 73 77

Исследуемые данные введите в рабочую таблицу Excel по столбцам: в столбец А — заполняемость гостиниц в центре города, в столбец В — гостиниц, находящихся на расстоянии от 3 до 5 км и т. д. (диапазон А1:С6).

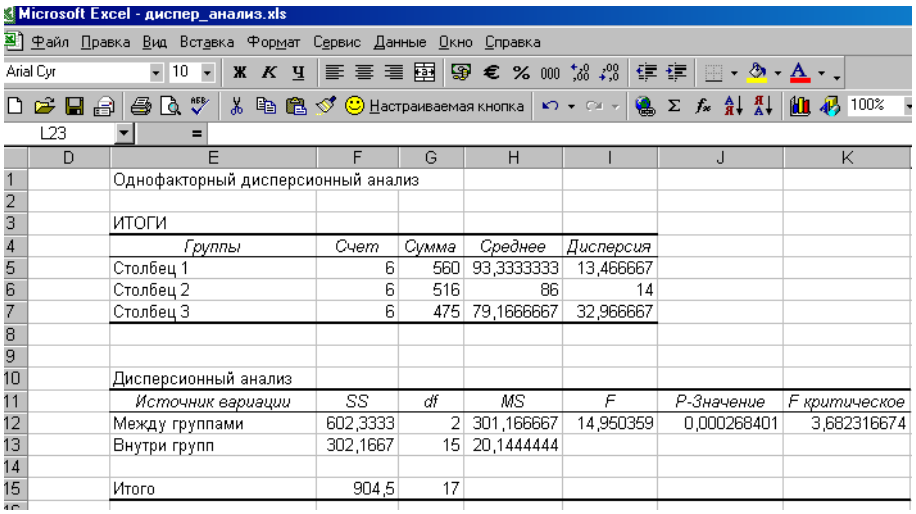
Выполните команду Сервис > Анализ данных. В появившемся диалоговом окне Анализ данных в списке Инструменты анализа щелчком мыши выберите процедуру Однофакторный дисперсионный анализ. Нажмите кнопку ОК.

В появившемся диалоговом окне Однофакторный дисперсионный анализ в поле Входной интервал задайте А1:С6.

В разделе Группировка переключатель установите в положение по столбцам.

Укажите выходной диапазон Е1 и нажмите Ок.

В результате будет получена следующая таблица рисунок 52. В таблице Дисперсионный анализ на пересечении строки Между группами и столбца Р-значение находится величина 0,0002684 <0,05, следовательно, критерий Фишера значим и влияние фактора расстояния от центра города на эффективность заполнения гостиниц доказана статистически.



Microsoft Excel - диспер_анализ.xls

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

Однофакторный дисперсионный анализ

ИТОГИ

Группы	Счет	Сумма	Среднее	Дисперсия
Столбец 1	6	560	93,3333333	13,466667
Столбец 2	6	516	86	14
Столбец 3	6	475	79,1666667	32,966667

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	602,3333	2	301,166667	14,950359	0,000268401	3,682316674
Внутри групп	302,1667	15	20,1444444			
Итого	904,5	17				

Рисунок 52 - Результаты примера 1

Задание для самостоятельной работы:

Определите влияет ли фактор образования на уровень зарплаты сотрудников фирмы на основании следующих данных (таблица 13).

Таблица 13 - Данные по зрплате

Образование	Зарплата сотрудников
Высшее	3200 3000 2600 2000 1900 1900
Среднее спец.	2600 2000 2000 1900 1800 1700
среднее	2000 2000 1900 1800 1700 1700

2. Исследователь сравнивает эффективность четырех разных методик обучения производственным навыкам. Для этой цели из всех выпускников ПТУ выбраны четыре группы учащихся, обучавшиеся, соответственно четырьмя разными методами. Эффективность методик оценивалась по сумме обработанных учащимися деталей в течение одного дня (таблица 14).

Таблица 14 - Сравнение методик

№ учащихся	1 группа	2 группа	3 группа	4 группа
1	60	75	60	95
2	80	66	80	85
3	75	85	65	100
4	80	80	60	80
5	85	70	86	
6	70	80	75	
7		90		

3. Проведите дисперсионный анализ для примера 1, задания 1 и задания 2 в статистическом пакете Stadia (лаб. 5). Для этого выбираем процедуру B=Однофакторный в окне Статистические методы – Дисперсионный анализ. При запросе метода выбираем 1=параметрический. Совпадают ли полученные значения.

Список использованных источников

1 Использование электронных таблиц Excel для вычисления выборочных характеристик данных / [Электронный ресурс] / Режим доступа: https://sites.google.com/site/ktnoscience/Home/lab/lab1w

2 Использование электронных таблиц Excel для построения распределений случайных величин (СВ) и генерации случайных чисел / [Электронный ресурс] / Режим доступа: https://sites.google.com/site/ktnoscience/Home/lab/lab2w

3/ Использование электронных таблиц Excel для обработки данных тестирования / [Электронный ресурс] /Режим доступа: https://sites.google.com/site/ktnoscience/Home/lab/lab3w

4 Использование электронных таблиц Excel для обработки данных тестирования / [Электронный ресурс] / Режим доступа: https://sites.google.com/site/ktnoscience/Home/lab/lab4w

5 Решение математических задач средствами Excel: Практикум/ В.Я. Гельман. – СПб.: Питер, 2013 - с. 168-172

6 Психологические тесты. Т.2. Под ред. А.А. Карелина. - М., ВЛАДОС, 2010, стр. 99