Программные средства обработки данных

Презентационные материалы лекционных занятий

В современных условиях наблюдается постоянный рост интенсивности информационных потоков и объемов обрабатываемой информации. Это требует непрерывного обновления знаний о состоянии предметной области и перспектив развития. При решении задач планирования и выбора стратегии развития предприятия, фирмы можно выделить ряд блоков задач, в решении которых используется статистическая информация:

− формирование стратегических целевых установок фирмы;
− прогнозирование потребности в материальных, энергетических, трудовых и финансовых ресурсах;
− анализ конкурентов и рынков сбыта;
− анализ спроса и предложений;
− оценка финансовой деятельности предприятия;
и многие другие.

Особенность решения таких задач заключается не только в обработке большого объема информации, но и в необходимости выявления причинно-следственных связей, построении формализованных моделей для анализа и прогноза.

К основным предпосылкам применения современных информационных технологий в области статистической обработки информации можно отнести следующие:

− большое количество объектов статистического наблюдения, многомерность данных;

− необходимость отслеживания динамики массива показателей во времени, формирование на их основе различных сводок;

− низкую оперативность обработки данных;

− высокие материальные и трудовые затраты на сбор и обработку статистической информации;

− территориальную разобщенность исходных данных, необходимость их интеграции и одновременной обработки;

− сложность математических методов анализа данных.

В последнее время получили широкое распространение программные средства или информационные системы, предназначенные для автоматизации работ статистической обработки данных, которые позволяют собирать, хранить и обрабатывать разнородные массивы данных с использованием единой информационной базы. Такие системы на предприятии ориентируются на потребности руководства при выполнении функций управления на основе внутренних и внешних статистических данных.

Достоинством таких систем является адаптация информационной базы и функций системы к условиям функционирования предприятия. Однако, в силу сложности реализации математических методов, такие системы, как правило, включают лишь ограниченный набор аналитических методов.

В настоящее время получили распространение статистические пакеты, которые могут быть легко подключены к существующей информационной системе обработки информации на предприятии. В нашей стране наибольшее распространение получили следующие статистические пакеты:

- STATISTICA;
- SPSS;
- Deductor.

Пакет прикладных программ STATISTICA – универсальная система анализа данных, разработанная компанией StatSoft, построенная по модульному принципу, каждый модуль выполняет определенный набор функций и может быть использован и автономно. Основные возможности пакета:

– реализует широкий набор математических методов;

– дает возможность представить графическую интерпретацию результатов (в графиках типа 2D, 3D, пиктограммах или в разработанных в собственном дизайне графиках);

– осуществляет поддержку всех стандартов современных офисных приложений (импорт данных из электронных таблиц, в том числе и их MS Excel, экспорт диаграмм в приложения MS Office и др.);

– позволяет расширять возможности пакета за счет встроенного языка программирования Statistica Visual Basic.

Пакет STATISTICA может применяться в разнообразных сферах деятельности:

– в банковской деятельности (для анализа кредитных рисков и прогнозирования финансовых показателей);

– торговой деятельности (для сравнительного анализа поставщиков и прогнозирования потребления товаров и ресурсов);

– маркетинговых исследованиях (для изучения сезонности спроса, классификации товара по потребительским свойствам);

– производственной деятельности (для прогнозирования потребности материальных ресурсов, выявления причинно-следственных связей между технологическими параметрами, анализа надежности и долговечности продукции);

– медицинском обслуживании (для анализа результатов обследования, диагностики);

– социологических исследованиях (для анализа опроса общественного мнения).

Кроме этого, пакет STATISTICA является базовым статистическим пакетом в большинстве вузов России, служит для обучения методам статистического анализа.

Пакет прикладных программ SPSS (Statistical Package for Social Science) – статистический пакет, разработанный компанией SPSS Inc, предназначенный для работы в операционной системе MS Windows. Является пакетом обработки и анализа социологических данных. Основные возможности пакета:

− реализует набор математических методов статистической обработки данных;

− осуществляет доступ к территориально распределенным данным и позволяет объединять несколько баз данных;

− формирует нестандартные отчеты, позволяющие оценить данные с разных точек зрения;

− осуществляет настройку интерфейса и процедур работы с данными с помощью встроенного языка сценариев;

− поддерживает связь с большинством форматов данных и обмен данными с другими приложениями MS Windows.

Пакет прикладных программ Deductor

– статистический пакет, разработанный фирмой Base Group Labs, состоит из 3-х частей: многомерного хранилища данных Deductor Warehouse, аналитического приложения Deductor Studio и рабочего места конечного пользователя Deductor Viewer.

Deductor Warehouse – многомерное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию.

Deductor Studio – программа, реализующая функции импорта, обработки, визуализации и экспорта данных. В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки, используя **Мастера обработки** (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты на сторону. Это полностью соответствует концепции извлечения знаний из баз данных.

Deductor Viewer – рабочее место конечного пользователя. Позволяет отделить процесс построения моделей от использования уже готовых моделей. Все сложные операции по подготовке моделей выполняются аналитиками-экспертами при помощи Deductor Studio, а Deductor Viewer обеспечивает пользователям простой способ работы с готовыми результатами.

Реализованные в Deductor обработчики покрывают основную потребность в анализе данных и создании законченных аналитических решений на базе Data Mining.

Анализ возможностей различных пакетов (для сравнения рассмотрен и пакет MS Excel) позволил сформулировать их преимущества и недостатки и дать рекомендации по их применению:

Функции и методы	Пакеты прикладных программ			
	MS Excel	STATISTICA	SPSS	Deductor
Описательные методы статистического анализа: 1) вычисления математических ожиданий, дисперсий изучаемых величин и др.	встроенные функции Excel	модуль Описательной статистики	команда Descriptives	при выполнении функции Линейная регрессия
2) проверка гипотез о равенстве математических ожиданий	функции пакета Анализа данных	модуль Описательной статистики	широкий спектр команд One sample T-test, Independent sample T-test и др. непараметрические методы	-
3) построение гистограмм	функции пакета Анализа данных	модуль Описательной статистики	команды FREQUECIES STATISTICS, HISTOGRAM	-
Построение модели временного ряда и прогнозирование с учетом сезонных колебаний и периодических трендов	требуется самостоятельно создавать шаблон на листе Excel	модуль Временные ряды и прогнозирование с поквартальной и месячной десонализацией	-	-
Построение многомерной линейной регрессионной модели	встроенная функция ЛИНЕЙН и функция пакета Анализ данных РЕГРЕССИЯ	модуль Множественная регрессия	линейная регрессия в процедуре REGRESSION	функция Линейная регрессия
Построение нелинейной регрессионной модели	встроенные функции позволяют построить полиномиальную и экспоненциальную модели	модуль Множественная регрессия дает большой выбор нелинейных моделей	логистическая регрессия в процедуре REGRESSION	-
Корреляционный анализ	встроенные функции Excel КОРРЕЛ, КОВАР, функции пакета Анализа данных	модули Описательной статистики, Непараметрический анализ	процедуры связи количественных переменных CORRELATIONS и неколичественных переменных CROSSTABS	функция Корреляционный анализ

Функции и методы	Пакеты прикладных программ			
	MS Excel	STATISTICA	SPSS	Deductor
Одномерный и двухмерный дисперсионный анализ	функции пакета Анализ данных	модуль Дисперсионный анализ	процедура ANOVA	-
Кластерный анализ	-	модуль Кластерный анализ	процедуры CLUSTER, QUICK CLUSTER или команда k-means	функции Дерево решений и Карта Кохонена
Факторный анализ	-	модуль Факторный анализ	процедура FACTOR	функция Факторный анализ
Дискриминантный анализ	-	модуль Дискриминантный функциональный анализ	-	-
Многомерное шкалирование	-	модуль Многомерное шкалирование	процедура Multidimentional scaling	-
Возможности графического отображения результатов	встроенные функции Мастер диаграмм	графики типа 2M, 3M, пиктограммы	графики, дендрограммы в процедуре PLOT DENDROGRAM	диаграммы, гистограммы, OLAP – многомерное представление данных в виде кросс-таблиц и кросс-диаграмм
Возможности импорта данных	из других приложений MS Office	из других приложений MS Office, в том числе из MS Excel	из других приложений MS Office, в том числе из MS Excel	из других приложений MS Office программой Deductor Studio
Возможности экспорта данных	таблицы и диаграммы в другие приложения MS Office	таблицы и диаграммы в другие приложения MS Office	таблицы и диаграммы в другие приложения MS Office	таблицы и диаграммы в другие приложения MS Office программой Deductor Studio
Возможности интеллектуализации данных	-	дополнительный модуль Нейронные сети	-	методы Мастера обработки: Нейросеть
Очистка и трансформация данных	-	модуль Временные ряды и прогнозирование	-	широкий спектр, в том числе: сглаживание (скользящее окно), очистка от шумов (фильтрация), группировка

1. Хотя пакет MS Excel не является статистическим пакетом, но он входит в MS Office, включает много статистических функций и дает возможность подключить встроенный пакет **Анализа данных**. Поэтому следует рассмотреть его возможности для статистического анализа. Для небольших предприятий, когда не требуется проводить кластеризации данных, а лишь необходимо установить некоторые зависимости, дать статистическое описание исследуемым переменным, данный пакет будет экономически выгодным.

2. Пакет STATISTICA является мощным средством статистического анализа, нашедший применение во многих сферах деятельности. Он включает большое количество методов, реализуемых в отдельных модулях, которые могут запускаться автономно. Но для реализации каждого метода не хватает методики их выполнения и толкований полученных результатов. Этот недостаток может затруднить внедрение пакета.

3. Пакет ППП SPSS включает широкий спектр команд и процедур, связанных с описательными методами статистики: описание распределения, анализ связи количественных и качественных переменных, наряду с параметрическими методами сравнения средних, большой набор непараметрических тестов. Такая обработка актуальна в ходе социологических исследованиях. Имеется возможность работать с данными, подготовленными в MS Excel.

4. Пакет Deductor имеет единое хранилище данных (а не отдельные файлы, как ППП STATISTICA), разработанные сценарии, включающие загрузку данных из хранилища или внешнего источника, восстановление пропущенных значений, установления незначимых факторов, построение моделей. В пакете при открытии файла с данными он проверяется на пропущенные данные, идет их восстановление, поэтому результаты дальнейшей обработки могут немного отличаться от других пакетов.

Microsoft Excel. Понятия и возможности.

Табличный процессор MS Excel (электронные таблицы) – одно из наиболее часто используемых приложений пакета MS Office, мощнейший инструмент, значительно упрощающий рутинную повседневную работу. Основное назначение MS Excel – решение практически любых задач расчетного характера, входные данные которых можно представить в виде таблиц. Применение электронных таблиц упрощает работу с данными и позволяет получать результаты без программирования расчётов. В сочетании же с языком программирования Visual Basic for Application (VBA), табличный процессор MS Excel приобретает универсальный характер и позволяет решить вообще любую задачу, независимо от ее характера.

Особенность электронных таблиц заключается в возможности применения формул для описания связи между значениями различных ячеек. Расчёт по заданным формулам выполняется автоматически. Изменение содержимого какой-либо ячейки приводит к пересчёту значений всех ячеек, которые с ней связаны формульными отношениями и, тем самым, к обновлению всей таблицы в соответствии с изменившимися данными.

Основные возможности электронных таблиц:

- проведение однотипных сложных расчётов над большими наборами данных;
- автоматизация итоговых вычислений;
- решение задач путём подбора значений параметров;
- обработка (статистический анализ) результатов экспериментов;
- проведение поиска оптимальных значений параметров (решение оптимизационных задач);
- подготовка табличных документов;
- построение диаграмм (в том числе и сводных) по имеющимся данным;
- создание и анализ баз данных (списков).

Анализ данных. Использование сценариев.

Данные - сведения:
- полученные путем измерения, наблюдения, логических или арифметических операций;
- представленные в форме, пригодной для постоянного хранения, передачи и (автоматизированной) обработки.

В Excel тип данных – тип значения, хранящегося в ячейке. Когда данные вводятся на рабочий лист, Excel автоматически анализирует их определяет тип данных. Тип данных, присваиваемый ячейке по умолчанию, определяет способ анализа данных, который можно применять к данной ячейке. Например, в большинстве инструментах анализа данных используются именно числовые значения. Если вы попробуете ввести текстовое значении, то программа отреагирует сообщением об ошибке.

Основные типы данных*:*

Текстовый	Числовой	Денежный
Финансовый	Дроби	Процентный
Даты	Время	Формулы

Анализ данных - область информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных.

Анализ данных – сравнение различной информации.

Работа с таблицей не ограничивается простым занесением в нее данных. Таблицы данных являются частью блока задач, который иногда называют инструментами анализа «что-если». Таблица данных представляет собой диапазон ячеек, показывающий, как изменение определенных значений в формулах влияет на результаты этих формул.

Excel представляет широкие возможности для проведения анализа данных, находящихся в списке. К средствам анализа относятся:

- Обработка списка с помощью различных формул и функций;
- Построение диаграмм и использование карт Excel;
- Проверка данных рабочих листов и рабочих книг на наличие ошибок;
- Структуризация рабочих листов;
- Автоматическое подведение итогов (включая мастер частичных сумм);
- Консолидация данных;
- Сводные таблицы;
- Специальные средства анализа выборочных записей и данных – подбор параметра, поиск решения, сценарии и др.

Одно из главных преимуществ анализа данных – предсказание будущих событий на основе сегодняшней информации.

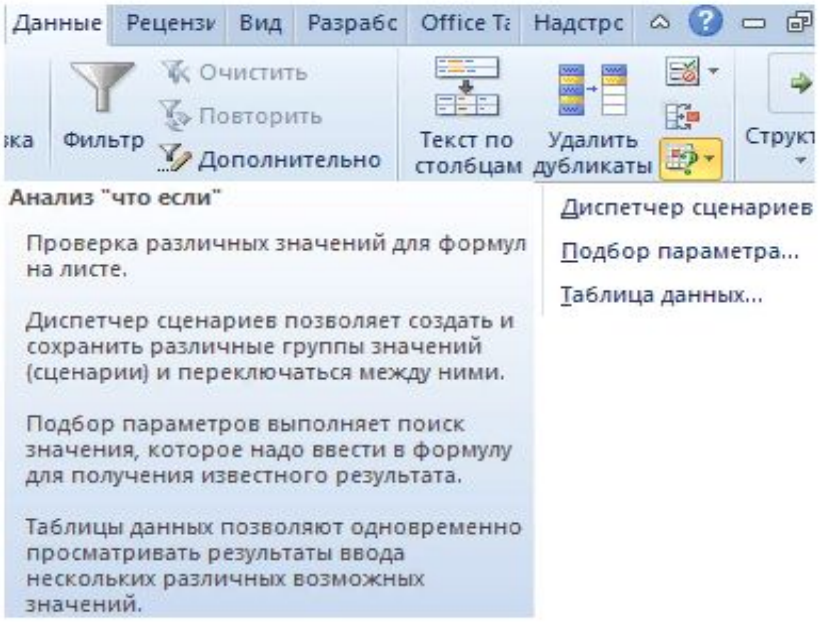
Сценарии являются частью блока задач, который иногда называют инструментами анализа "что-если" (процесс изменения значений ячеек и анализа влияния этих изменений на результат вычисления формул на листе).

Сценарий — это набор значений, которые в приложении Excel сохраняются и могут автоматически подставляться в лист. Сценарии можно использовать для прогноза результатов моделей расчетов листа. Существует возможность создать и сохранить в листе различные группы значений, а затем переключаться на любой из этих новых сценариев, чтобы просматривать различные результаты.

При разработке сценария данные на листе будут меняться. По этой причине перед началом работы со сценарием придется создать сценарий, сохраняющий первоначальные данные, или же создать копию листа Excel.

Инструменты анализа Excel.

Одним из самых привлекательных анализов данных является «Что-если». Он находится: «Данные»-«Работа с данными»-«Что-если».

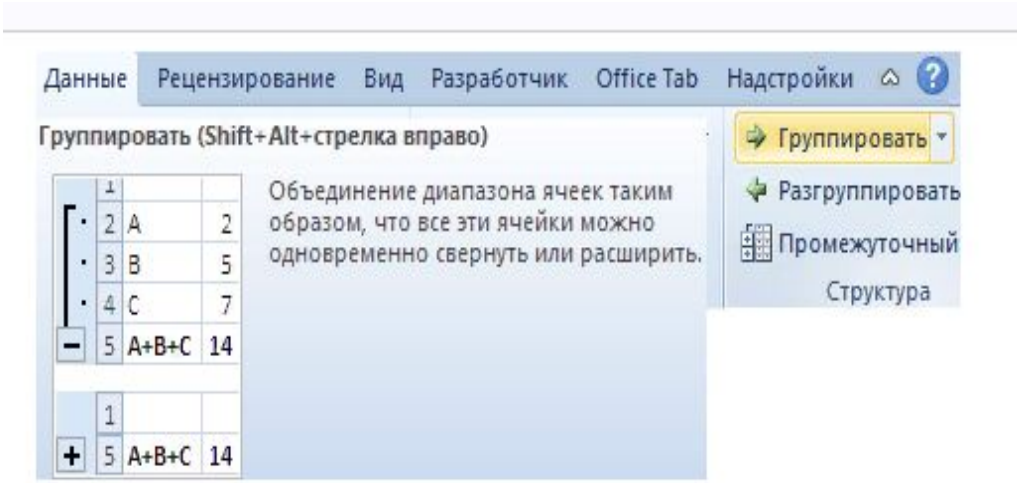


Средства анализа «Что-если»:

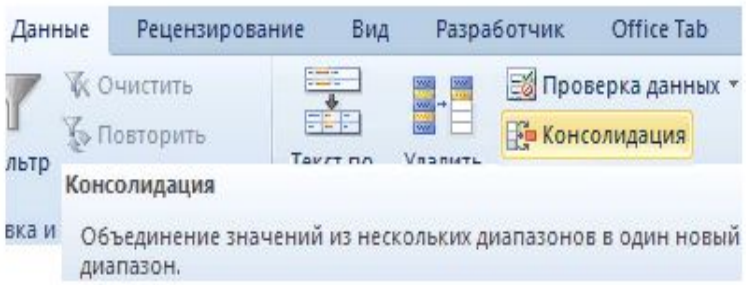
1. «Подбор параметра». Применяется, когда пользователю известен результат формулы, но неизвестны входные данные для этого результата.
2. «Таблица данных». Используется в ситуациях, когда нужно показать в виде таблицы влияние переменных значений на формулы.
3. «Диспетчер сценариев». Применяется для формирования, изменения и сохранения разных наборов входных данных и итогов вычислений по группе формул.
4. «Поиск решения». Это надстройка программы Excel. Помогает найти наилучшее решение определенной задачи.

Другие инструменты анализа Excel.

- Группировка данных:

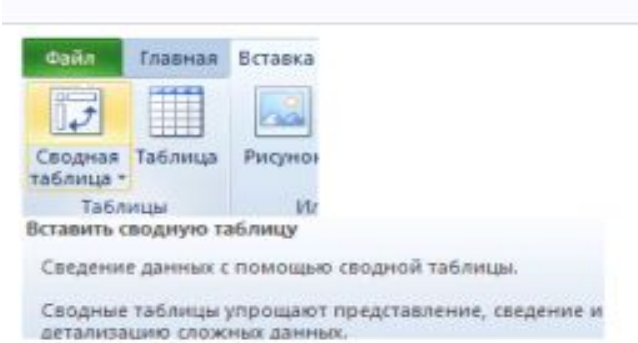


- консолидация данных (объединение нескольких наборов данных):

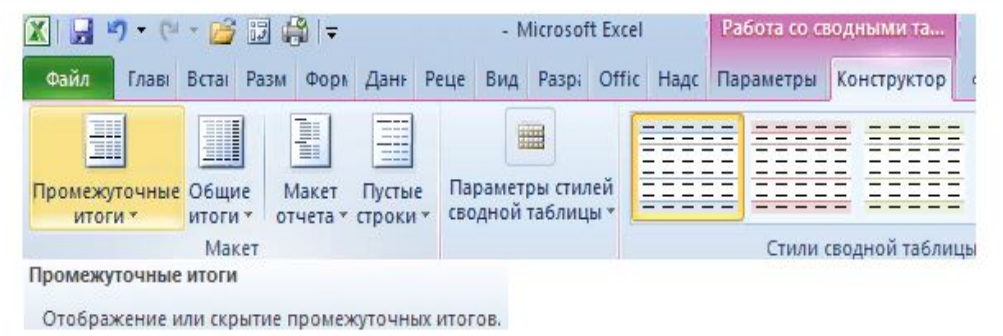


- сортировка и фильтрация (изменение порядка строк по заданному параметру):

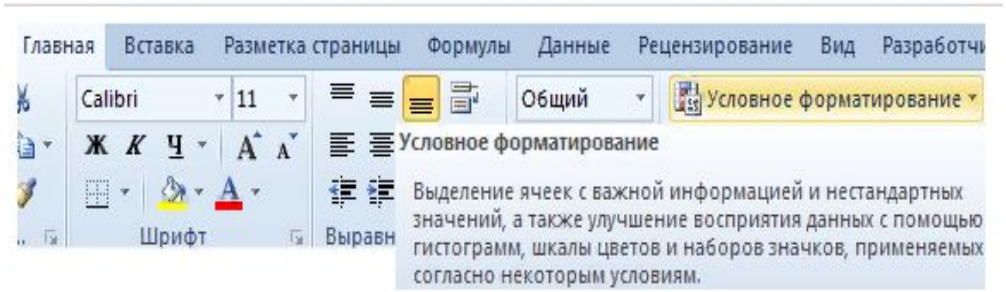
- работа со сводными таблицами:



- получение промежуточных итогов (часто требуется при работе со списками):



- условное форматирование:



- работа с графиками и диаграммами: