

Министерство науки и высшего образования
Российской Федерации
Московский государственный университет
геодезии и картографии
(МИИГАиК)

А.А. Кудлаев

Учебное пособие

**Технологии обработки информации средствами
электронных таблиц. Математическая статистика**



**Москва
2022**

Рецензенты:

Доцент, к.т.н. Н.К. Шавенько (МИИГАиК).

Доцент кафедры экономики предприятия и инноватики Донецкого
национального технического университета, к.э.н. Бондарева И.А.

Составитель: А. А. Кудлаев

Технологии обработки информации средствами электронных таблиц.
Математическая статистика. М. МИИГАиК, 2022.-59 стр.

Изложенный материал соответствует программе по дисциплине «Технологии обработки информации» для технических и экономических специальностей ВУЗов. Теоретический материал сопровождается примерами, нацеленными на развитие у студентов навыков решения практических задач, а также вопросами и упражнениями для самостоятельного выполнения.

Предназначено для студентов Высших учебных заведений, обучающихся по направлениям «Информационные системы и технологии», «Прикладная информатика», «Информационная безопасность». Также может быть рекомендовано для студентов, обучающихся по направлениям «Геодезия и дистанционное зондирование». «Картография и геоинформатика», «Землеустройство и кадастр»

ОГЛАВЛЕНИЕ

Введение. Группы статистических пакетов	4
1. Выборочный метод. Лабораторная работа №1	16
2. Выборочные характеристики. Лабораторная работа №2	21
3. Проверка статистических гипотез. Лабораторная работа №3	27
4. Проверка статистических гипотез. Лабораторная работа №4	33
5. Корреляционный анализ. Лабораторная работа №5	45
6. Регрессионный анализ. Лабораторная работа №6	49
7. Дисперсионный анализ. Лабораторная работа №7	56
 Литература	 59

Введение

Понятие математической статистики

Математическая статистика - раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей. Математическая статистика имеет дело с массовыми явлениями. Она связана с теорией вероятностей и базируется на ее математическом аппарате.

Целью статистического исследования является обнаружение и исследование соотношений между статистическими данными, и их использование для изучения, прогнозирования и принятия решений.

Статистические данные представляют собой данные, полученные в результате обследования большого числа объектов или явлений.

Математическая статистика подразделяется на две основные области:

1. описательную статистику,
2. аналитическую статистику.

Описательная статистика охватывает методы описания статистических данных, представления их в форме таблиц, распределений и т. п.

Аналитическая статистика ориентирована на обработку данных, полученных в ходе эксперимента, и формулирование выводов, имеющих прикладное значение.

Группы статистических пакетов

Потребность в средствах статистического анализа данных в различных областях деятельности, особенно в науке, велика, что и послужило причиной развития рынка компьютерных программ для статистической обработки данных. За последние 30 лет активное развитие получили компьютерные программы, позволяющие проводить статистический анализ больших объемов данных с целью выявления закономерностей, построения прогнозов развития событий, обнаружения связей между различными явлениями и процессами.

На сегодняшний день на рынке представлено большое количество компьютерных программ для статистической обработки данных (статистические пакеты). Разнообразие статистических пакетов обусловлено многообразием задач обработки данных с применением различных типов статистических процедур анализа.

Перед пользователями различных категорий встает вопрос выбора оптимального статистического пакета. Очевидно, что оптимальным является вариант, сочетающий в себе необходимые функциональные возможности, удобство пользования и умеренную цену. При выборе пакета учитываются следующие параметры:

- соответствие характеру решаемых задач;
- объем обрабатываемых данных;
- требования, предъявляемые к квалификации пользователя;
- имеющееся в наличии компьютерное оборудование.

Статистические пакеты по признаку функциональности могут быть разделены на 3 основные группы.

1. *Универсальные пакеты, или пакеты общего назначения* (SPSS, Stata, STATISTICA, Stadia, STATGRAPHICS, Minitab).

Эти пакеты могут применяться для анализа данных из различных областей деятельности. Они предлагают широкий диапазон статистических методов. С такими пакетами рекомендуется работать начинающим пользователям, а также опытным пользователям на начальных этапах работы с данными, когда еще четко не определены статистические методы, которые будут применяться для решения поставленной задачи. Большинство существующих универсальных пакетов имеют много пересечений по составу встроенных статистических процедур.

Универсальный статистический пакет должен удовлетворять требованиям:

- содержать широкий набор стандартных статистических методов;
- быть простым в освоении и использовании непрофессиональным пользователем;
- работать с большими базами данных, отвечать высоким требованиям к вводу, преобразованию и организации хранения данных;
- осуществлять обмен данными с широко распространенными пакетами и базами данных;
- иметь обширный набор средств графического представления данных и результатов их анализа;
- иметь подробное документационное сопровождение и справочную систему.

2. *Профессиональные пакеты* (SAS).

Профессиональные пакеты отличаются от универсальных тем, что позволяют работать со сверхбольшими объемами данных, применять узкоспециализированные методы анализа, создавать собственную систему обработки данных. Подобные пакеты сложны в освоении для непрофессионалов. Подготовленным пользователям работа с профессиональным пакетом предоставит больше возможностей для глубокого и детального анализа данных, построения сложных моделей и адаптации системы к собственным потребностям. Профессиональные пакеты более дорогостоящи, чем универсальные.

3. Специализированные пакеты (BioStat).

В некоторых областях деятельности анализируемые данные настолько специфичны, что к ним следует применять особые методы статистического анализа, как правило, не представленные в универсальных пакетах.

Специализированные пакеты позволяют проводить анализ с использованием ограниченного числа специализированных статистических методов или применимы к использованию для решения вопросов, относящихся к отдельно взятой предметной области. Как правило, с подобными статистическими пакетами работают специалисты, хорошо знакомые с методами анализа данных в той области, на которую ориентирован пакет. Так, статистический пакет BioStat создан для анализа данных в области биологии и медицины. Российский статистический пакет MESOSAUR специализируется на анализе одномерных и многомерных временных рядов и построении регрессионных моделей. Еще один российский статистический пакет DATASCOPE специализируется на проведении анализа многомерных данных.

Большинство представленных на рынке статистических пакетов обладают гибкой модульной структурой, которая может пополняться и расширяться за счет пользовательских модулей, дополнительно закупаемых или находящихся в свободном доступе в Интернете. Подобная гибкость позволяет адаптировать большинство пакетов к потребностям конкретного пользователя.

Статистический пакет должен удовлетворять следующему минимальному набору требований:

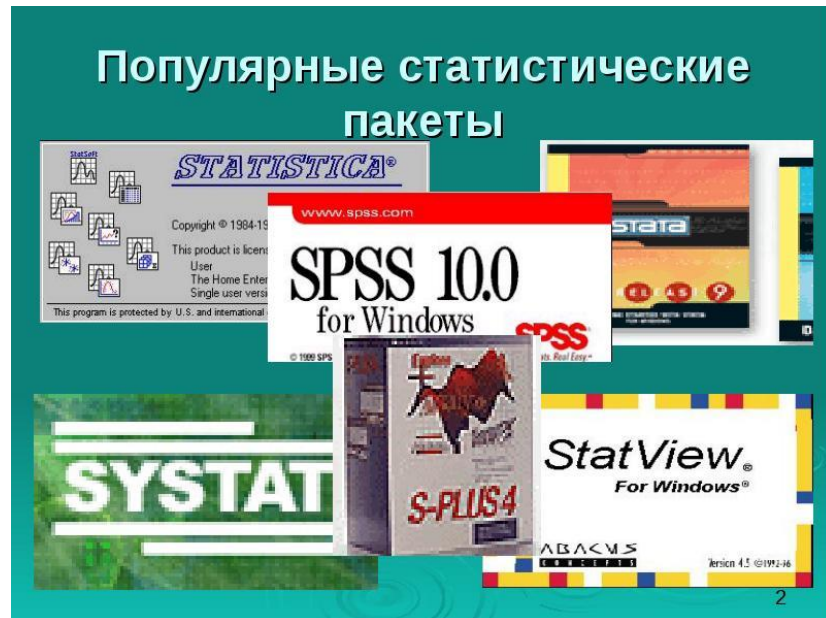
- модульность;
- ассистирование при выборе способа обработки данных;
- использование простого проблемно-ориентированного языка для формулировки задания пользователя;
- автоматическая организация процесса обработки данных;
- ведение банка данных пользователя и составление отчета о результатах проделанного анализа;
- диалоговый режим работы пользователя с пакетом;
- совместимость с другим программным обеспечением.

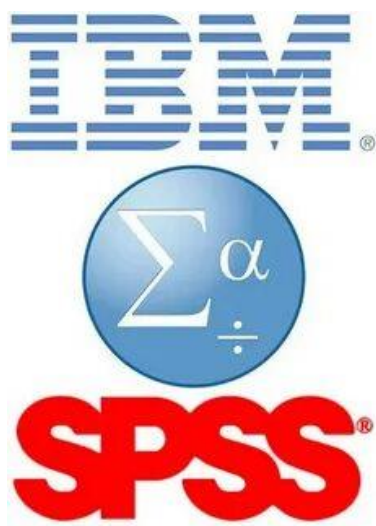
Существует минимальный набор статистических методов анализа, который включен во все рассмотренные пакеты:

- описательная статистика (базовые статистические методы, проверка закона распределения данных);
- дисперсионный анализ;
- непараметрическая статистика (анализ таблиц сопряженности, непараметрические сравнения, дисперсионный анализ);
- контроль качества;
- анализ выживаемости;

- кластерный анализ;
- факторный анализ;
- дискриминантный анализ;
- регрессионный анализ;
- обработка данных (сортировка, отбор, трансформация данных).

Проведем обзор наиболее известных программных продуктов, разработанных для статистической обработки данных.





Пакет **SPSS** (Statistical Package for the Social Sciences) – универсальный статистический пакет компании SPSS. Первая версия пакета была выпущена в 1968 г. В 2009 г. компания IBM поглотила SPSS Inc.

SPSS является одним из лидирующих программных продуктов в области статистического анализа данных для решения вопросов в государственной, академической и бизнессфере.

SPSS является модульной программой. Ее основу составляет базовый модуль (SPSS Base), позволяющий осуществлять управление данными и содержащий наиболее распространенные методы статистического анализа данных:

- проведение описательной статистики;
- построение линейных и нелинейных моделей;
- осуществление преобразования данных;
- проведение факторного, кластерного, дисперсионного анализов;
- вычисление корреляций;
- построение графиков; подготовка отчетов и пр.

Для проведения расширенного и углубленного анализа данных могут быть установлены дополнительные модули пакета.

Достоинства SPSS:

- универсальность;
- широкий набор статистических и графических процедур анализа данных, а также процедур создания отчетов;
- высокая скорость вычислений, простой и удобный интерфейс;
- детальная контекстно-ориентированная справочная система;
- возможность свободного скачивания демонстрационной версии продукта на официальном сайте компании, наличие версий продукта на различных языках;
- совместимость с операционными системами Windows, Mac, Linux;
- наличие значительного количества литературы по работе с пакетом.
- импорт данных из Excel и SAS;
- экспорт результатов в MS Office, PDF; сохранение результатов в формате HTML;
- одновременная работа с несколькими наборами данных;
- построение диаграммы для переменных с множественными ответами;
- построение диаграммы с двумя осями Y.



Пакет **STATA** – универсальный статистический пакет компании StataCorp. Первая версия пакета была выпущена в 1985 г. Пакет STATA является инструментом анализа, предназначенного для специалистов, занимающихся научными исследованиями.

STATA предоставляет следующие возможности статистического анализа данных:

- динамические факторные модели, линейные, обобщенные линейные модели и нелинейные модели, многоуровневые смешанные модели, модели выбытия;
- обобщенные оценки уравнений (GEE);
- непараметрические методы, методы повторных выборок и статистического эксперимента;
- проверка гипотез и работа с оцененными моделями;
- максимизация функций правдоподобия, заданных пользователем;
- матричные команды.

Существуют четыре различных решения при установке STATA.

1. Stata/IC представляет стандартную версию пакета.
2. Stata/SE позволяет работать с более обширными базами данных по сравнению со стандартной версией.
3. Stata/MP обладает многопроцессорными возможностями, производительная, работает быстрее, других версий пакета.
4. Small Stata используется в образовательных целях.

Достоинства STATA:

- широкий набор средств статистического анализа данных;
- многообразные возможности по управлению данными;
- возможность программирования последовательности команд;
- возможность написания собственных программных модулей;
- использование координатного указателя;
- создание графики полиграфического качества;
- совместимость с ОС Windows, Macintosh и Unix (включая Linux);
- интерактивный справочник;
- полное документационное сопровождение ;
- экспорт результатов в MS Office и SAS;
- для работы программы требуется лишь 512 МВ оперативной памяти, сама программа занимает 250 МВ на жестком диске компьютера;



Пакет **Minitab** – универсальный статистический пакет, разработанный Государственным университетом Пенсильвании в 1972 году. Выпускается компанией Minitab Inc. Последняя версия Minitab представляет собой статистический пакет для аналитической

работы на современных предприятиях.

Программа предоставляет следующие возможности статистического анализа данных:

- управление процессом статистической обработки данных;
- оценка мощности и объема выборки;
- планирование экспериментов;
- матричные функции;
- анализ измерительных систем, анализ надежности/выживаемости, анализ временных рядов и прогнозирование, многомерный анализ.

Достоинства Minitab:

- система консультационной поддержки пользователя (интерактивное дерево принятия решений позволяет отыскать правильный подход к анализу данных для поиска ответа на поставленные вопросы), настройка меню и панелей инструментов для быстрого доступа к используемым методам;
- интерпретация полученных результатов;
- осуществление импорта данных из текстовых и табличных процессоров, HTML-файлов, сохранение результатов анализа в форматах TIFF, JPEG, PNG, BMP, GIF, EMF;
- удобный инструмент для размещения нескольких графиков на одной странице, автоматическое обновление графиков при изменении исходных данных, создание и вращение трехмерных графиков;
- автоматизация заданий и создание новых функций с помощью языка макропрограммирования;
- для работы программы требуется лишь 512 МВ оперативной памяти, сама программа занимает 160 МВ на жестком диске компьютера;
- работа с файлами, содержащими до 4000 переменных и неограниченное число наблюдений.

Основной недостаток Minitab – несовместимость с операционными системами, отличными от Windows.



Пакет **STATISTICA** – универсальный статистический пакет компании StatSoft Inc. Первая версия пакета (для DOS) была выпущена в 1991 г. Стандартную комплектацию пакета STATISTICA составляют три модуля, которые могут приобретаться как единым пакетом, так и отдельно.

1. Базовый пакет STATISTICA Base предоставляет обширные возможности выбора основных типов статистического анализа.
2. Модуль *Линейные и нелинейные модели* содержит большой набор инструментов для моделирования и прогнозирования.
3. Модуль *Многомерные разведочные технологии анализа* служит для применения разведочного анализа различных типов данных в сочетании с интерактивными средствами визуализации.

Помимо общих статистических и графических средств в пакете имеются специализированные модули: для проведения социологических или биомедицинских исследований, решения технических и промышленных задач, – карты контроля качества, модули анализа процессов и планирования эксперимента.

Возможны различные варианты установки пакета в зависимости от целей и задач пользователя:

- однопользовательская версия;
- сетевая версия для работы в локальных вычислительных сетях;
- версия для применения в вычислительных системах и крупных организациях
- версия для использования в крупных сетях через веббраузер.

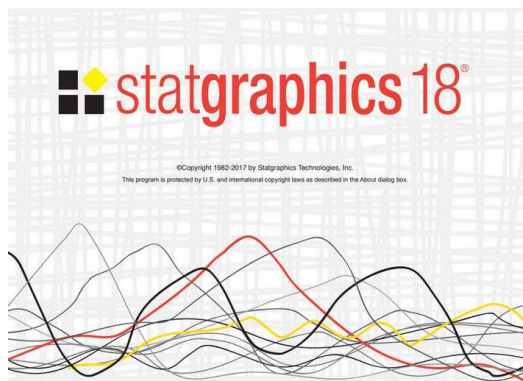
Пакет предоставляет пользователям следующие возможности статистического анализа данных:

- исследование корреляций между переменными;
- диаграмма рассеяния, матричная диаграмма рассеяния;
- быстрые основные статистики и блоковые статистики;
- интерактивный калькулятор вероятностных распределений;
- анализ многомерных откликов, многомерное шкалирование;
- анализ при помощи временных рядов и прогнозирование временных зависимостей, в том числе анализ сезонных колебаний.

Достоинства STATISTICA:

- реализован обмен данными между STATISTICA и Windows-приложениями;
- результаты анализа в виде графиков, таблиц и текста могут быть сохранены в файле с форматом RTF;
- расширяемость системы при помощи программ, созданных на встроенном языке программирования;
- запись макросов для автоматизации выполнения задач;
- возможность обработки больших массивы данных/

Для начинающих пользователей существует версия для обучения основам статистических методов – Student Edition of STATISTICA. Эта версия представляет собой урезанный вариант пакета и позволяет анализировать файлы данных, включающих не более 400 наблюдений.



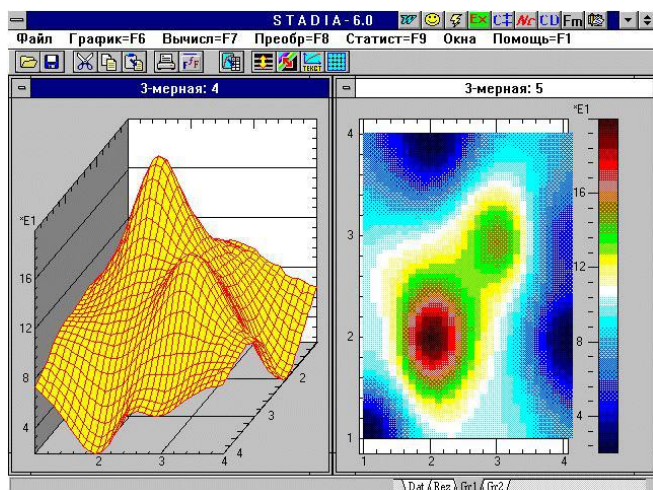
Пакет **STATGRAPHICS** (STATistical GRAPHICS System) – универсальный статистический пакет компании Manugistics. Первая версия пакета была выпущена в середине 80-х годов.

Достоинства STATGRAPHICS:

- сочетание методов обработки разнотипных данных с возможностью создания высококачественной интерактивной графики;
- широкие возможности взаимодействия с другими программными продуктами (электронными таблицами, базами данных);
- высококачественная двумерная и трехмерная графика,

В STATGRAPHICS графика из средства представления результатов анализа превращается в аналитический инструмент: можно идентифицировать точку на графическом отображении и определить ее местонахождение в файле данных; вращать и рассматривать с разных сторон трехмерные изображения; осуществлять разгонку точек на диаграммах рассеяния и т. п.

В STATGRAPHICS существует возможность сохранения результатов работы и создания собственных статистических проектов. После завершения анализа можно сохранить последовательность выбранных методов, параметры статистических процедур, виды графических отображений результатов анализа, табличные формы, комментарии и прочее в отдельном файле. Сохраненную схему анализа можно автоматически применять к другому множеству данных.



Пакет **STADIA** (Statistical Dialogue System) – универсальный статистический пакет, разработанный специалистами Московского государственного университета им. М.В. Ломоносова совместно с НПО «Информатика и компьютеры». Первая версия пакета была выпущена в 1989 году.

Пакет ориентирован на массового пользователя, имеющего небольшой опыт, но нуждающегося

в быстром и удобном средстве оформления и обработки данных.

Пакет предоставляет пользователям широкий набор методов статистического анализа данных: описательная статистика, дисперсионный, корреляционный и спектральный анализ, сглаживание, прогнозирование, простая, нелинейная регрессия, кластерный и факторный анализ, методы контроля качества, анализ и замена пропущенных значений. Также возможно построение и редактирование 2-х, 3-х и многомерной графики: зависимости, прогнозы, диаграммы рассеяния, карты, гистограммы, столбиковые, башенные и круговые диаграммы, установка размеров, надписей по осям и под рисунком и пр.

Достоинства STADIA:

- наличие системы контекстной экранной помощи, включающей объемный гипертекстовый справочник и экспертную систему по выбору метода статистического анализа;
- обработка больших объемов данных (до 32 000 наблюдений);
- наличие режима выдачи оглавления архива данных с комментариями;
- селективный поиск файлов по контексту комментариев, присвоенных архиву с данными;
- экспорт данных и результатов.

Недостатки STADIA:

- несовместимость с операционными системами, отличными от Windows.



Пакет SAS (Statistical Analysis System) — профессиональный статистический пакет, разработанный компанией SAS Institute Inc. Первая версия пакета была выпущена в 1976 году.

Пакет эффективно работает с данными различных типов. Встроенные инструменты статистического анализа могут быть применены для решения широкого круга вопросов, относящихся к различным областям деятельности:

- промышленное производство;
- телекоммуникации;
- правительство (применение статистических выборочных методов для разработки опросов общественного мнения);
- исследование окружающей среды (построение моделей загрязнения воздуха с использованием пространственной статистики);
- биотехнологии;
- розничная торговля (моделирование поведения потребителей).

Пакет предоставляет возможности статистического анализа данных:

- смешанные модели;
- анализ категориальных данных;
- байесовский анализ;
- многомерный анализ;
- анализ выживаемости;
- психометрический анализ;
- анализ пропущенных значений.

Пакет позволяет проводить следующую работу с данными:

- ввод, восстановление, модификация и управление данными;
- написание отчетов и подготовка графиков;
- статистический анализ;
- бизнес планирование, прогнозирование и поддержка принятия решений;
- исследование операций и управление проектами;
- совершенствование качества;
- разработка приложений.

Достоинства SAS:

- быстрая обработка больших объемов данных;
- возможность преобразования математических формул в программный код;
- создание пользовательских модулей;
- получение консультативной помощи в выборе методов анализа и в интерпретации его результатов;
- совместимость с операционными системами Linux и Windows.



Пакет **BioStat** – специализированный статистический пакет компании AnalystSoft. Первая версия пакета была выпущена в середине 80х годов.

Пакет создан для специалистов, работающих в области медицины, биологии и химии. Основан на базе известного пакета StatPlus и включает в себя почти весь его функционал, в том числе регрессионный анализ, анализ выживаемости, дисперсионный анализ. Программа использует электронную таблицу, совместимую с Microsoft Excel как по формату данных, так и по интерфейсу, что делает процесс изучения программы легким и быстрым.

Достоинства BioStat:

- экспорт/импорт данных в MS Excel;
- требуемая память на жестком диске компьютера 32 MB;
- проверка правописания;
- совместимость с операционными системами Windows и Mac.

Далее в учебно-методическом пособии описывается применение пакета MS EXCEL для статистической обработки данных. И хотя EXCEL уступает специализированным статистическим пакетам обработки данных, тем не менее, этот раздел математики представлен в EXCEL достаточно полно. В него включены часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы, позволяющие проводить статистический анализ различных типов данных.

Важно отметить, что использование пакета MS EXCEL и его программных средств разработки позволяет студентам глубже овладеть содержанием учебного курса «Теория вероятностей и математическая статистика» и получить дополнительные практические навыки по курсу «Технология программирования».

Выборочный метод

По охвату статистической совокупности исследование может быть сплошным или выборочным. Множество всех единиц наблюдения, охватываемых сплошным наблюдением, называется *генеральной совокупностью*.

Выборочный метод применяется в том случае, если исследуемая совокупность слишком многочисленна, либо ее элементы малодоступны. Тогда прибегают к изучению какой-то части этой совокупности - выборки.

Выборка - это группа элементов, выбранная для исследования из всей совокупности элементов. Задача выборочного метода состоит в том, чтобы сделать правильные выводы относительно всей совокупности объектов.

Необходимо стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, то есть была бы репрезентативной.

Для получения репрезентативной выборки необходимо четко определять, что понимается под генеральной совокупностью. Например, если хотят получить данные о поступающих во все вузы города, то абитуриенты данного института есть выборка из более широкой генеральной совокупности - всех абитуриентов вузов города.

В тех случаях, когда генеральная совокупность недостаточно известна, для получения представительной выборки применяется случайный выбор.

Выборочная функция распределения

Характеристики случайной величины определяются ее законом распределения $F(x)$. Для практических задач закон распределения обычно неизвестен, или известен с точностью до некоторых неизвестных параметров. В частности, часто невозможно рассчитать точное значение соответствующих вероятностей, так как нельзя определить количество как общего числа исходов, так и число благоприятных исходов. Поэтому вводится *статистическое определение вероятности*. По этому определению вероятность равна отношению числа испытаний (m), в которых событие появилось, к общему количеству произведенных испытаний (n). Такая вероятность называется *статистической частотой*.

В результате на практике сведения о законе распределения случайной величины получают независимыми многократными повторениями опыта, в котором измеряются значения случайной величины. На основе информации из полученной выборки можно построить приблизительные значения для функции распределения и других характеристик случайной величины.

Связь между эмпирической функцией распределения $F_n(x)$ и теоретической функцией распределения $F(x)$ такая же, как связь между частотой события и его вероятностью: функция $F_n(x) \rightarrow F(x)$ при $n \rightarrow \infty$.

Для построения выборочной функции распределения весь диапазон изменения случайной величины X разбивают на ряд интервалов одинаковой ширины. Число интервалов обычно выбирают не менее 5 и не более 15. Затем

определяют число значений случайной величины X , попавших в каждый интервал. Поделив эти числа на общее количество наблюдений n , находят относительную частоту попадания случайной величины X в заданные интервалы. По найденным относительным частотам строят гистограмму выборочной функции распределения.

При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график функции плотности распределения.

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

- Функция *ЧАСТОТА* вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив цифр.
- Процедура *Гистограмма* используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений. Процедура выводит результаты в виде таблицы и гистограммы.

Лабораторная работа №1

Пример 1. Используя средства MS Excel постройте эмпирическое распределение веса для следующей выборки: 64, 57, 63, 62, 58, 61, 63, 60, 60, 61, 65, 62, 62, 60, 64, 61, 59, 59, 63, 61, 62, 58, 58, 63, 61, 59, 62, 60, 60, 58, 61, 60, 63, 63, 58, 60, 59, 60, 59, 61, 62, 62, 63, 57, 61, 58, 60, 64, 60, 59, 61, 64, 62, 59, 65.

Решение:

1. В ячейку *A1* введите слово *Наблюдения*, в диапазон *A2:E12* - значения веса.
2. Выберите ширину интервала 1 кг. Получите 9 интервалов. В ячейки *G1* и *G2* введите названия интервалов: *Вес* и *кг*. В диапазон *G4:G12* введите граничные значения интервалов (57, 58, 59, 60, 61, 62, 63, 64, 65).
3. Введите заголовки создаваемой таблицы: *Абсолютные частоты*, *Относительные частоты*, *Накопленные частоты*.

	A	B	C	D	E	F	G	H	I	J
1	Наблюдения						Вес	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	64	62	58	63	61		кг			
3	57	62	63	58	58					
4	63	60	61	60	60		57			
5	62	64	59	59	64		58			
6	58	61	62	60	60		59			
7	61	59	60	59	59		60			
8	63	59	60	61	61		61			
9	60	63	58	62	64		62			
10	60	61	61	62	62		63			
11	61	62	60	63	59		64			
12	65	58	63	57	65		65			
13										

Рис. 1.1. Ввод исходных данных

4. Заполните столбец абсолютных частот с помощью *Мастера функций* (f_x). В диалоговом окне *Мастер функций* выберите категорию *Статистические* и функцию *ЧАСТОТА*. В рабочее поле *Массив данных* введите диапазон данных наблюдений. В рабочее поле *Массив интервалов* введите диапазон интервалов. Последовательно нажмем комбинацию клавиш *Ctrl+Shift+Enter*, В столбце *H* появится массив абсолютных частот.

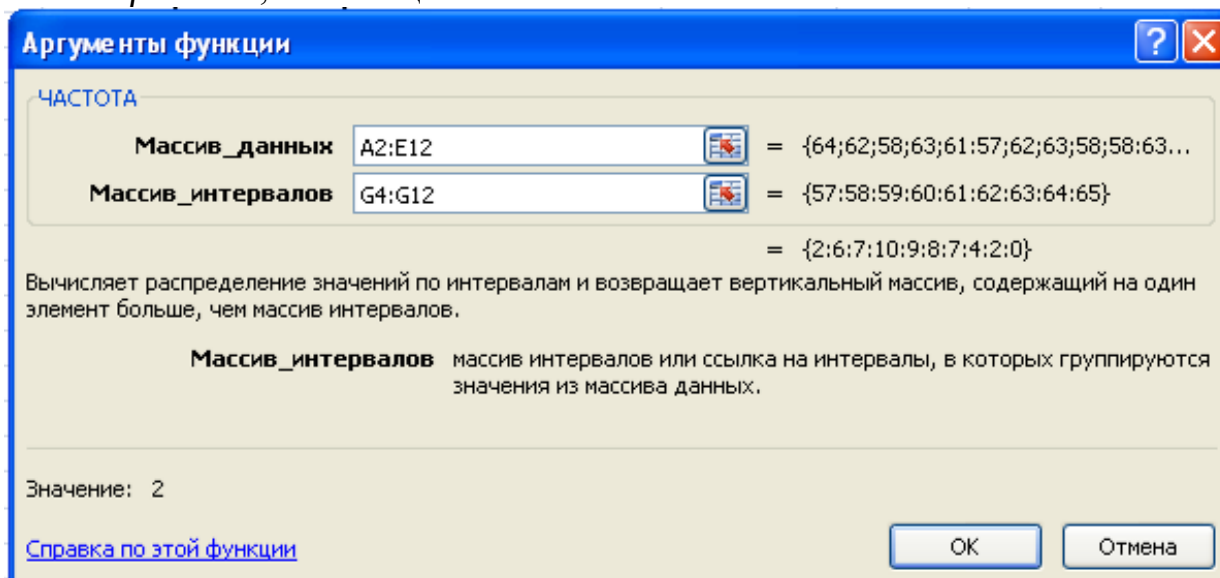


Рис. 1.2. Диалоговое окно функция *ЧАСТОТА*

5. В ячейке *H13* найдите общее количество наблюдений.

6. Заполните столбец относительных частот. В ячейку *I4* введите формулу для вычисления относительной частоты: $=H4/H\$13$. Скопируйте формулу на весь диапазон. Получите массив относительных частот.

7. Заполните столбец накопленных частот. В ячейку *J4* скопируйте значение относительной частоты из ячейки *I4*. В ячейку *J5* введем формулу: $=J4 + 15$. Скопируйте введенную формулу на весь диапазон. Получите массив накопленных частот.

	A	B	C	D	E	F	G	H	I	J
1	Наблюдения						Вес	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	64	62	58	63	61		кг			
3	57	62	63	58	58					
4	63	60	61	60	60		57	2	0,036	0,036
5	62	64	59	59	64		58	6	0,109	0,145
6	58	61	62	60	60		59	7	0,127	0,273
7	61	59	60	59	59		60	10	0,182	0,455
8	63	59	60	61	61		61	9	0,164	0,618
9	60	63	58	62	64		62	8	0,145	0,764
10	60	61	61	62	62		63	7	0,127	0,891
11	61	62	60	63	59		64	4	0,073	0,964
12	65	58	63	57	65		65	2	0,036	1,000
13								55		

Рис. 1.3. Результат вычислений относительных и накопленных частот

8. Постройте диаграмму относительных и накопленных частот.

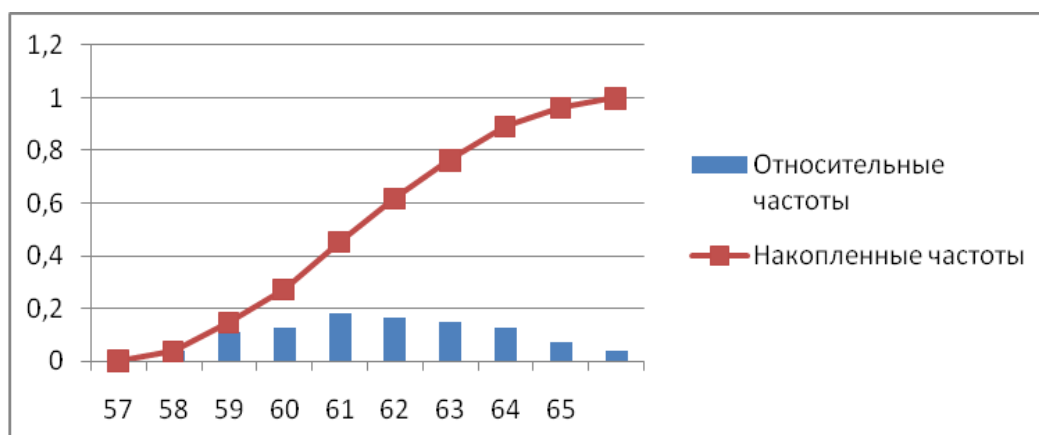


Рис. 1.4. Диаграмма относительных и накопленных частот

Накопленные частоты — это есть значения функции распределения.

Относительные частоты — это есть значения функции плотности распределения.

Пример 2. Для данных из примера 1. постройте эмпирические распределения, воспользовавшись процедурой *Гистограмма*.

Решение:

1. Процедуру *Гистограмма* выберите из меню *Данные* подпункт *Анализ данных*. В поле *Инструменты анализа* укажите процедуру *Гистограмма*.

2. В появившемся окне *Гистограмма* заполните рабочие поля:

- во *Входной диапазон* введите диапазон исследуемых данных,
- в *Выходной диапазон* - ссылку на левую верхнюю ячейку выходного диапазона,
- Установите переключатели в положение *Интегральный процент* и *Вывод графика*;

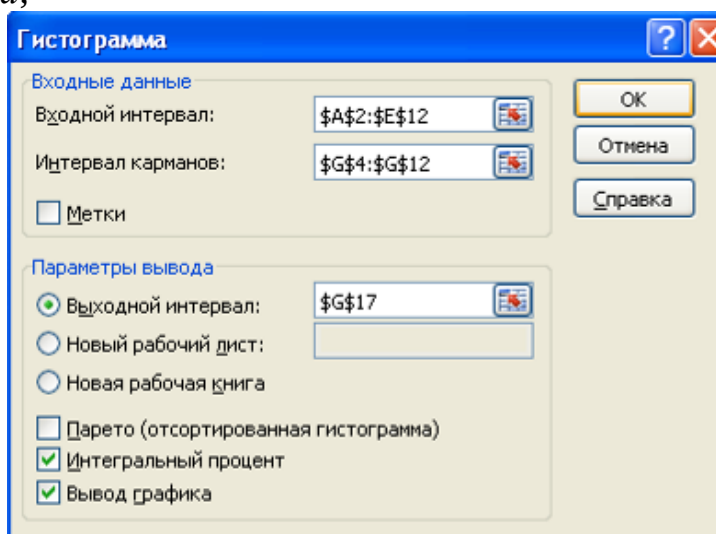


Рис. 1.5. Заполнение диалогового окна *Гистограмма*

3. Нажмите кнопку *OK*.

4. В результате появится таблица и диаграмма:

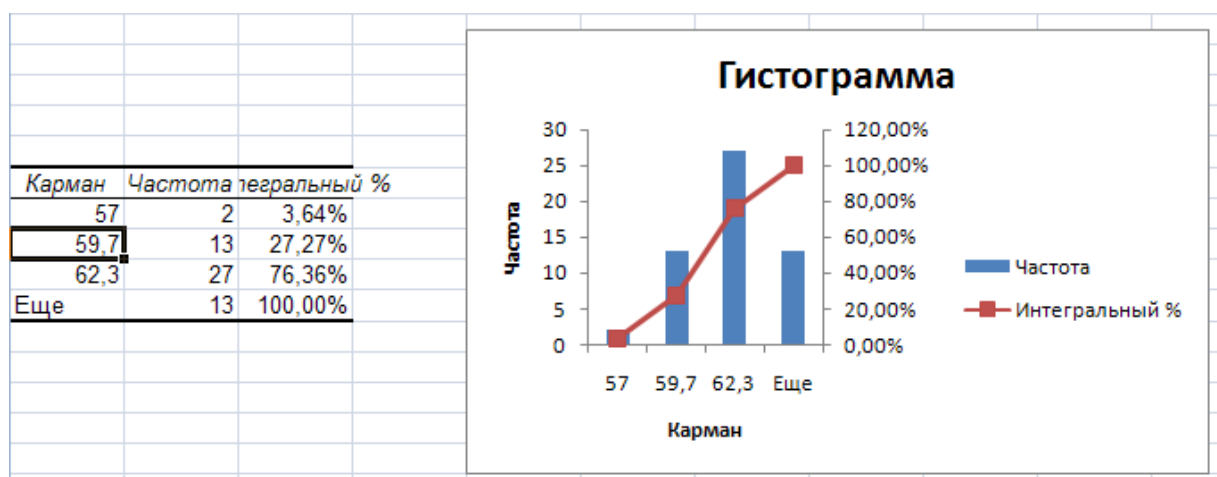


Рис. 1.6. Таблица и диаграмма примера 2

Диаграмма на рис. 1.6. отличается от диаграммы на рис. 1.4. Это объясняется тем, что диапазон карманов не был введен. Количество и границы интервалов определялись в процедуре *Гистограмма* автоматически.

Задания:

1. Постройте эмпирические функции распределения (относительные и накопленные частоты) выборки из 20 величин; 181, 169, 178, 178, 171, 179, 172, 181, 179, 168, 174, 167, 169, 171, 179, 181, 181, 183, 172, 176.

2. Найдите распределение по абсолютным частотам для следующих результатов тестирования в баллах: 79, 85, 78, 85, 83, 81, 95, 88 и 97 (используйте границы интервалов 70, 79, 89).

3. Постройте эмпирические функции распределения (абсолютные и накопленные частоты) успеваемости в группе из 20 студентов: 4, 4, 5, 3, 4, 5, 4, 5, 3, 5, 3, 3, 5, 4, 5, 4, 3, 5, 3, 5.

Выборочные характеристики

Замена теоретической функции распределения $F(x)$ на ее выборочный аналог $F_n(x)$ в определении математического ожидания, дисперсии, стандартного отклонения и т. п. приводят к выборочному среднему, выборочной дисперсии, выборочному стандартному отклонению и т. д. Выборочные характеристики являются оценками соответствующих характеристик генеральной совокупности. Эти оценки должны удовлетворять определенным требованиям. Оценки должны быть:

- несмещенными - стремиться к истинному значению характеристики генеральной совокупности при увеличении количества испытаний;
- состоятельными - с ростом размера выборки оценка должна стремиться к значению соответствующего параметра генеральной совокупности с вероятностью, приближающейся к 1;
- эффективными - для различных выборок равного объема оценка должна иметь минимальную дисперсию.

В результате наблюдений или эксперимента получают наборы данных, называемые выборками. Для проведения их анализа данные подвергаются статистической обработке. Первое, что всегда делается при обработке данных, это вычисление элементарных статистических характеристик выборок (среднего, среднеквадратичного отклонения, ошибки среднего) по каждому параметру и по каждой группе. Полезно также вычислить эти характеристики для объединения родственных групп и суммарно по всем данным.

Среди выборочных характеристик выделяют показатели, относящиеся к центру распределения, рассеяния и формы распределения.

Лабораторная работа №2

Показатели, характеризующие центр распределения

В мастере функций *Excel* имеется ряд функций, предназначенных для вычисления выборочных характеристик центра распределения.

- Функция *СРЗНАЧ* вычисляет среднее арифметическое значение своих аргументов. Если выборка содержит числа {10;14;5;6;10;12;13}, то $СРЗНАЧ(10;14;5;6;10;12;13) = 10$.

Средним значением выборки, или выборочным аналогом математического ожидания, является величина:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

При увеличении числа наблюдений среднее арифметическое значение приближается к математическому ожиданию.

- Функция *СРГАРМ* позволяет получить среднее гармоническое множества данных. *СРГАРМ(10;14;5;6;10;12;13)* равняется 8,317.

Среднее гармоническое - величина, обратная к среднему арифметическому обратных величин.

$$H_n = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

- Функция *СРГЕОМ* вычисляет среднее геометрическое значений массива положительных чисел. Функцию *СРГЕОМ* можно использовать для вычисления средних показателей динамического ряда. *СРГЕОМ(10;14;5;6;10;12;13)* равняется 9,414.

Средняя геометрическая

$$x_{\text{геом}} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n}$$

- Функция *МЕДИАНА* позволяет получать медиану заданной выборки. *МЕДИАНА(10;14;5;6;10;12;13)* равняется 10.

Медиана - элемент выборки, число элементов выборки со значениями больше которого и меньше которого равно.

Для нахождения медианы обычно выборку ранжируют - располагают элементы в порядке возрастания. Если количество членов ранжированного ряда нечетное, медианой является значение ряда, которое расположено посередине, то есть элемент с номером $(n + 1)/2$. Если число членов ряда четное, то медиана равна среднему членов ряда с номерами $n/2$ и $n/2 + 1$.

- Функция *МОДА* позволяет получать моду заданной выборки. *МОДА(10;14;5;6;10;12;13)* равняется 10.

Мода - элемент выборки с наиболее часто встречающимся значением.

Показатели рассеяния вариант

Функции *Excel*, предназначенные для вычисления величин рассеяния элементов относительно центра распределения.

Интервал (вариационный размах) - это разница между максимальным и минимальным значениями элементов выборки. Интервал является простейшей мерой рассеяния элементов в выборке. Более точно отражают рассеяние показатели, учитывающие не только крайние, но и все значения элементов выборки.

- Функция *ДИСП* позволяет оценить дисперсию по выборочным данным. Например, *ДИСП*(10;14;5;6;10;12;13) равняется 11,667.

Дисперсия выборки - параметр, характеризующий степень разброса элементов выборки относительно среднего значения.

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

- Функция *СТАНДОТКЛОН* вычисляет стандартное отклонение (среднее квадратическое отклонение).

Например, *СТАНДОТКЛОН*(10;14;5;6;10;12;13) равняется 3,416.

Среднее квадратичное отклонение

$$\sigma = \sqrt{D}$$

Это параметр характеризует степень разброса элементов выборки относительно среднего значения. Параметр используется в тех случаях, когда необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и среднее значение этой случайной величины. Часто выборочное стандартное отклонение обозначают буквой σ (сигма).

Стандартная ошибка :

$$m = \frac{\sigma}{\sqrt{n}}$$

Стандартная ошибка - параметр, характеризующий степень возможного отклонения среднего значения, полученного на исследуемой выборке, от истинного среднего значения всей совокупности элементов. С помощью стандартной ошибки задается доверительный интервал. 95%-ный доверительный интервал, равный $\bar{x} \pm 2m$, обозначает диапазон, в который с вероятностью $p = 0,95$ при достаточно большом числе наблюдений ($n > 30$) попадает среднее генеральной совокупности MX .

- Функция *ПЕРСЕНТИЛЬ* позволяет получить квантили заданной выборки. Квантилью со значением 0,1 чисел {10; 14; 5; 6; 10; 12; 13}, является *ПЕРСЕНТИЛЬ*(10; 14; 5; 6; 10; 12; 13;0,1) = 5,6.

Выборочной квантилью называется решение уравнения [1]

$$F_n(x) = p$$

В частности, выборочная медиана есть решение уравнения [1]

$$F_n(x) = 0,5$$

Показатели, характеризующие форму распределения

Асимметрия - величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает значения от -1 до 1. В случае симметричного распределения асимметрия равна 0.

$$S = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\sigma^3}$$

Экссесс - это степень выраженности «хвостов» распределения, то есть частоты появления удаленных от среднего значений.

$$E = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\sigma^4} - 3$$

Форму эмпирического распределения позволяют оценить специальные функции *СКОС* и *ЭКССЕСС*.

- Функция *СКОС* позволяет оценить асимметрию выборочного распределения. Например, *СКОС*(10;14;5;6;10;12;13) равняется -0,527.
- Функция *ЭКССЕСС* вычисляет оценку эксцесса по выборочным данным. Например, *ЭКССЕСС*(10;14;5;6;10;12;13) равняется -1,169.

Значения асимметрии и эксцесса используют для проверки гипотезы о том, что данные (выборка) принадлежат к определенному теоретическому распределению, в частности, нормальному распределению. Для нормального распределения асимметрия = 0, а эксцесс = 3.

Задания:

1. Найдите среднее значение и стандартное отклонение для выборки: 12,8; 13,2; 13,0; 12,9; 13,5; 13,1.
2. Найдите выборочные среднее, медиану, моду, дисперсию и стандартное отклонение для выборки: 26, 35, 29, 27, 33, 35, 30, 33, 31, 29.
3. Определите верхнюю (0,75) и нижнюю (0,25) квартили для выборки: 164, 160, 157, 166, 162, 160, 161, 159, 160, 163, 170, 171.
4. Определите выборочные асимметрию и эксцесс для данных из задания 3.

Использование инструментов Пакета анализа

В пакете *MS Excel* имеется набор инструментов для анализа данных, называемый *Пакет анализа*, который может быть использован для решения задач статистической обработки выборочных данных.

Нахождение основных выборочных характеристик

Для определения характеристик выборки используется процедура *Описательная статистика*. Процедура позволяет получить статистический отчет, содержащий информацию о центральной тенденции и изменчивости входных данных.

В результате анализа в выходном диапазоне для каждого столбца данных выводятся следующие статистические характеристики: среднее, стандартная ошибка (среднего), медиана, мода, стандартное отклонение, дисперсия выборки, эксцесс, асимметрия, интервал, минимум, максимум, сумма, счет, наибольшее, наименьшее, уровень надежности.

Пример. Рассматривается стипендия различных специальностей студентов. Были получены следующие данные:
Специальность 1 = {4500, 4000, 3700, 3000, 2500},
Специальность 2 = {2100, 2100, 2000, 2000, 2000, 1900, 1800, 1800},
Специальность 3 = {3200, 3000, 2500, 2000, 1900, 1800}

Необходимо определить основные статистические характеристики в группах данных.

Решение:

1. Для использования инструментов анализа исследуемые данные представьте в виде таблицы (рис. 2.1), где столбцами являются соответствующие показатели:

	А	В	С
1	Специальность 1	Специальность 2	Специальность 3
2	4500	2100	3200
3	4000	2100	3000
4	3700	2000	2500
5	3000	2000	2000
6	2500	2000	1900
7		1900	1800
8		1800	
9		1800	

Рис. 2.1. Таблица данных

2. Проведите статистическую обработку. Для этого в пункте меню *Данные*, выберите команду *Анализ данных*. В списке *Инструменты анализа* выберите строку *Описательная статистика*.

Описательная статистика

Входные данные

Входной интервал:

Группирование: ☒ по столбцам ☐ по строкам

☒ Метки в первой строке

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист

☐ Новая рабочая книга

☒ Итоговая статистика

☐ Уровень надежности: %

☐ К-ый наименьший:

☐ К-ый наибольший:

OK Отмена Справка

Рис. 2.2. Заполнения диалогового окна Описательная статистика

В диалоговом окне (рис. 2.2) в рабочем поле *Входной интервал* укажите входной диапазон данных. Активируйте рабочее поле *Выходной интервал*. Укажите выходной диапазон – адрес верхней левой ячейки. В разделе *Группировка* переключатель установите в положение *по столбцам*. Установите флажок в поле *Итоговая статистика*. Нажмем кнопку *OK*.

В результате анализа (рис. 2.3) в указанном выходном диапазоне для каждого столбца данных будут получены соответствующие результаты.

Е	F	G	Н	I	J
Специальность 1		Специальность 2		Специальность 3	
Среднее	3540	Среднее	1962,5	Среднее	2400
Стандартная ошибка	355,8089375	Стандартная ошибка	41,99277149	Стандартная ошибка	243,5843
Медиана	3700	Медиана	2000	Медиана	2250
Мода	#Н/Д	Мода	2000	Мода	#Н/Д
Стандартное отклонение	795,6129712	Стандартное отклонение	118,7734939	Стандартное отклонение	596,6574
Дисперсия выборки	633000	Дисперсия выборки	14107,14286	Дисперсия выборки	356000
Эксцесс	-1,29384635	Эксцесс	-1,229290178	Эксцесс	-2,06887
Асимметричность	-0,245024547	Асимметричность	-0,394325716	Асимметричность	0,457606
Интервал	2000	Интервал	300	Интервал	1400
Минимум	2500	Минимум	1800	Минимум	1800
Максимум	4500	Максимум	2100	Максимум	3200
Сумма	17700	Сумма	15700	Сумма	14400
Счет	5	Счет	8	Счет	6

Рис. 2.3. Результат работы инструмента Описательная статистика

Задания:

5. Найдите наиболее популярный туристический маршрут из четырех: 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.

6. По полученной выборке: 12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16. Необходимо определить основные выборочные характеристики.

Проверка статистических гипотез

Проверка соответствующих статистических гипотез выполняется с помощью аналитической статистики, которая ориентирована на обработку данных, полученных в ходе эксперимента, с целью формулировки выводов, имеющих прикладное значение: отражают ли наблюдаемые данные объективно существующую реальность.

Принятие статистических решений

Статистическая гипотеза - это предположение о виде или отдельных параметрах распределения вероятностей, которое подлежит проверке на имеющихся данных. Методы проверки гипотез называются критериями.

В большинстве случаев рассматривают нулевую гипотезу H_0 , состоящую в том, что все события произошли случайно. Альтернативная гипотеза H_1 состоит в том, что события случайным образом произойти не могли, имело место воздействие некоего фактора.

Обычно нулевая гипотеза формулируется таким образом, чтобы на основании эксперимента ее можно было отвергнуть с заранее заданной вероятностью ошибки α , называемой уровнем значимости.

Уровень значимости - максимальное значение вероятности появления события, при котором событие считается практически невозможным. В статистике наибольшее распространение получил уровень значимости, равный $\alpha = 0,05$. В наиболее ответственных случаях, уровень значимости принимают равным $\alpha = 0,01$ или даже $\alpha = 0,001$.

Величину P , равную $1 - \alpha$, называют доверительной вероятностью, то есть вероятностью, признанной достаточной для того, чтобы уверенно судить о принятом статистическом решении. Соответственно, в качестве доверительных вероятностей выбирают значения 0,95, 0,99 и 0,999.



Рис. 3.1. 95%-ный доверительный интервал для среднего значения

Интервал, в котором с заданной доверительной вероятностью $P = 1 - \alpha$ находится оцениваемый параметр, называется доверительным интервалом. Граничные точки доверительного интервала называют доверительными пределами (рис. 3.1).

Анализ одной выборки

Важным вопросом, возникающим при анализе выборки, является вопрос: относится та или иная выборка к данной статистической совокупности? Решение вопроса не представляет сложности, если распределение в этой совокупности является нормальным. Для этого используют правило трех сигм. Согласно этому правилу, в пределах $M \pm 3\sigma$ находится 99,7% всех выборок. Поэтому, если выборка попадает в этот интервал, то она считается принадлежащей к данной совокупности.

Другой важной задачей, возникающей при анализе одной выборки, является сравнение выборочного среднего арифметического со средним значением генеральной совокупности. Эта задача решается с помощью статистических критериев. При этом выясняется, значительно ли отличие выборочного среднего значения от среднего значения генеральной совокупности, из которой предположительно взята выборка, или наблюдаемое различие является случайным.

Средние значения, получаемые по выборочным данным, обычно не совпадают с генеральным средним. В связи с этим возникает вопрос: можно ли по результатам выборочной оценки судить о свойствах всей генеральной совокупности?

Поскольку каждую оценку, полученную в отдельной выборке, можно рассматривать как случайную величину, то при увеличении числа выборок распределение отдельных оценок будет принимать вид нормального распределения.

Отсюда следует, что 68,3% всех выборочных средних находятся в пределах $\Delta = M \pm m$, где Δ - предельная ошибка выборки, M - среднее выборочное, m - стандартное отклонение среднего значения.

Здесь 0,683 - доверительная вероятность, $1 - 0,683 = 0,317$ - уровень значимости α , $\Delta = M \pm m$ - 68% доверительный интервал.

Для принятой в большинстве исследований доверительной вероятности 0,95, доверительный интервал для средних при достаточно большом числе наблюдений ($n > 30$) примерно равен $\pm 2m$. При доверительной вероятности 0,99, доверительный интервал составит примерно $\pm 3m$. Для более точного определения границ доверительного интервала можно воспользоваться формулой:

$$\left[M - t_{n,p} \frac{s}{\sqrt{n}}; M + t_{n,p} \frac{s}{\sqrt{n}} \right],$$

где M - среднее значение, s - стандартное отклонение, $t_{n,p}$ - табличное значение распределения Стьюдента с числом степеней свободы n и доверительной вероятностью p , n - количество элементов в выборке.

Лабораторная работа №3

В *MS Excel* для более точного вычисления границ доверительного интервала и при числе элементов в выборке $n < 30$ можно воспользоваться функцией *ДОВЕРИТ* или процедурой *Описательная статистика*.

Функция *ДОВЕРИТ*(альфа; станд_откл; размер) определяет полуширину доверительного интервала и содержит следующие параметры:

- *Альфа* - уровень значимости, используемый для вычисления доверительной вероятности, альфа, равное 0,05, означает 95%-ный уровень доверительной вероятности;
- *Станд_откл* - стандартное отклонение генеральной совокупности для интервала данных, предполагается известным;
- *Размер* - это размер выборки.

Пример 1. Пусть имеется выборка, содержащая числовые значения: 13, 15, 17, 19, 22, 25, 19. Необходимо определить границы 95%-ного доверительного интервала для среднего значения.

Решение:

1. В диапазон *A1:A7* введем исходный ряд чисел.
2. Далее вызовем процедуру *Описательная статистика*: *Данные ► Анализ данных ► Инструменты анализа ► Описательная статистика*.
3. В появившемся диалоговом окне в рабочем поле *Входной интервал*: укажем входной диапазон - *A1:A7*, Переключателем активизируем *Выходной интервал* и укажем выходной диапазон. В разделе *Группировка* переключатель установим в положение по столбцам. Установим флажок в левое поле *Уровень надежности*: и в правом поле (%) - 95. *ОК*.

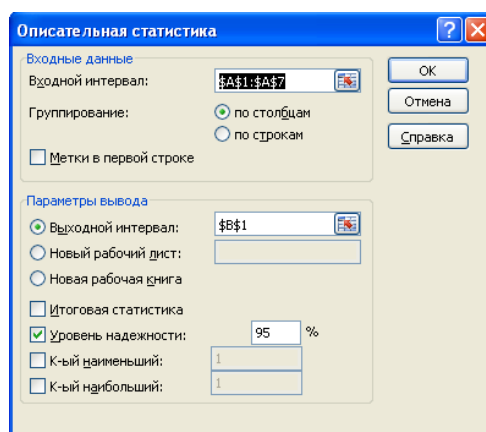


Рис. 3.2. Заполнение диалогового окна *Описательная статистика*

4. В результате анализа в указанном выходном диапазоне для доверительной вероятности 0,95 будет получено значение доверительного интервала (рис. 3.3).

	A	B	C
1	13	Столбец1	
2	15		
3	17	Среднее	18,57143
4	19	Стандартная ошибка	1,540828
5	22	Медиана	19
6	25	Мода	19
7	19	Стандартное отклонение	4,076647
8		Дисперсия выборки	16,61905
9		Эксцесс	-0,36748
10		Асимметричность	0,288456
11		Интервал	12
12		Минимум	13
13		Максимум	25
14		Сумма	130
15		Счет	7
16		Уровень надежности(95,0%)	3,770269
17			

Рис. 3.3. Исходная выборка и результат вычислений

Уровень надежности - это половина доверительного интервала для генерального среднего арифметического. Из полученного результата следует, что с вероятностью 0,95 среднее арифметическое для генеральной совокупности находится в интервале $18,571 \pm 3,77$.

Проверка соответствия теоретическому распределению

В большинстве случаев при решении реальных задач закон распределения и его параметры неизвестны. В то же время применяемые статистические методы в качестве предпосылок часто требуют определенного закона распределения. Наиболее часто проверяется предположение о нормальном распределении генеральной совокупности.

Для оценки соответствия имеющихся экспериментальных данных нормальному закону распределения обычно используют графический метод, выборочные параметры формы распределения и критерии согласия.

Графический метод позволяет давать ориентировочную оценку расхождения или совпадений распределений.

При большом числе наблюдений ($n > 100$) достоверные результаты дает вычисление выборочных параметров формы распределения: эксцесса и асимметрии. Предположение о нормальном законе распределения не противоречит имеющимся данным, если асимметрия близка к нулю, то есть лежит в диапазоне от -0,2 до 0,2, а эксцесс - от 2 до 4.

Достоверные результаты дает использование критериев согласия. Так называют статистические критерии, предназначенные для проверки согласия опытных данных и теоретической модели. Здесь нулевая гипотеза H_0 представляет собой утверждение о том, что распределение генеральной совокупности, из которой получена выборка, не отличается от нормального. Среди критериев согласия большое распространение получил непараметрический критерий χ^2 .

Уверенно о нормальном законе распределения данных можно судить, если имеется не менее 50 наблюдений. В случаях меньшего числа наблюдений можно говорить, что данные не противоречат нормальному закону, в этом случае обычно используют графические методы оценки соответствия.

Использование критерия согласия хи-квадрат

Для применения критерия желательно, чтобы объем выборки $n \geq 50$, выборочные данные были сгруппированы в интервальный ряд с числом равномерных интервалов не менее 7, а в каждом интервале находилось не менее 5 наблюдений.

Как и любой другой статистический критерий, критерий χ^2 не доказывает справедливость нулевой гипотезы, а лишь может позволить ее отвергнуть с определенной вероятностью.

В *MS Excel* критерий χ^2 реализован в функции *ХИ2ТЕСТ*. Функция *ХИ2ТЕСТ* вычисляет вероятность совпадения фактических значений и теоретических значений. Если вычисленная вероятность ниже уровня значимости (0,05), то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют нормальному закону распределения. Если вычисленная вероятность близка к 1, то можно говорить о высокой степени соответствия экспериментальных данных нормальному закону распределения.

Пример 2. Проверить соответствие эмпирических данных из ЛР№1 нормальному закону распределения.

Решение:

1. Повторите пункты 1-7 решения примера 1 ЛР №1. В результате получится таблица (рис. 1.3).

2. Найдите частоты нормального распределения. Для этого предварительно найдите среднее значение (ячейка *E13*) и стандартное отклонение (ячейка *E14*) выборки.

Столбцу *K* дайте название - *Относительные частоты (теоретические)*. С помощью функции *НОРМРАСП* найдите значения вероятности нормальной функции распределения в диапазоне ячеек *K4:K12*

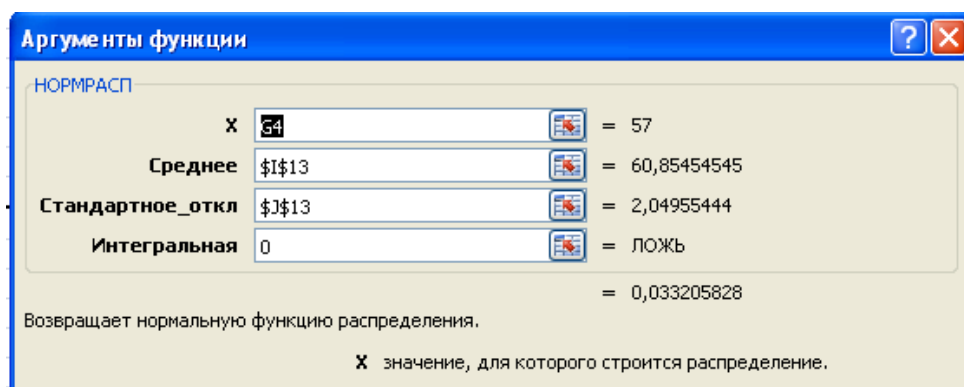


Рис. 3.4. Пример заполнения рабочих полей функции *НОРМРАСП*

	A	B	C	D	E	F	G	H	I	J	K	L
1	Наблюдения											
2	64	62	58	63	61		Вес (кг)	Абсолютные частоты (эмпирические)	Относительные частоты (эмпирические)	Накопленные частоты (эмпирические)	Относительные частоты (теоретические)	Абсолютные частоты (теоретические)
3	57	62	63	58	58							
4	63	60	61	60	60		57	2	0,036	0,036	0,033	1,826
5	62	64	59	59	64		58	6	0,109	0,145	0,074	4,059
6	58	61	62	60	60		59	7	0,127	0,273	0,129	7,109
7	61	59	60	59	59		60	10	0,182	0,455	0,178	9,814
8	63	59	60	61	61		61	9	0,164	0,618	0,194	10,679
9	60	63	58	62	64		62	8	0,145	0,764	0,167	9,158
10	60	61	61	62	62		63	7	0,127	0,891	0,113	6,190
11	61	62	60	63	59		64	4	0,073	0,964	0,060	3,297
12	65	58	63	57	65		65	2	0,036	1,000	0,025	1,384
13	Среднее значение					60,855		55				0,984
14	Стандарт отклонен					2,05						

3. С помощью функции *XI2TEST*, помещенной в ячейку *L13*, определите соответствие данных нормальному закону распределения. В ячейке *L13* появится значение вероятности того, что выборочные данные соответствуют нормальному закону распределения - 0,9842.

Фактический_интервал H4:H12 = {2;6;7;10;9;8;7;4;2}

Ожидаемый_интервал L4:L12 = {1,82632054961487}

4. Поскольку полученная вероятность соответствия экспериментальных данных $p = 0,98$ много больше, чем уровень значимости $\alpha = 0,05$, то можно утверждать, что нулевая гипотеза не может быть отвергнута, следовательно, данные не противоречат нормальному закону распределения. Поскольку полученная вероятность $p = 0,98$ близка к 1, можно говорить о высокой степени вероятности того, что экспериментальные данные соответствуют нормальному закону.

1. Определите, лежит ли значение 19 внутри границ 95%-ного доверительного интервала выборки 2, 3, 5, 7, 4, 9, 6, 4, 9, 10, 4, 7, 19.

3. Найдите соответствие экспериментальных данных нормальному закону распределения для следующей выборки весов (кг): 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27.

Лабораторная работа №4

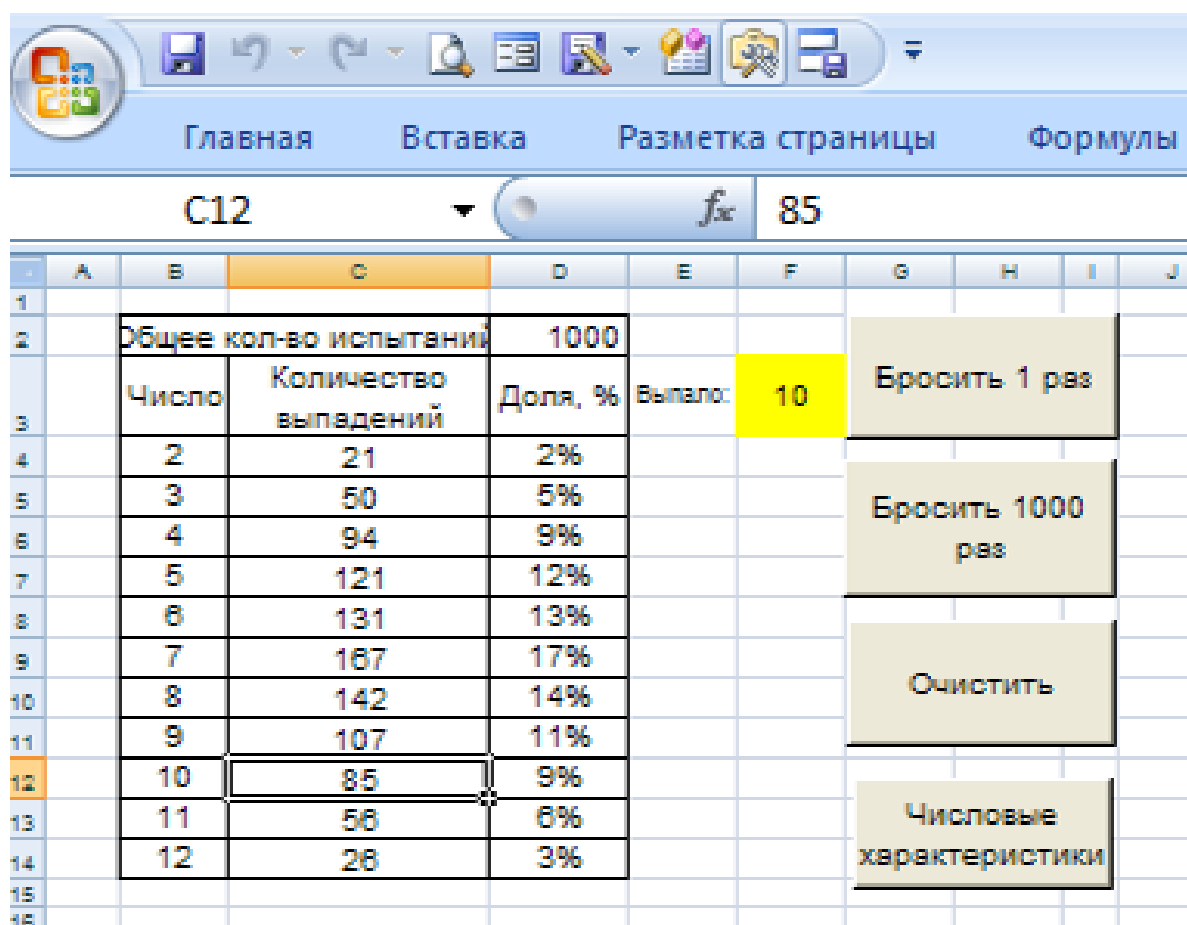
Цель работы: обработка данных, полученных в ходе численного моделирования. Определение соответствия эмпирического распределения заданному теоретическому распределению.

Задача №1

Отладка программы численного моделирования распределения случайной величины

Для моделирования распределения случайной величины написать и отладить программу на VBA, в которой случайная величина задается суммой двух случайных целых величин, находящихся в диапазоне от 1 до 6.

На рабочем листе MS Excel «Экспериментальное распределение» создать таблицу (рис. 4.1), в которую из программы будут поступать данные об общем количестве испытаний, количестве абсолютных и относительных частот наступления события - суммы двух случайных целых величин.



	A	B	C	D	E	F	G	H	I	J
1										
2		Общее кол-во испытаний		1000						
3		Число	Количество выпадений	Доля, %	Выпало:	10	Бросить 1 раз			
4		2	21	2%			Бросить 1000 раз			
5		3	50	5%						
6		4	94	9%						
7		5	121	12%						
8		6	131	13%			Очистить			
9		7	167	17%						
10		8	142	14%						
11		9	107	11%			Числовые характеристики			
12		10	85	9%						
13		11	56	6%						
14		12	26	3%						
15										
16										

Рис. 4.1. Результаты численного моделирования

На этом же рабочем листе MS Excel создать:

1. кнопку «Бросить 1 раз» и привязать к ней процедуру **Sub Бросить2Кубика_1_раз_Click()**, имитирующую бросание двух кубиков, нахождение суммы выпавших значений и внесение в таблицу обновленной информации (рис. 4.2).

2. кнопку «Бросить 1000 раз», вызывающую на исполнение процедуру **Sub Бросить2_1000_раз_Click()**, имитирующую бросание двух кубиков 1000 раз (рис. 4.2).

3. кнопку «Очистить» вызывающую на исполнение процедуру **Sub Очистить_Click()**, очищающую содержимое ячеек таблицы в диапазоне **C4: C14** (рис.4.2).

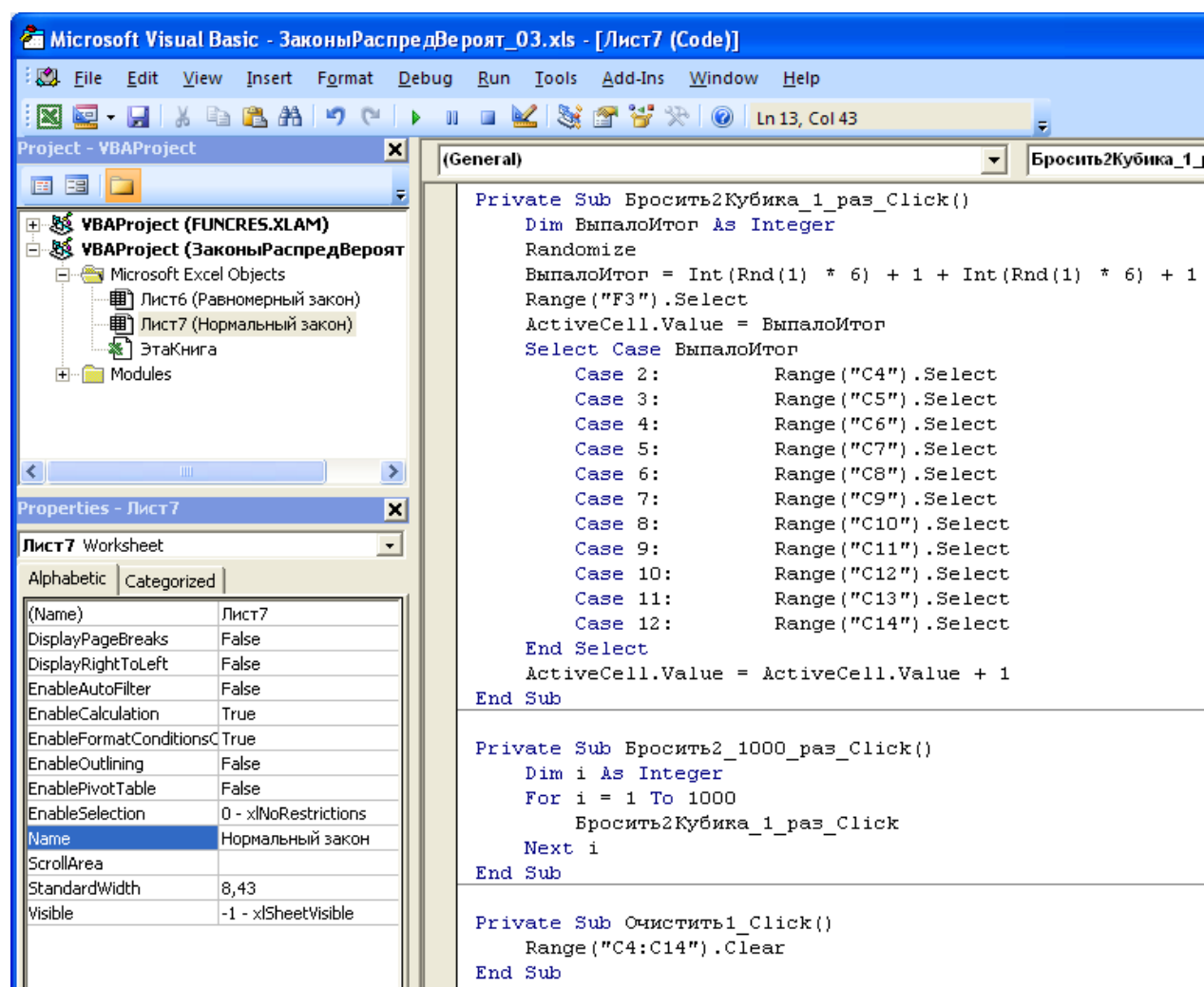


Рис. 4.2. Окно редактора VBA MS EXCEL с кодом программы

По результатам численного моделирования построить диаграмму распределения экспериментальных абсолютных частот появления случайной величины, расположив ее ниже таблицы результатов (рис. 4.3).

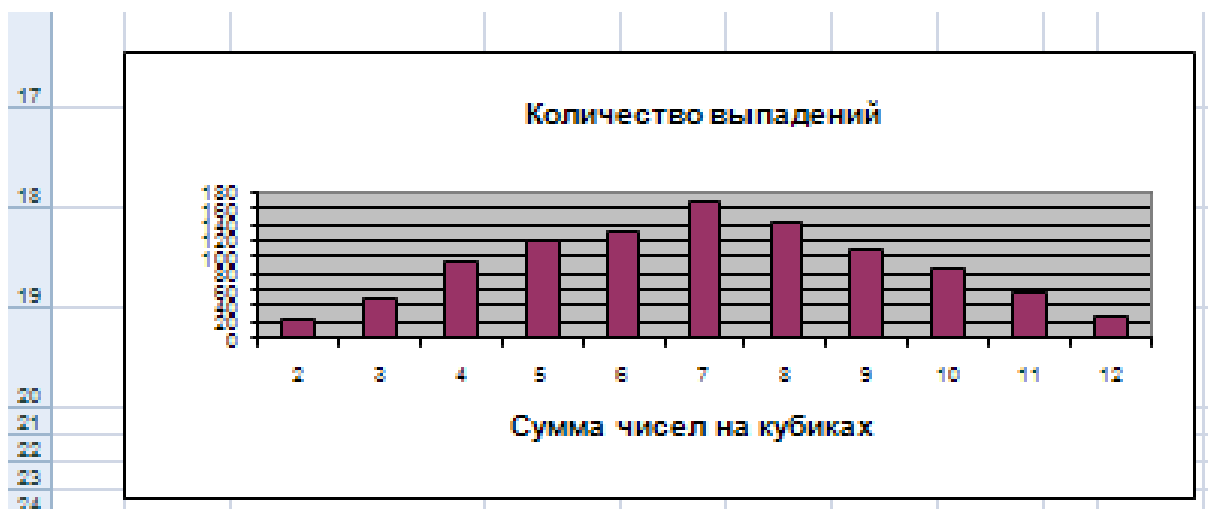


Рис. 4.3. Диаграмма распределения случайной величины

Ниже диаграммы (строки 26 – 29) расположить таблицу (рис. 4.4), в которую при нажатии кнопки «Числовые характеристики» выводятся характеристики распределения случайной величины, вычисленные с помощью встроенных функций MS Excel и функций, написанных на VBA (ExpVal, Variance, Skewness, Kurtosis):

		Excel	VBA	
	Математическое ожидание	7,005	7,005	$MX = \sum_k x_k p_k$
	Дисперсия	5,667	5,667	$DX = MX^2 - (MX)^2$
	Асимметрия	0,018	0,018	$\frac{M(X - MX)^3}{(DX)^3}$
	Экцесс	2,333	2,333	$\frac{M(X - MX)^4}{(DX)^4}$

Рис. 4.4. Вычисленные числовые характеристики распределения случайной величины

Математическое ожидание дискретной случайной величины X , принимающей конечное число значений x_i с вероятностями p_i :

$$MX = \sum_k x_k p_k ;$$

Дисперсия случайной величины X – число, характеризующее рассеяния значений случайной величины X :

$$DX = MX^2 - (MX)^2;$$

Асимметрия - центральный момент третьего порядка служит для оценки асимметрии распределения. Если распределение симметрично относительно точки $x = MX$, то асимметрия равна нулю:

$$As = \frac{M(X - MX)^3}{(DX)^2}$$

Экспесс - центральный момент четвертого порядка служит для оценки степени крутизны кривой распределения вблизи центра распределения по отношению к кривой нормального распределения

$$Ex = \frac{M(X - MX)^4}{(DX)^2}$$

Коды функций вычисления числовых характеристик распределения случайной величины:

```
Function ExpVal(Values, Prob)
```

```
    ExpVal = 0
```

```
    For i = 1 To UBound(Values)
```

```
        ExpVal = ExpVal + Values(i) * Prob(i)
```

```
    Next i
```

```
End Function
```

```
Function Variance(Values, Prob)
```

```
    Dim tmpVal()
```

```
    ReDim tmpVal(UBound(Values))
```

```
    For i = 1 To UBound(Values)
```

```
        tmpVal(i) = Values(i) ^ 2
```

```
    Next i
```

```
    Variance = ExpVal(tmpVal, Prob) - ExpVal(Values, Prob) ^ 2
```

```
End Function
```

```

Function Skewness(Values, Prob)
    Dim tmpVal()
    ReDim tmpVal(UBound(Values))
    For i = 1 To UBound(Values)
        tmpVal(i) = (Values(i) - ExpVal(Values, Prob)) ^ 3
    Next i
    If Variance(Values, Prob) <> 0 Then
        Skewness = ExpVal(tmpVal, Prob) / (Variance(Values, Prob) ^ 2)
    End If
End Function

```

```

Function Kurtosis(Values, Prob)
    Dim tmpVal()
    ReDim tmpVal(UBound(Values))
    For i = 1 To UBound(Values)
        tmpVal(i) = (Values(i) - ExpVal(Values, Prob)) ^ 4
    Next i
    If Variance(Values, Prob) <> 0 Then
        Kurtosis = ExpVal(tmpVal, Prob) / (Variance(Values, Prob) ^ 2)
    End If
End Function

```

```

Sub Числ_Хар()
    Dim Vals(11)
    Dim Prob(11)
    For i = 1 To 11
        Vals(i) = Cells(i + 3, 2): Prob(i) = Cells(i + 3, 4)
    Next i
    Cells(26, 6) = ExpVal(Vals, Prob)
    Cells(27, 6) = Variance(Vals, Prob)
    Cells(28, 6) = Skewness(Vals, Prob)
    Cells(29, 6) = Kurtosis(Vals, Prob)
End Sub

```

Задача №2

Проверка соответствия экспериментального распределения нормальному закону распределения

При решении реальных задач конкретный закон распределения случайной величины и его параметры заранее неизвестны. При этом применяемые статистические методы разработаны для конкретных видов закона распределения, что требует отнесение экспериментально полученного закона распределения к какому-либо виду известных распределений.

Среди непрерывных распределений случайной величины наиболее важным является *нормальное распределение*. Среди дискретных распределений - *биномиальное распределение*.

Наиболее часто проверяется предположение о нормальном распределении. Нормальное распределение непрерывных случайных величин в одномерном случае задаётся функцией, совпадающей с функцией Гаусса:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ — математическое ожидание, а σ — среднеквадратическое отклонение.

Для проверки соответствия полученной экспериментальной модели распределения случайной величины нормальному закону распределения можно применить графический метод и вычислить выборочные параметры формы распределения.

В MS Excel на рабочем листе «Теоретическое распределение» создать таблицу (рис 4.5) для сравнения *экспериментальных относительных частот* результатов численного моделирования и *теоретических относительных частот*. Данные для столбцов А, В и С таблицы взять из таблицы результатов численного моделирования (рис. 4.1). Столбец D таблицы заполнить *накопленными частотами*.

	A	B	C	D	E	F
1	Распределение дискретных случайных величин					
2	описываемое нормальным законом распределения					
3						
4	Число	Абсолютные частоты	Относительные частоты	Накопленные частоты	Относительные частоты	Абсолютные частоты
5		(экспериментальные)	(экспериментальные)		(теоретические)	(теоретические)
6	2	21	0,021	0,021	0,018	18,381
7	3	50	0,050	0,071	0,041	40,702
8	4	94	0,094	0,165	0,076	75,548
9	5	121	0,121	0,286	0,118	117,543
10	6	131	0,131	0,417	0,153	153,297
11	7	167	0,167	0,584	0,168	167,584
12	8	142	0,142	0,726	0,154	153,567
13	9	107	0,107	0,833	0,118	117,958
14	10	85	0,085	0,918	0,076	75,949
15	11	56	0,056	0,974	0,041	40,990
16	12	26	0,026	1,000	0,019	18,544
17		1000	1,000		0,980	980,062
18						
19			Среднее	7,005		
20			Стандартное отклонен	2,381		
21						

Рис. 4.5. Результаты вычисления экспериментальных и теоретических относительных частот

Теоретические относительные частоты (столбец *E*) получить с помощью встроенной функции MS Excel *НОРМРАСП* (рис.4.6), вычисляющей *вероятность нормальной функции распределения*:

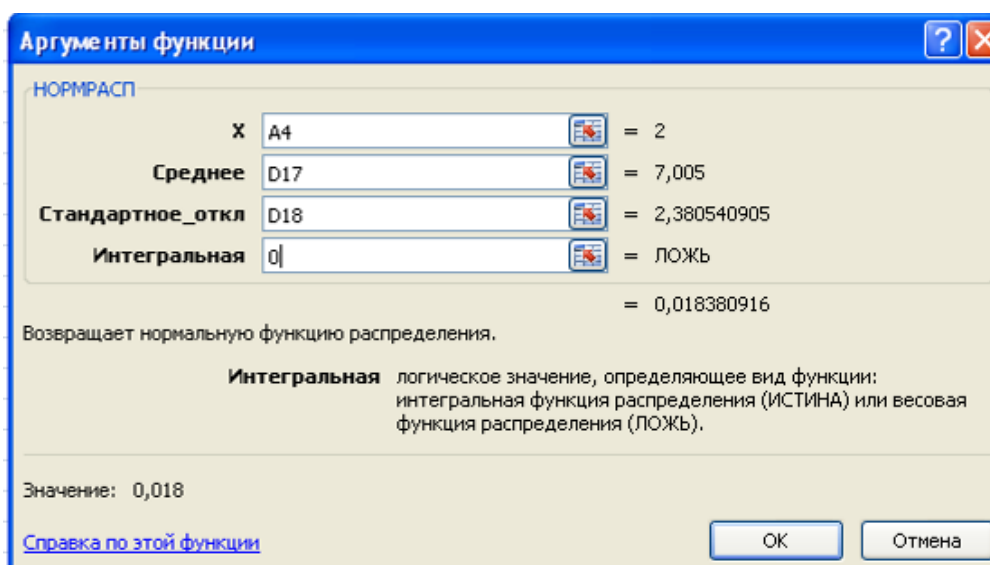


Рис. 4.6. Заполнения диалогового окна функции *НОРМРАСП*

Теоретические абсолютные частоты (столбец *F*) подсчитать по формуле $=E6* \$B\$17$.

По полученным данным построить графики функций плотности распределения экспериментальных и теоретических абсолютных частот:

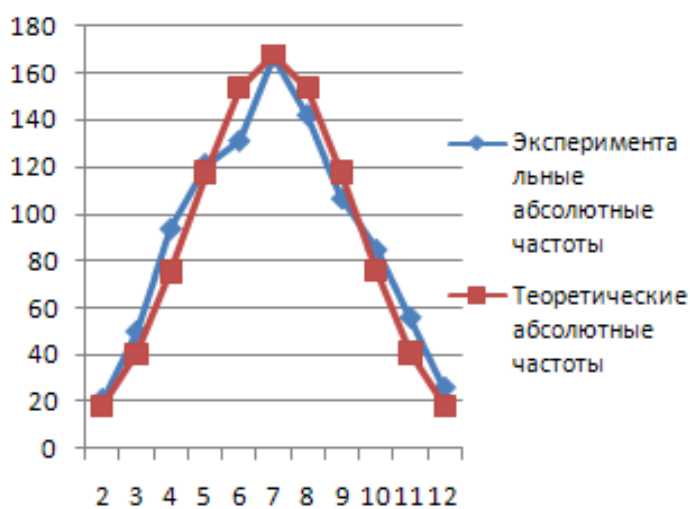


Рис. 4.7. Графики функций плотности распределения экспериментальных и теоретических абсолютных частот

Графический метод показал близкое совпадение распределений.

При большом числе наблюдений (в эксперименте их было 1000) достоверные результаты дает вычисление параметров формы распределения: *эксцесса* и *асимметрии*.

Предположение о нормальности распределения не противоречит полученным результатам (рис. 4.4): *асимметрия* близка к нулю, и лежит в диапазоне от -0,2 до 0,2, *эксцесс* находится в диапазоне от 2 до 4.

Предложенный алгоритм моделирования распределения дискретных случайных величин показал, что его результаты могут быть описаны нормальным законом распределения.

Задача №3

Проверка соответствия экспериментального распределения биномиальному закону распределения

Биномиальное распределение - одно из самых распространенных дискретных распределений. Оно возникает в тех случаях, когда ищут, сколько раз происходило некоторое событие в серии из известного числа независимых наблюдений, выполняемых в одинаковых условиях.

Для проверки соответствия полученной экспериментальной модели распределения случайной величины биномиальному закону распределения повторить применение графического метода.

В MS Excel на рабочем листе «Теоретическое распределение» создать вторую таблицу, расположив ее ниже (рис 4.8) для сравнения *экспериментальных относительных частот* результатов численного моделирования и *теоретических относительных частот*. Данные столбцов A, B, C и D таблицы дублируются из таблицы (рис. 4.5) задачи №2.

	A	B	C	D	E	F
22	Распределение дискретных случайных величин					
23	описываемое биномиальным законом распределения					
24		Абсолютные частоты	Относительные частоты	Накопленные частоты	Вероятность успеха	Абсолютные частоты
25	Число	(экспериментальные)	(экспериментальные)			(теоретические)
26						
27	2	21	0,021	0,021	0,028	21
28	3	50	0,050	0,071	0,056	49
29	4	94	0,094	0,165	0,083	78
30	5	121	0,121	0,286	0,111	108
31	6	131	0,131	0,417	0,139	138
32	7	167	0,167	0,584	0,167	168
33	8	142	0,142	0,726	0,139	143
34	9	107	0,107	0,833	0,111	117
35	10	85	0,085	0,918	0,083	91
36	11	56	0,056	0,974	0,056	65
37	12	26	0,026	1,000	0,028	21
38		1000	1,000		1,000	999,000

Рис. 4.8. Результаты вычисления экспериментальных и теоретических относительных частот

Выполнить подсчет вероятности успеха для событий 2 – 12:

Число	2	3	4	5	6	7
Вероятность	$(1/6*1/6)$	$(1/6*1/6)*2$	$(1/6*1/6)*3$	$(1/6*1/6)*4$	$(1/6*1/6)*5$	$(1/6*1/6)*6$
Число	8	9	10	11	12	
Вероятность	$(1/6*1/6)*5$	$(1/6*1/6)*4$	$(1/6*1/6)*3$	$(1/6*1/6)*2$	$(1/6*1/6)$	

Рис. 4.9. Подсчет вероятности успеха

В MS Excel для вычисления значения случайной величины по заданной вероятности используется функция *КРИТБИНОМ*.

Функция *КРИТБИНОМ* вычисляет наименьшее значение числа успешных исходов случайной величины, для которого интегральное биномиальное распределение больше или равно заданной величине (критерию).

Применить функцию *КРИТБИНОМ* для подсчета теоретических абсолютных частот распределения (рис. 4.10):

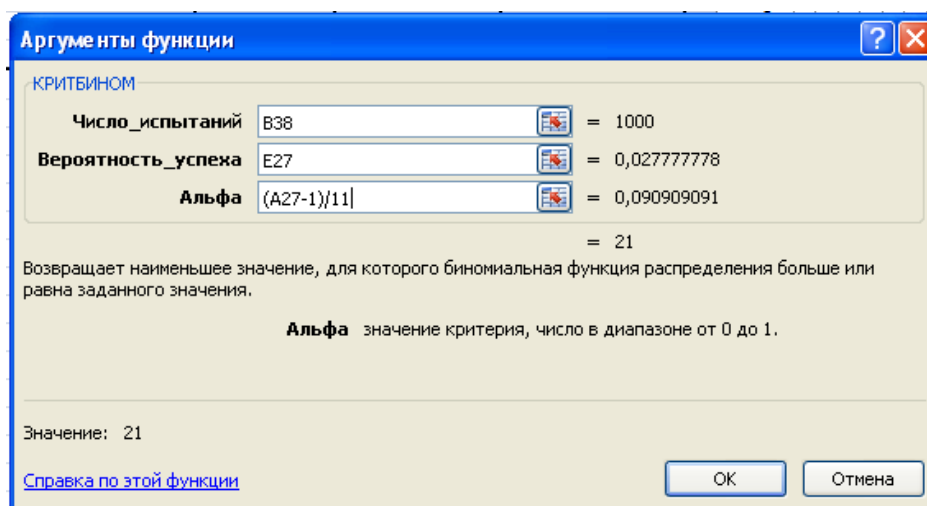


Рис. 4.10. Заполнения диалогового окна функции *КРИТБИНОМ*

По полученным данным построить графики функций плотности распределения экспериментальных и теоретических абсолютных частот (рис. 4.11):

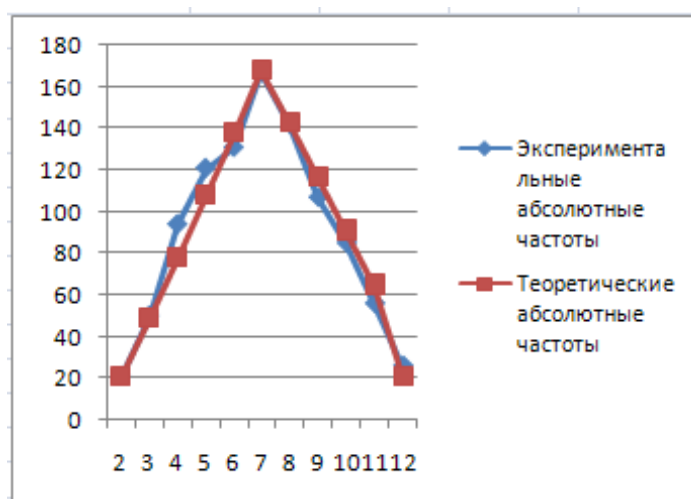


Рис. 4.11. Графики функций плотности распределения экспериментальных и теоретических абсолютных частот

Графический метод показал близкое совпадение распределений экспериментальных данных и теоретической модели.

Выводы

Для оценки соответствия имеющихся экспериментальных данных теоретической модели обычно используют графический метод, выборочные параметры формы распределения и критерии согласия.

Графический метод позволяет дать ориентировочную оценку расхождения или совпадений распределений.

Достоверные результаты дает использование критериев согласия. Так называют статистические критерии, предназначенные для проверки согласия экспериментальных данных и теоретической модели. Среди критериев согласия большое распространение получил непараметрический критерий χ^2 .

В MS EXCEL критерий хи-квадрат реализован в функции *ХИ2ТЕСТ*. Функция *ХИ2ТЕСТ* вычисляет вероятность совпадения фактических значений и теоретических значений. Если вычисленная вероятность ниже уровня значимости (0,05), то утверждается, что наблюдаемые значения не соответствуют теоретической модели. Если вычисленная вероятность близка к 1, то можно говорить о высокой степени соответствия экспериментальных данных теоретической модели.

С помощью функции *ХИ2ТЕСТ*, помещенной в ячейку F19 (рис. 4.14), определить соответствие фактических данных нормальному закону распределения:

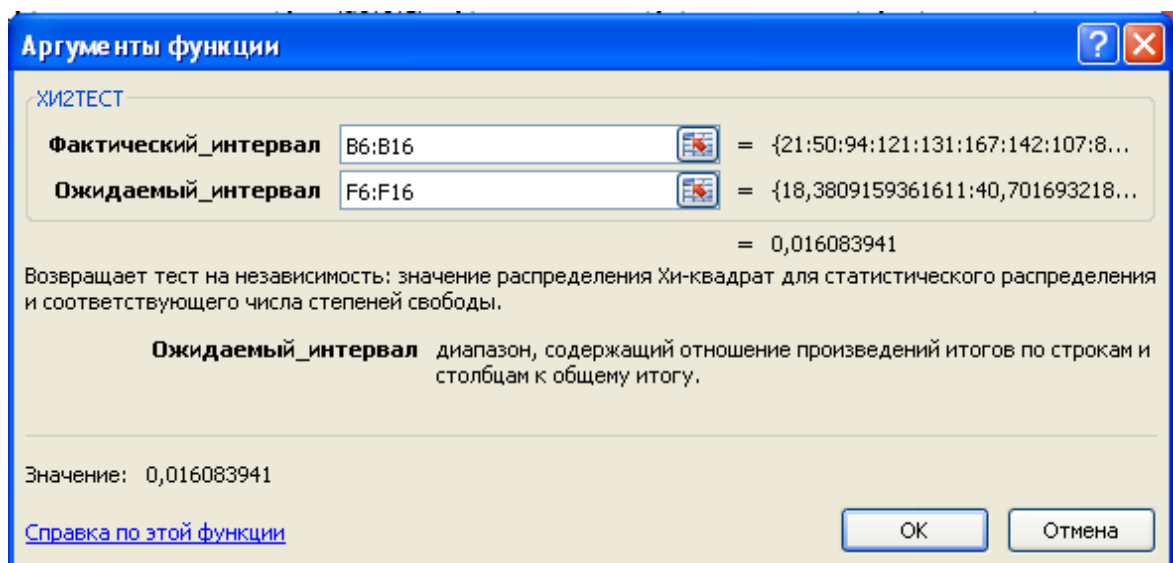


Рис. 4.12. Пример заполнения рабочих полей функции *ХИ2ТЕСТ*

Поскольку полученная вероятность соответствия экспериментальных данных $p = 0,02$ меньше уровня значимости $\alpha = 0,05$, следовательно, экспериментальные данные противоречат нормальному закону распределения.

С помощью функции *ХИ2ТЕСТ*, помещенной в ячейку *F40* (рис. 4.14), определить соответствие фактических данных биномиальному закону распределения:

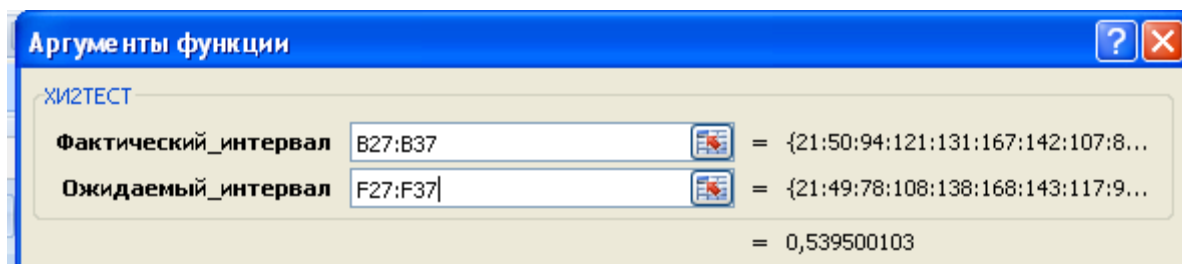


Рис. 4.13. Пример заполнения рабочих полей функции *ХИ2ТЕСТ*

Поскольку полученная вероятность соответствия экспериментальных данных $p = 0,54$ больше, чем уровень значимости $\alpha = 0,05$, то можно утверждать, что данные не противоречат биномиальному закону распределения.

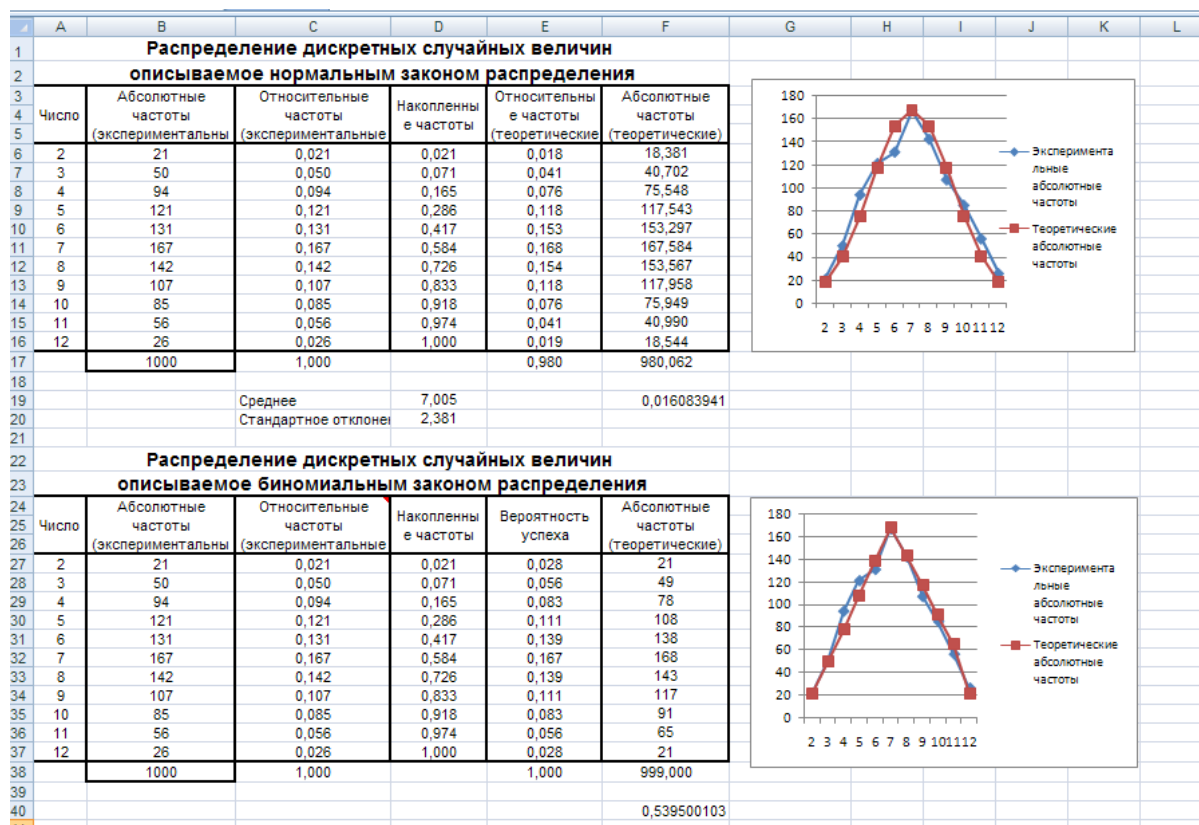


Рис. 4.14. Вычисление критериев согласия

Задание:

Выполните задачи №№ 1-3 для $100 + 20 \cdot N$ испытаний, где N – номер по журналу.

Корреляционный анализ

Одна из наиболее распространенных задач статистического анализа состоит в изучении связи между отдельными наблюдаемыми случайными величинами. Обычно взаимосвязь между выборками носит не функциональный, а вероятностный (стохастический) характер. В этом случае нет строгой 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 объемом n связанных пар наблюдений (x_i, y_i) из совместной генеральной совокупности X и Y . Существует несколько видов коэффициентов корреляции, применение которых зависит от предположений о совместном распределении величин X и Y ,

Для оценки степени взаимосвязи наибольшее распространение получил коэффициент линейной корреляции, предполагающий нормальный закон распределения выборок.

Коэффициент корреляции - параметр, характеризующий степень линейной взаимосвязи между двумя выборками. Коэффициент корреляции изменяется от -1 (строгая обратная линейная зависимость) до 1 (строгая прямая пропорциональная зависимость). При значении 0 линейной зависимости между двумя выборками нет. Под прямой зависимостью понимают зависимость, при которой увеличение или уменьшение значения одного признака выборки ведет к увеличению или уменьшению значения второго признака выборки.

Выборочный коэффициент линейной корреляции между двумя случайными величинами X и Y рассчитывается по формуле:

$$r = \frac{\sum_k (x_k - M_x)(y_k - M_y)}{\sqrt{\sum_k (x_k - M_x)^2 (y_k - M_y)^2}}.$$

Коэффициент корреляции является безразмерной величиной и его значение не зависит от единиц измерения случайных величин X и Y .

На практике коэффициент корреляции принимает некоторые промежуточные значения между -1 и 1 (рис. 5.1). Для оценки степени взаимосвязи руководствуются следующими эмпирическими правилами. Если коэффициент корреляции (r) по абсолютной величине больше $0,95$, то принято считать, что между параметрами существует практически линейная зависимость. Если коэффициент корреляции r лежит в диапазоне от $0,8$ до $0,95$, говорят о сильной степени линейной связи между параметрами. Если $0,6 < |r| < 0,8$, говорят о наличии линейной связи между параметрами. При $|r| < 0,4$ считают, что линейную взаимосвязь между параметрами выявить не удалось.

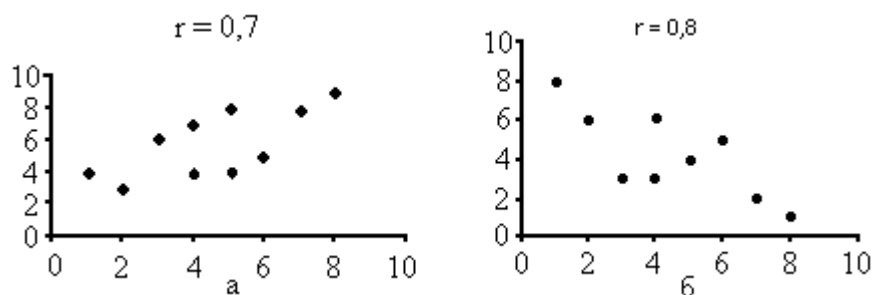


Рис. 5.1. Примеры прямой (а) и обратной (б) корреляционной зависимостей

В *MS Excel* для вычисления парных коэффициентов линейной корреляции используется функция *KORPEЛ* (*массив1*; *массив2*). Параметрами функции являются: *массив1* - диапазон ячеек первой случайной величины; *массив2* – диапазон ячеек второй случайной величины.

Лабораторная работа №5

Пример 1. Имеются результаты рентабельности предприятий и коэффициентов сменности оборудования.

Необходимо определить, имеется ли взаимосвязь между рентабельностью и коэффициентом сменности оборудования.

Решение: Введем данные в рабочую таблицу. Затем вычислим значение коэффициента корреляции между выборками.

	А	В	С
	Номер предприятия	Рентабельность, %	Коэффициент сменности оборудования
1			
2	1	13,28	1,14
3	2	22,31	1,85
4	3	15,27	1,14
5	4	12,99	1,33
6	5	25,78	1,74
7	6	28,47	1,9
8	7	12,97	1,16
9	8	23,47	1,86
10	9	10,47	1,17
11	10	13,58	1,23
12			
13	0,935		

Рис. 5.2. Вычисления коэффициента корреляции

В ячейку *A13* поместим функцию *KORPEЛ* (*массив1*; *массив2*). Введем параметры: *массив1* = значения рентабельности и *массив2* = коэффициент сменности оборудования.

Появится значение коэффициента корреляции - 0,935. Значение коэффициента корреляции находится в диапазоне от 0,8 до 0,95. Значит, имеется сильная степень линейной связи между параметрами ($r = 0,935$).

Корреляционная матрица

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

Корреляционная матрица - это квадратная (или прямоугольная) таблица, в которой на пересечении соответствующих строк и столбцов находится коэффициент корреляции между соответствующими параметрами.

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

Пример 2. Имеются данные наблюдений за состоянием погоды и посещаемостью музеев и парков: число ясных дней {8, 14, 20, 25, 20, 15}, количество посетителей музеев {495, 503, 380, 305, 348, 465}, количество посетителей парков {132, 348, 643, 865, 743, 541}.

Необходимо определить, существует ли взаимосвязь между состоянием погоды и посещаемостью музеев и парков.

Решение: Для выполнения корреляционного анализа введем в диапазон *B1:G3* исходные данные (рис. 5.3).

	A	B	C	D	E	F	G
1	Ясные дни	8	14	20	25	20	15
2	Посещаемость музея	495	503	380	305	348	465
3	Посещаемость парка	132	348	643	865	743	541

Рис. 5.3. Исходные данные

Затем в меню *Данные* выберем пункт *Анализ данных* и укажем *Корреляция*. В появившемся диалоговом окне укажем *Входной интервал* *A1:G3*. Укажем, что данные рассматриваются по строкам. Зададим выходной диапазон. *ОК*.

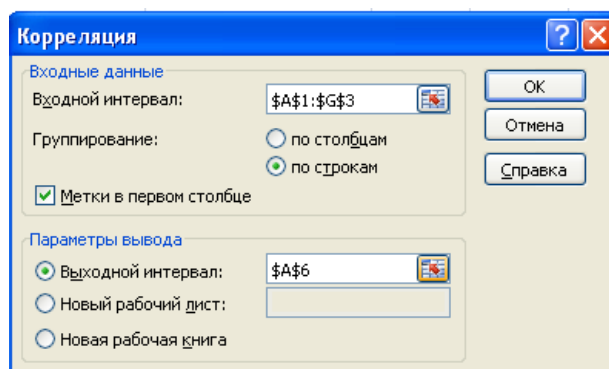


Рис. 5.4. Установка параметров корреляционного анализа

Результаты анализа. В выходной диапазон выводится корреляционная матрица, в которой на пересечении каждой 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 с самим собой.

6		Ясные дни	Посещаемость музея	Посещаемость парка
7	Ясные дни	1		
8	Посещаемость музея	-0,921854339	1	
9	Посещаемость парка	0,974575588	-0,919375244	1

Рис. 5.5. Результаты вычисления корреляционной матрицы

Интерпретация результатов. Из таблицы 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 $r = -0,92$.

Таким образом, в результате анализа выявлено:

- сильная степень обратной линейной взаимосвязи между посещаемостью музея и количеством солнечных дней ($r = -0,92$),
- практически линейная (очень сильная прямая) связь между посещаемостью парка и состоянием погоды ($r = 0,97$),
- между посещаемостью музея и парка имеется сильная обратная взаимосвязь ($r = -0,92$).

Задания:

1. Определите, имеется ли взаимосвязь между температурой($^{\circ}\text{C}$), ощущением температуры($^{\circ}\text{C}$), давлением(мм рт.ст.), влажностью воздуха(%), ветром(м/с) и наличием осадков по прогнозу погоды на середину дня на следующие 10 дней портала Яндекс.Погода.

Температура	2	-2	-2	-3	0	-3	-6
Ощущение темп.	-3	-9	-5	-9	-4	-7	-10
Давление	737	741	749	751	748	751	756
Влажность	90	80	91	82	90	88	87
Ветер	4,1	6,1	2,3	4,6	2,3	2	1,4
Осадки	1	1	1	1	1	1	0

2. Определите, имеется ли взаимосвязь между количеством осадков {5, 12, 11, 7, 3, 16}, урожайностью {138, 224, 205, 154, 110, 254} и ценой {754, 523, 544, 612, 876, 489}.

3. Определите, имеется ли взаимосвязь между количеством проведенных мировых турниров по киберспорту, количеству призовых мест российских команд по годам. Годы {2016,2017,2018,2019,2020,2021}; кол-во турниров {220, 201, 217, 176, 53, 47}; кол-во побед {147,142,168,153,48,42}

Регрессионный анализ

При исследовании взаимосвязей между выборками помимо процедуры корреляции применяют также и процедуру регрессии. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных.

Регрессионный анализ устанавливает формы зависимости между случайной величиной Y (зависимой) и значениями одной или нескольких переменных величин (независимых). Значения независимых величин считаются заданными. Такая зависимость определяется математической моделью, называемой уравнением регрессии, содержащей несколько неизвестных параметров. В ходе регрессионного анализа на основании выборочных данных находят оценки этих параметров, определяются статистические ошибки оценок и проверяется соответствие принятой математической модели экспериментальным данным.

В линейном регрессионном анализе связь между случайными величинами предполагается линейной. В самом простом случае в линейной регрессионной модели имеются две переменные X и Y . И требуется по n парам наблюдений $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ подобрать прямую линию, называемую линией регрессии. Уравнение этой линии $Y = aX + b$. С помощью регрессионного уравнения можно предсказать ожидаемое значение зависимой величины Y_0 , соответствующее значению независимой переменной X_0 .

Когда рассматривается зависимость между одной зависимой переменной Y и несколькими независимыми $X_1, X_2, \dots, X_n$, имеет место множественная линейная регрессия, уравнение которой имеет вид:

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots + a_nX_n,$$

где $a_1, a_2, \dots, a_n$ – искомые коэффициенты регрессии при независимых переменных $X_1, X_2, \dots, X_n$; a_0 – константа.

Мерой зависимости случайных величин регрессионной модели является коэффициент детерминации R^2 . Коэффициент детерминации определяет, с какой степенью точности полученное регрессионное уравнение описывает исходные данные.

Исследуется также регрессионная модель с помощью F -критерия (Фишера). Если величина F -критерия значима ($p < 0,05$), то регрессионная модель является значимой.

Достоверность отличия коэффициентов $a_0, a_1, a_2, \dots, a_n$ от нуля проверяется с помощью критерия Стьюдента. В случаях, когда $p > 0,05$, коэффициент может считаться нулевым, это означает, что влияние соответствующей независимой переменной на зависимую переменную недостоверно, и эта переменная может быть исключена из уравнения.

В *MS Excel* экспериментальные данные аппроксимируют линейным уравнением с 16 переменными:

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots + a_{16}X_{16},$$

Для получения коэффициентов регрессии используется процедура *Регрессия* из пакета анализа. Кроме того, могут быть использованы функция *ЛИНЕЙН* для получения параметров регрессионного уравнения и функция *ТЕНДЕНЦИЯ* для получения предсказанных значений Y в требуемых точках.

Для реализации процедуры *Регрессия* необходимо:

- выполнить команду *Данные ► Анализ данных*;
- в появившемся диалоговом окне *Анализ данных* в списке *Инструменты анализа* выбрать строку *Регрессия. ОК*;
- в появившемся диалоговом окне задать *Входной интервал Y* - ввести ссылку на диапазон анализируемых зависимых данных;
- указать *Входной интервал X* - ввести ссылку на диапазон независимых данных, содержащий до 16 столбцов анализируемых данных;
- указать выходной диапазон - ввести ссылку на ячейки, в которые будут выведены результаты анализа. Размер выходного диапазона будет определен автоматически;

Регрессия

Входные данные

Входной интервал Y:

Входной интервал X:

☐ Метки

☐ Константа - ноль

☐ Уровень надежности: %

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист:

☐ Новая рабочая книга

Остатки

☐ Остатки

☐ Стандартизованные остатки

☐ График остатков

☒ График подбора

Нормальная вероятность

☐ График нормальной вероятности

ОК

Отмена

Справка

Рис. 6.1. Заполнение диалогового окна *Регрессия*

- если необходимо визуально проверить отличие экспериментальных точек от предсказанных по регрессионной модели, следует установить флажок в поле *График подбора*;
- нажать кнопку *ОК*.

Результаты анализа. Выходной диапазон будет включать в себя результаты дисперсионного анализа, коэффициенты регрессии, стандартную погрешность вычисления Y , среднеквадратичные отклонения, число наблюдений, стандартные погрешности для коэффициентов.

Лабораторная работа №6

Пример 1. Заданы значения зависимой переменной $Y\{5, 7, 9, 12, 15, 16, 20, 25\}$, сопоставляя их со значениями независимой переменной $X\{6,3; 9; 12; 15; 19; 21; 25; 29,3\}$, построить регрессионное уравнение, позволяющее по заданному значению X определять предполагаемое значение Y .

Решение:

1. Введем данные в рабочую таблицу:

	A	B
1	X	Y
2	6	5
3	9	7
4	12	9
5	15	12
6	19	15
7	21	16
8	25	20
9	29	25

2. В пункте меню *Данные* выберем *Анализ данных*, укажем на строку *Регрессия*.

3. В появившемся диалоговом окне (рис. 6.1) зададим *Входной интервал Y* и *Входной интервал X*

4. Зададим выходной диапазон. Для этого установим переключатель в положение *Выходной интервал*, мышью в правом поле ввода *Выходной интервал* активируем левую верхнюю ячейку выходного диапазона.

5. Установим флажок в поле *График подбора*.

6. *ОК*.

В выходном диапазоне появятся следующие результаты:

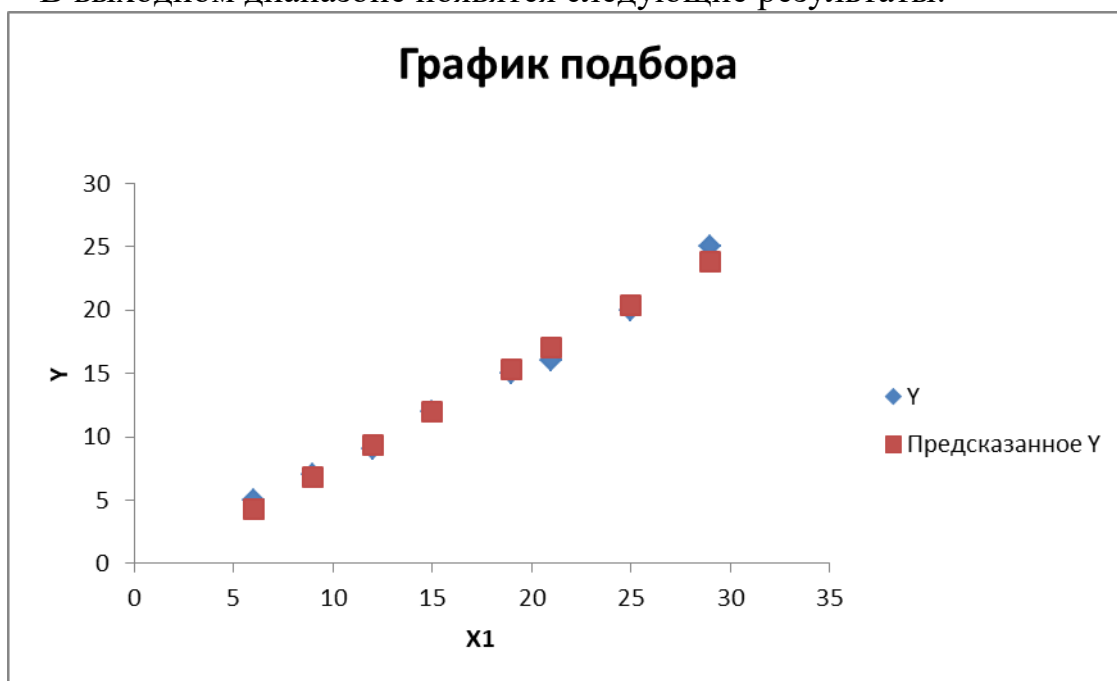


Рис. 6.2. График соответствия экспериментальных точек и предсказанных по регрессионной модели

Интерпретация результатов

L	M	N	O	P	Q
ВЫВОД ИТОГОВ					
<i>Регрессионная статистика</i>					
Множественный R	0,994650071				
R-квадрат	0,989328764				
Нормированный R-квадрат	0,987550225				
Стандартная ошибка	0,754261398				
Наблюдения	8				
<i>Дисперсионный анализ</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	1	316,4615385	316,4615385	556,259155	3,81276E-07
Остаток	6	3,413461538	0,568910256		
Итого	7	319,875			
	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>
Y-пересечение	-0,759615385	0,665652736	-1,14115866	0,29729464	-2,388408953
Переменная X 1	0,846153846	0,035876556	23,58514691	3,8128E-07	0,758367076

Оценка качества полученной модели:

- по уровню значимости критерия Фишера из таблицы *Дисперсионный анализ* (строка *Регрессия*, столбец *Значимость*) $F = 1.58E-07 \ll 0,05$, т.е. модель значима,

- по степени точности описания моделью процесса (*Регрессионная статистика*) R -квадрат = 0,989. Поскольку R -квадрат > 0,95, можно говорить о высокой точности аппроксимации. Модель хорошо описывает явление.

Значения коэффициентов модели определяются из таблицы в столбце *Коэффициенты* - в строке *Y-пересечение* приводится свободный член; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных.

В столбце *p-значение* приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда $p > 0,05$, коэффициент может считаться нулевым. Это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную и коэффициент может быть убран из уравнения.

Отсюда выражение для определения Y будет иметь следующий вид:

$$Y = 0,85 * X - 0,76$$

Если R -квадрат лежит в диапазоне от 0,8 до 0,95, говорят об удовлетворительной аппроксимации (модель адекватна описываемому явлению). Если R -квадрат < 0,6, принято считать, что точность аппроксимации недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейностей и т. д.).

Пример 2. Построить регрессионную модель для предсказания значения Y в зависимости от X_1 X_2 . В таблице приведены данные наблюдений.

X_1	X_2	Y
1	1,3	1160
1	1,3	1155
1,1	1,4	1158
1,1	1,4	1157
1,1	1,5	1160
1,1	1,5	1161
1	1,4	1157
1	1,5	1159
1,2	1,6	1256
1,2	1,7	1260
0,6	1	1040
0,6	1	1039
0,7	1,1	1039
0,7	1,15	1040
0,75	1,2	1040
0,7	1,2	1039
0,7	1,3	1040
0,7	1,3	1039
0,8	1,4	1040
0,8	1,4	1138
0,78	1,5	1240
0,8	1,5	1239
0,78	1,5	1241
0,78	1,6	1240
0,8	1,7	1239
0,8	1,8	1239
0,75	1,8	1240
0,78	1,9	1238
0,75	1,9	1238

Решение

1. Введем данные в диапазон $A1:C30$ рабочей таблицы *Excel*.
2. Далее *Данные ► Анализ данных ► Регрессия. ОК*.
3. В диалоговом окне зададим *Входной интервал* Y .
4. Укажем *Входной интервал* X_1 и X_2 непрерывной областью.
5. Установим флажок в поле *График подбора*.
6. Укажем выходной диапазон: поставим переключатель в положение *Выходной интервал*, затем указатель мыши наведем на левую верхнюю ячейку выходного диапазона ($E1$). *ОК*.

В выходном диапазоне появятся результаты регрессионного анализа и графики предсказанных точек (рис. 6.3).

E	F	G	H	I	J
ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,890				
R-квадрат	0,792				
Нормированный R-квадрат	0,775				
Стандартная ошибка	39,293				
Наблюдения	29,000				
Дисперсионный анализ					
	df	SS	MS	F	Значимость F
Регрессия	2	152412,899	76206,450	49,35846261	1,40483E-09
Остаток	26	40142,411	1543,939		
Итого	28	192555,310			
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%
Y-пересечение	681,9	50,293	13,559	2,67E-13	578,528
Переменная X 1	90,9	43,893	2,071	4,84E-02	0,684
Переменная X 2	274,7	31,834	8,628	4,16E-09	209,231

Интерпретация результатов.

- по уровню значимости критерия Фишера *Дисперсионный анализ* (строка *Регрессия*, столбец *Значимость*) $F, 1,4 \cdot 10^{-9} \ll 0,05$, т.е. модель значима,

- по степени точности описания моделью процесса (*Регрессионная статистика*) R -квадрат = 0,79. Поскольку R -квадрат близок к диапазону от 0,8 до 0,95 можно говорить о довольно высокой точности аппроксимации.

Значения коэффициентов модели определяются из таблицы в столбце *Коэффициенты* - в строке *Y-пересечение* приводится свободный член $a_0 = 682$; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных $a_1 = 91$ и $a_2 = 275$. В столбце *p-значение* приводится достоверность отличия соответствующих коэффициентов от нуля. Так как $p < 0,05$, значит все коэффициенты значимы.

Поэтому выражение для определения значения Y в зависимости от значений X_1 и X_2 имеет вид:

$$Y = 682 + 91 \cdot X_1 + 275 \cdot X_2.$$

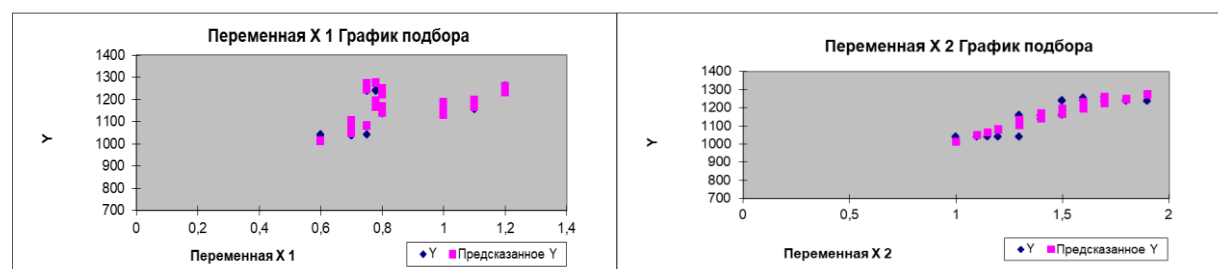


Рис. 6.3. Графики расположения фактических и предсказанных точек

Задания:

1. Постройте зависимость зарплаты (руб.) от возраста сотрудника по следующим данным:

Возраст	Зарплата
20	8000
50	25000
45	25000
40	20000
25	12000
30	18000

2. Постройте зависимость Y от X_1 и X_2 :

X_1	X_2	Y
1,85	18	5,4
1,8	25	5,7
1,75	20	4,8
1,7	24	5,1
1,68	21	4,5
1,73	19	4,8
1,77	22	5,1
1,81	23	5,6
1,76	18	4,7

Дисперсионный анализ

При необходимости оценить достоверность различия между значениями нескольких групп наблюдения (выборками) используют методы дисперсионного анализа.

Дисперсионный анализ предназначен для исследования воздействия на исследуемую случайную величину одного или нескольких независимых факторов. Влияющие на результат факторы считаются известными, и речь идет об оценке этого влияния.

Применение дисперсионного анализа возможно, если можно предположить соответствие выборочных групп генеральным совокупностям с нормальным распределением и независимость распределений наблюдений в группах.

Задача заключается в том, чтобы сравнить дисперсию, обусловленную случайными причинами, с дисперсией, вызываемой наличием исследуемого фактора. Если они значимо различаются, то считают, что фактор оказывает статистически значимое влияние на исследуемую переменную. Проверка различий выполняется по критерию Фишера.

Влияние случайной составляющей характеризует внутригрупповая дисперсия, а влияние изучаемого фактора - межгрупповая.

Внутригрупповая дисперсия рассчитывается по формуле:

$$s_2^2 = \frac{1}{m(n-1)} \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - M_i)^2,$$

межгрупповая:

$$s_1^2 = \frac{1}{m-1} \sum_{i=1}^m (M_i - M)^2.$$

M - общее среднее, $M_i = \frac{1}{n} \sum_{j=1}^m x_{ij}$, m - количество групп, n - количество

элементов в группе.

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

Лабораторная работа №7

Пример 1. Необходимо выявить, влияет ли расстояние от центра города на степень заполняемости гостиниц. 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:

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

Решение:

- Данные введем в рабочую таблицу *Excel* по столбцам:
 - заполняемость в центре города,
 - заполняемость на расстоянии от 3 до 5 км от центра,
 - заполняемость на расстоянии свыше 5 км от центра.
- Выполним команду *Данные ► Анализ данных*. В появившемся диалоговом окне *Анализ данных* в списке *Инструменты анализа* выберем процедуру *Однофакторный дисперсионный анализ*. *ОК*.
- В появившемся диалоговом окне *Однофакторный дисперсионный анализ* в поле *Входной интервал* задайте *A1:C7*.
- В разделе *Группировка* переключатель установим в положение по столбцам.
- Укажем выходной диапазон. Для этого поставим переключатель в положение *Выходной интервал*, затем щелкнем указателем мыши в правом поле ввода *Выходной интервал*, и щелчком мыши на ячейке *E1* укажем расположение выходного диапазона (рис. 7.1). *ОК*.

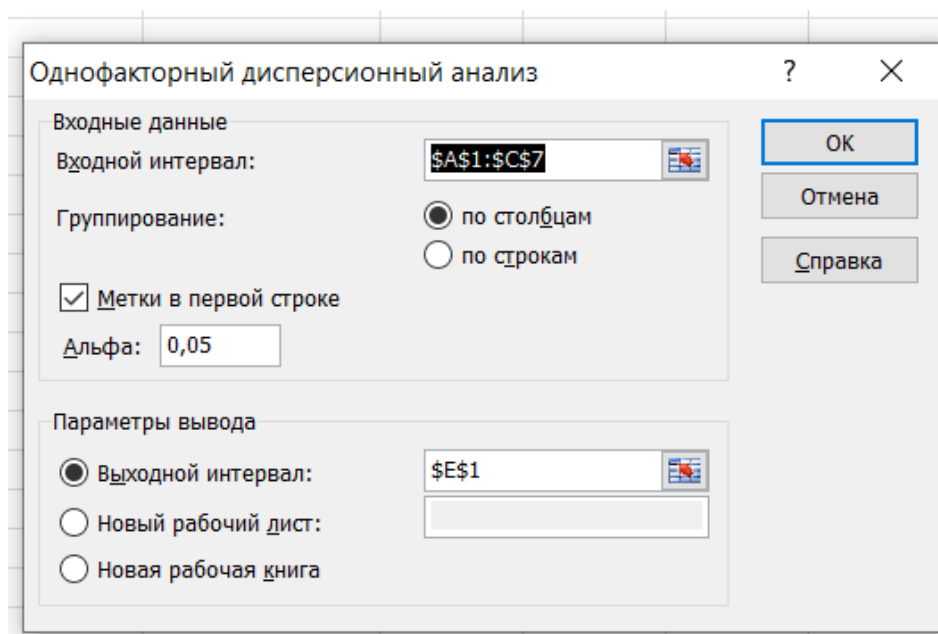


Рис. 7.1. Заполнение диалогового окна
Однофакторный дисперсионный анализ

Е	Ф	Г	Н	И	Ж	К
Однофакторный дисперсионный анализ						
ИТОГИ						
<i>Группы</i>	<i>Счет</i>	<i>Сумма</i>	<i>Среднее</i>	<i>Дисперсия</i>		
До 3 км	6	560	93,33	13,47		
От 3 до 5 км	6	516	86,00	14,00		
Свыше 5 км	6	475	79,17	32,97		
Дисперсионный анализ						
<i>Источник вариации</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-Значение</i>	<i>F критическое</i>
Между группами	602,33	2	301,17	14,95	0,000268401	3,68
Внутри групп	302,17	15	20,14			
Итого	904,50	17				

Рис.7.2. Результат работы инструмента
Однофакторный дисперсионный анализ

Результаты анализа. Выходной диапазон включает в себя результаты дисперсионного анализа: средние, дисперсии, критерий Фишера и другие показатели.

Интерпретация результатов. Влияние исследуемого фактора определяется по величине значимости критерия Фишера, которая находится в таблице *Дисперсионный анализ* на пересечении строки *Между группами* и столбца *P-Значение*. В случаях, когда $P\text{-Значение} < 0,05$, критерий Фишера значим, и влияние исследуемого фактора можно считать доказанным.

В данном примере эта величина равна $0,0002684 < 0,05$, следовательно, критерий Фишера значим, и влияние фактора расстояния от центра города на эффективность заполнения гостиниц доказано статистически.

Задания:

1. Определите, влияет ли фактор образования на уровень зарплаты на основании следующих данных:

Образование	Зарплата сотрудника					
высшее	32000	30000	26000	20000	19000	19000
среднее специальное	26000	20000	20000	19000	18000	17000
среднее	20000	20000	19000	18000	17000	17000

ЛИТЕРАТУРА

1. Кудлаев А.А. Технологии обработки информации средствами электронных таблиц. Теория вероятностей. М. МИИГАиК, 2020.-42стр.
2. Кудлаев А.А. Программирование в среде MS EXCEL. Учебно-методическое пособие по курсу «Технология программирования». М., МИИГАиК, 2016, - 37 стр.
3. Гаврилова В.В., Кудлаев А.А., Семов А.М. Информационные технологии. Офисные приложения. Учебное пособие. М.: МИИГАиК, 2012. – 60 стр.
4. Шень А. Вероятность: примеры и задачи. – М.: МЦНМО, 2012. – 72 с.
5. Зайцев А.А. Практические работы по математической статистике в среде MS EXCEL. Учебно-методическое пособие по курсу «Экономико-математические методы и моделирование». М., МИИГАиК, 2010. – 24 стр.
6. Ниворожкина Л.И., Морозова З.А. Теория вероятностей и математическая статистика. – М.: Эксмо, 2008. – 432 стр. (Техническое образование).
7. Гельман В.Я. Решение математических задач средствами EXCEL: Практикум. СПб.: Питер, 2003. – 240 стр.
8. Таннант-Смит Дж. Бейсик для статистиков. – М.: Мир, 1988. – 208 стр.
9. Справочник геодезиста: в 2-х книгах. Кн. 1/ Под редакцией Большакова В.Д. и Левчука Г.П. – М.: Недра, 1985. – 455 стр.