Рязанский государственный педагогический университет им С. А. Есенина

П Р А К Т И К У М

по курсу «Информатика»

Изучение табличного процессора «MS Excel» для специальностей «Социология» и «Управление персоналом»

Автор: Парадела В. Д.

Рязань 2008

Обязательная записка

Настоящий практикум подготовлен для студентов специальностей «Социология» и «Управление персоналом» с учетом дидактического принципа профессиональной направленности. При его проведении учитывается, что студенты данных специальностей уже изучили в первом семестре первого курса «MS Word», и поэтому обладают общими умениями работы с файлами прикладного пакета «MS Office». Кроме того, они также уже умеют обращаться к справочной системе, если не знают, как решить какой-то конкретный вопрос, поскольку приобретению такого навыка было уделено достаточное внимание при работе в «MS Word».

Данный лабораторный курс является пропедевтической основой изучения методов обработки результатов анкетирования. Без него не мысленно успешное обоснование нулевых гипотез в математической статистике. Но он не заменяет изучение методов обработки данных при использовании специализированных пакетов. Он только подготовит студентов к более глубокому анализу математической обработки статистических данных, показывая при этом некоторые возможности «MS Excel».

В пособии, кроме описания лабораторных работ рассматриваются основные понятия описательной и аналитической статистик. Оно создано на основе шестой главы пособия В. Я. Гельмана «Решение математических задач средствами «MS Excel»», издательство Питер, 2003.

Основные понятия и определения «MS Excel»

1. Файл – в «MS Excel» файл называется «книгой» и имеет расширение «xls».
2. Каждая книга состоит из трех листов по умолчанию и каждый лист из 256 столбцов и 65536 строк.

3. Пересечение одного столбца и одной строки называется ячейкой.
4. Каждый столбец обозначается латинской буквой или двумя латинскими буквами от A до IV и каждая строка числом от 1 до 65536 по умолчанию.
5. Каждая ячейка имеет название, определяемое её столбцом и строкой. Например, A4, $B25, C$34 или D44. Знак доллара «$» фиксирует столбец или ячейку (в зависимости от его положения) при ссылках.
6. Ячейки можно заполнять автоматически при помощи маркера заполнения (маленького черного квадратика в правом нижнем углу выделенной ячейки).
7. Множество ячеек, расположенных непрерывно формирует таблицу.
8. Используя данные в таблицах можно построить диаграммы.
9. В «MS Excel» можно проводить вычисления в ячейках при помощи формул.
10. Каждая формула начинается знаком равенства «=».
11. Формулы могут создаваться пользователем. Можно также использовать встроенные формулы (Функции).
12. В «MS Excel» есть пакеты формул для специалистов разных профессии, например для социологов и специалистов, управляющих персоналом.

Мы будем изучать «MS Excel» на примере использования формул для специальностей «Социология» и «Управление персоналом».

Основные понятия и определения математической статистики

Приступая к изучению статистики в среде «MS Excel» сначала нужно смотреть определенные понятия математической статистики, методы статистического исследования, а также некоторые выборочные функции распределения и основные выборочные характеристики.

Понятие математической статистики

Раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей, называется математической статистикой.

Математическая статистика имеет дело с массовыми явлениями. Она тесно связана с теорией вероятностей и базируется на ее математическом аппарате.

Целью статистического исследования является обнаружение и исследование соотношений между статистическими данными и их использованием 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: описательную и аналитическую статистику. Описательная статистика охватывает методы описания статистических данных, представления их в форме таблиц, распределений и т. п.

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

Пакет «MS Excel» оснащен средствами статистической обработки данных. И хотя он существенно уступает специализированным статистическим пакетам обработки данных, тем не менее, этот раздел математики представлен в «MS Excel» наиболее полно. В него включены

основные, наиболее часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы, позволяющие проводить необходимый статистический анализ данных об обществе и других системах.

При рассмотрении применения методов обработки статистических данных ограничимся только простейшими и наиболее часто используемыми методами, реализованными в мастере функций и пакете анализа ««MS Excel»».

Выборочный метод

По охвату статистической совокупности исследование может быть сплошное или не сплошное. При сплошном статистическом исследовании группа наблюдения формируется путем полного охвата всех единиц изучаемого явления. Множество всех единиц наблюдения, охватываемых таким сплошным наблюдением, называется генеральной совокупностью.

Основным методом не сплошного наблюдения является выборочный метод. Если интересующая нас совокупность слишком многочисленна, либо ее элементы малодоступны, а также, если имеются другие причины (организационные, финансовые, физические и т. п.), не позволяющие изучать сразу все ее элементы, прибегают к изучению какой-то части этой совокупности. Эта выбранная для полного исследования группа элементов называется выборкой или выборочной совокупностью.

Выборка − это группа элементов, выбранная для исследования из всей совокупности элементов. Задача выборочного метода состоит в том, чтобы сделать правильные выводы относительно всего собрания объектов, их совокупности. Например, пробуя пищу, повар по одной ложке делает заключения о качестве приготавливаемого во всей кастрюле.

Конечной целью изучения выборочной совокупности всегда является получение информации о генеральной совокупности. Поэтому естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, то есть была бы

репрезентативной или представительной. Для получения репрезентативной выборки необходимо четко определять, что понимается под генеральной совокупностью. Ее состав и численность зависят от объектов и целей проводимого исследования. Например, если мы хотим получить данные о поступающих во все вузы города, то абитуриенты данного института есть выборка из более широкой генеральной совокупности − всех абитуриентов вузов города, и эта выборка не обязательно будет являться представительной.

В тех случаях, когда генеральная совокупность недостаточно известна, обычно не удается предложить лучшего способа получения представительной выборки, чем случайный выбор. При этом случайная выборка формируется случайным отбором − из генеральной совокупности наудачу извлекается по одному объекту.

Выборочная функция распределения

По определению классическая вероятность равна отношению числа испытаний (т), в которых событие появилось, к общему количеству произведенных испытаний (п). Такая вероятность также называется статистической частотой.

На практике, сведения о законе распределения случайной величины можно получить независимыми многократными повторениями опыта, в котором измеряются значения интересующей исследователей случайной величины (варианты). На основе информации из полученной выборки можно построить приблизительные значения для функции распределения и других характеристик случайной величины.

Выборочной (эмпирической) функцией распределения случайной величины ξ построенной по выборке x1, x2, ...,xn,,, называется функция Fn(x), равная доле таких значений xi что xi <x, i= 1, ...,п.

Другими словами, Fn(x) есть частота события xi < x в ряду x1,, x2,…, xn.

Связь между эмпирической функцией распределения и функцией распределения (теоретической функцией распределения) такая же, как связь

между частотой события и его вероятностью: функция Fn(x)→F(x) при n → ∞.

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 гистограммы выборочных функций распределения. Если соответствующие точки относительных частот соединить ломаной линией, то полученная диаграмма будет называться полигоном частот. Кумулятивная кривая будет получена, если по оси абсцисс, откладывать интервалы, а по оси ординат − число или доли элементов совокупности, имеющих значение, меньшее или равное заданному.

При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график плотности распределения, а кумулятивная кривая − в график функции распределения.

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

О Функция «ЧАСТОТА» вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив цифр. Функция задается в качества формулы массива. ЧАСТОТА(массив данных;массив_карманов). Здесь:

массив данных − это массив или ссылка на множество данных, для которых вычисляются частоты.

массив карманов −это массив или ссылка на множество интервалов, в

которые группируются значения аргумента массив данных.

Отметим, что количество элементов в возвращаемом массиве на единицу больше числа элементов в массив_карманов. Дополнительный элемент в возвращаемом массиве содержит количество значений, больших, чем максимальное значение в интервалах.

О Процедура «Гистограмма» используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений. Процедура выводит результаты в виде таблицы и гистограммы.

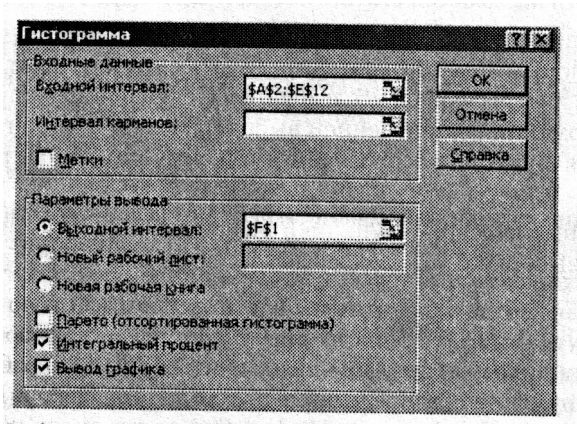


Рис. 1.1. Пример заполнения диалогового окна Гистограмма

Параметры диалогового окна «Гистограмма» представлены на рис. 1.1.

во Входной диапазон вводится диапазон исследуемых данных;

в поле Интервал карманов (необязательный параметр) может вводиться диапазон ячеек или необязательный набор граничных значений, определяющих выбранные интервалы (карманы). Эти значения должны быть введены в возрастающем порядке. В «MS Excel» вычисляется число попаданий данных между началом интервала и соседним большим по порядку. При этом включаются значения на нижней границе интервала и не включаются значения верхней границе.

Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически;

рабочее поле Выходной диапазон предназначено для ввода ссылки на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически;

переключатель Интегральный процент позволяет установить режим генерации интегральных процентных отношений и включения в гистограмму графика интегральных процентов;

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

Пример 1.1. Построить эмпирическое распределение веса студентов в килограммах 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, а в диапазон A2:E12 − значения веса студентов.
2. Выберите ширину интервала 1 кг. Тогда при крайних значениях веса 57 кг

и 65 кг получится 9 интервалов. В ячейки G1 и G2 введите названия интервалов Вес и кг, соответственно. В диапазон G4:G12 введите граничные значения интервалов (57, 58, 59, 60, 61, 62, 63, 64, 65).

3. Введите заголовки создаваемой таблицы: в ячейки H1:H2 − Абсолютные частоты, в ячейки I1:I2 − Относительные частоты, в ячейки J1;J2 − Накопленные частоты.
4. Заполните столбец абсолютных частот. Для этого выделите для них блок ячеек H4: H12 (используемая функция «ЧАСТОТА» задается в виде формулы массива). С панели инструментов «Стандартная» вызовите Мастер функций (кнопка fx). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ЧАСТОТА, после чего нажмите кнопку ОК. Появившееся диалоговое окно ЧАСТОТА необходимо за серое поле мышью отодвинуть вправо на 1-2 см от данных (при нажатой левой кнопке). Указателем мыши в рабочее поле Массив_ данных введите диапазон данных наблюдений (A2:E12). В рабочее поле Двоичный массив, мышью введите диапазон интервалов (G4:G12). Последовательно нажмите комбинацию клавиш Ctrl+Shift+Enter. В столбце H4:H12 появится массив абсолютных частот.
5. В ячейке H13 найдите общее количество наблюдений. Табличный курсор установите в ячейку H13. На панели инструментов Стандартная нажмите кнопку Автосумма. Убедитесь, что диапазон суммирования указан правильно (H4:H12), и нажмите клавишу Enter. В ячейке H13 появится число 55.
6. Заполните столбец относительных частот. В ячейку 14 введите формулу для вычисления относительной частоты: =H4/H$13. Нажмите клавишу Enter. Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон I5:I12. Получим массив относительных частот.
7. Заполните столбец накопленных частот. В ячейку J4 введите вручную значение относительной частоты из ячейки I4 (0,036364). В ячейку J5

введите формулу: =J4 +I5. Нажмите клавишу Enter. Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон J6:12. Получим массив накопленных частот.

8. В результате после форматирования получим таблицу, представленную на рис 1.2.

Наблюдения			Вес	Абсолютные	Относи-	Накопленные
63	61		кг	частоты	тельные	частоты
58	58				частоты	
60	60		57	2	0,036364	0,036364
59	64		58	6	0,109091	0,145454909
60	60		59	7	0,127273	0,272727636
59	59		60	10	0,181818	0,454545818
61	61		61	9	0,163636	0,618182182
62	64		62	8	0,145455	0,763636727
62	62		63	7	0,127273	0,890909455
63	59		64	4	0,072727	0,963636727
57	65		65	2	0,036364	1,000000364
				55		

Рис. 1.2. Результат вычислений относительных и накопленных частот из примера 1.1

9. Постройте диаграмму относительных и накопленных частот. Щелчком указателя мыши по кнопке на панели инструментов вызовите Мастер диаграмм. В появившемся диалоговом окне выберите вкладку Нестандартные и тип диаграммы График/гистограмма2. После нажатия кнопки Далее укажите диапазон данных −I1:J12 (с помощью мыши). Проверьте положение переключателя Ряды в: столбцах. Выберите вкладку Ряд и с помощью мыши введите в рабочее поле Подписи оси X диапазон подписей оси X: G4:G12. Нажав кнопку Далее, введите названия осей X и У: в рабочее поле Ось X (категорий) − Вес; Ось У (значений) — Относ. частота; Вторая ось Y (значений) − Накоплен. частота. Нажмите кнопку Готово.
10. После минимального редактирования диаграмма будет иметь такой вид, как на рис. 1.3.

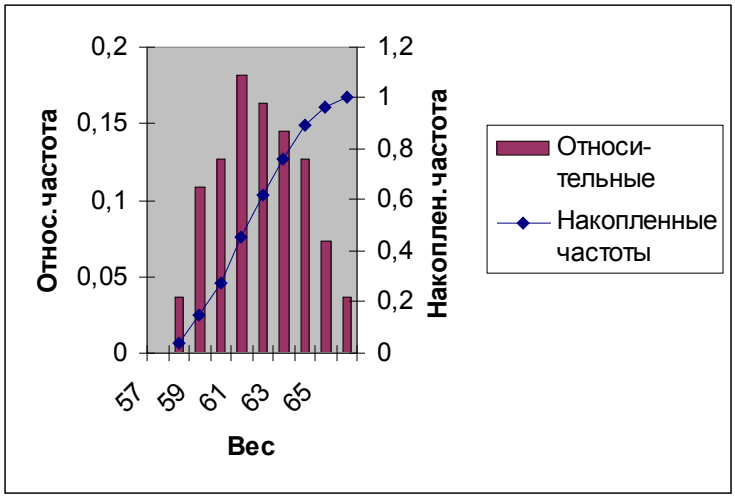


Рис. 1.3. Диаграмма относительных и накопленных частот из примера 1.1

Пример 1.2. Для данных из примера 1.1. построить эмпирические распределения, воспользовавшись процедурой «Гистограмма».

Решение

1. В ячейку A1 введите слово *Наблюдения,* а в диапазон A2:E12 − значений веса студентов.

2. Для вызова процедуры Гистограмма выберите из меню Сервис подпункт Анализ данных и в открываемся окне в поле Инструменты анализа укажите процедуру Гистограмма.

3. В появившемся окне «Гистограмма» заполните рабочие поля (см. рис. 1.1):

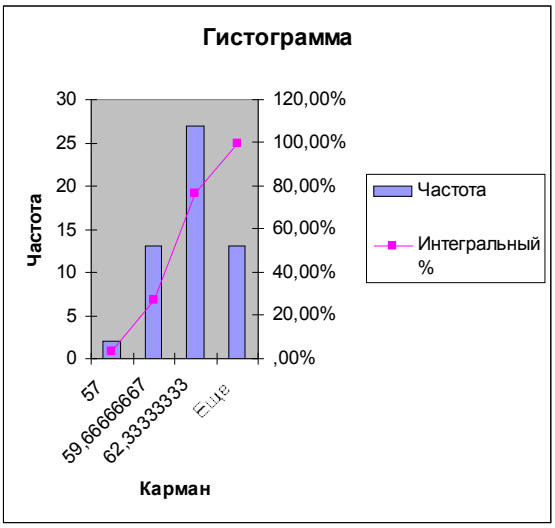
- во «Входной диапазон» введите диапазон исследуемых данных *(A2:E12);*
- в «Выходной диапазон» − ссылку на левую верхнюю ячейку выходного диапазона *(F1).* Установите переключатели в положение Интегральный процент и Вывод графика;

После этого нажмите кнопку ОК.

В результате появляется таблица и диаграмма (рис. 1.4).

Карман	*Частота*	*Интегральный %*
57	2	3,64%
59,66666667	13	27,27%
62,33333333	27	76,36%
Еще	13	100,00%

Рис. 1.4. Таблица и диаграмма из примера 1.2



Как видно, диаграмма на рис. 1.4. несколько отличается от диаграммы на рис. 1.3. Это объясняется тем, что диапазон карманов не был введен. Количество и границы интервалов определялись в процедуре ГИСТОГРАММА автоматически. Если бы в рабочее поле Интервал карманов был бы введен диапазон ячеек, определяющих выбранные интервалы, как в примере 1.1 *(57,58,59,..., 65),* то полученная диаграмма была бы идентична предыдущей.

Упражнения

1. Постройте эмпирические функции распределения (относительные и накопленные частоты)для роста (в см) группы из 20 мужчин: 181, 169,178,178,171,179, 172, 181, 179, 168, 174, 167, 169, 171, 179, 181, 181, 183, 172, 171.

2. Найдите распределение по абсолютным частотам для следующих результатов тестирования в баллах; 79, 85, 78г 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. В соответствии с важнейшими требованиями, оценки должны быть:

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

Среди выборочных характеристик выделяют показатели, относящиеся к центру распределения (меры положения), показатели рассеяния вариант (меры рассеяния) и меры формы распределения. К показателям, характеризующим центр распределения, относят различные виды средних (арифметическое, геометрическое и т.. п.), а также моду и медиану.

Простейшим показателем, характеризующим центр выборки, является мода.

Мода − это элемент выборки с наиболее часто встречающимся значением (наиболее вероятная величина).

Средним значением выборки, или выборочным аналогом математического ожидания, называется величина

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

Иначе говоря, среднее значение − это центр выборки, вокруг которого группируются элементы выборки. При увеличении числа наблюдении среднее приближается к математическому ожиданию. Среднее значение обозначается также буквой *M*.

Выборочная медиана − это число, которое является серединой выборки, то есть половина чисел имеет значения большие, чем медиана, а половина чисел имеет значения меньшие, чем медиана. 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.

Основными показатели рассеяния вариант являются интервал дисперсия выбор

ки, стандартное отклонение и стандартная; ошибка.

Интервал (амплитуда, вариационный размах) − это разница между максимальным и минимальным значениями элементов выборки. Интервал является простейшей и наименее надежной мерой вариации или рассеяния элементов выборке.

Более точно отражают рассеяние показатели, учитывающие не только крайние, но и все значения элементов выборки.

Дисперсией выборки, или выборочным аналогом дисперсии, называется величина

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

Дисперсия выборки − это параметр, характеризующий степень разброса элементов выборки относительно среднего значения. Чем больше дисперсия, тем дальше отклоняются значения элементов выборки от среднего значения.

Выборочным стандартным отклонением (среднее квадратичное отклонение) называется величина

$$s = \sqrt{s^2}$$

Это параметр, также характеризующий степень разброса элементов выборки относительно среднего значения. Чем больше среднее квадратичное отклонение, тем дальше отклоняются значения элементов выборки от среднего значения. Параметр аналогичен дисперсии и 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 или **ошибка среднего** находится из выражения

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

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

Выборочной квантилью называется решение уравнения

$$F_n(x) = p$$

В. частности, выборочная медиана есть решение уравнения

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

Показателями, характеризующими форму распределения, являются выборочные эксцесс и асимметрия.

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

Асимметрия − величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает значения от -1 до 1. В

случае симметричного распределения асимметрия равна 0.

Часто 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 равна нулю, а эксцесс − трем.

Определение основных статистических характеристик

В результате наблюдений или эксперимента получаются 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 (как минимум: 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.

Использование специальных функций

В мастере функций «MS Excel» имеется ряд специальных функций, предназначенных для вычисления выборочных характеристик. Прежде всего, это функции, характеризующие центр распределения.

- Функция СРЗНАЧ вычисляет среднее арифметическое из нескольких массивов (аргументов) чисел. Аргументы *число1, число2, ...* − это от 1 до 30 массивов, для которых вычисляется среднее. Например, если ячейки A1:A7 содержат числа 10,14,5,6,10,12 и 13, то средним арифметическим СРЗНАЧ(*A1:A7)* является 10 (рис. 1.5.).
- Функция СРГАРМ позволяет получить среднее гармоническое множества данных. Среднее гармоническое − это величина, обратная к среднему арифметическому обратных величин. Например, СРГАРМ(*10;14;5;6;10;12;13)* равняется 8,317.
- Функция СРГЕ0М вычисляет среднее геометрическое значений массива положительных чисел. Функцию СРГЕОМ можно использовать для

вычисления средних показателей динамического ряда.

Например, *СРГЕ0М(10;14;5;6;10;12;13)* равняется 9,414.

- Функция МЕДИАНА позволяет получать медиану заданной выборки. Медиана − это элемент выборки, число элементов выборки со значениями больше которого и меньше которого равно. Например, МЕДИАНА(*10;14; 5;6;10;12;13)* равняется 10.

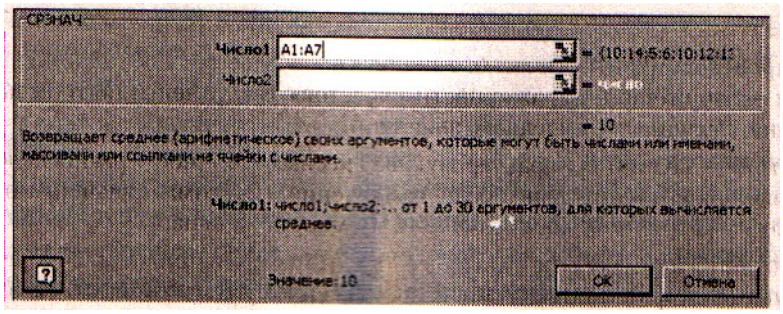
- 

РИС. 1.5. Диалоговое окно функции СРЗНАЧ

- Функция МОДА вычисляет наиболее часто встречающееся значение в выборке. Например, МОДА(*10;14;5;6;10;12;13)* равняется 10.

К специальным функциям, вычисляющим выборочные характеристики, характеризующие рассеяние вариант, относятся ДИСП, СТАНД0ТКЛ0Н, ПЕРСЕНТИЛЬ.

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

О Функция СТАНДОТКЛОН вычисляет стандартное отклонение. Например, СТАНДОКЛОН(10;14;5;6;10;12;13) равняется 3,411.

О Функций ПЕРСЕНТШЬ позволяет получить квантили заданной выборки. Например, если ячейки А1:А7 содер-жат числа 10,14,5,6,10,12 и 13, то квантилю со значением 0,1 является ПЕРСЕНТИЛЬ(А1А7;0,1), равная 5,6

Форму эмпирического распределения позволяют оценить специальные функции ЭКСЦЕСС и СКОС.

О Функция ЭКСЦЕСС вычисляет оценку эксцесса по выборочным данным.

Напри-, мер, ЭКСЦЕСС*(10;14;5;6;10;12;13)* равняется -1,169.

О Функция СКОС позволяет оценить асимметрию выборочного распределения.

Например, СКОС*(10;14;5,-6;10;12:13)* равняется -0,527.

Пример 1.3. Рассматриваются ежемесячные количества реализованных турфирмой путевок за периоды до и после начала активной рекламной компании. Ниже приведены количества реализованных путевок по месяцам.

С рекламой	**Без рекламы**
162	135
156	126
144	115
137	140
125	121
145	112
151	130

Требуется найти средние значения и стандартные отклонения этих данных.

Решение

1. Для проведения статистического анализа прежде всего необходимо ввести данные в рабочую таблицу. Откройте новую рабочую таблицу. Введите в ячейку A1 слово *Реклама,* затем в ячейки A2:A8 − соответствующие значения числа реализованных путевок. В ячейку B1 введите слова *Без рекламы,* а в B2:B8 − значения числа реализованных путевок до начала рекламной компании. Отметим, что рассматриваемые группы данных со статистической точки зрения являются выборками.

2. При статистическом анализе прежде всего необходимо определить характеристики выборки, и важнейшей характеристикой является среднее значение. Для определения среднего значения в контрольной группе необходимо установить табличный курсор в свободную ячейку (A9). На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию СРЗНАЧ, после чего нажмите кнопку ОК. Появившееся диалоговое окно СРЗНАЧ за серое поле мышью отодвиньте вправо на 1-2 см отданных

(при нажатой левой кнопке). Указателем мыши введите диапазон данных контрольной группы для определения среднего значения (A2:A8). Нажмите кнопку ОК. В ячейке А9 появится среднее значение выборки − *145,714.*

В качестве упражнения определите в ячейке В9 среднее значение числа реализованных путевок без активной рекламы. Для этого табличный курсор установите в ячейку В9. На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне выберите категорию Статистические и функцию СРЗНАЧ , после чего нажмите кнопку ОК. Появившееся диалоговое окно СРЗНАЧ за серое поле мышью отодвиньте вправо на 1 -2 см от данных (при нажатой левой кнопке). Указателем мыши введите диапазон данных для определения среднего значения *(B2:B8).* Нажмите кнопку ОК. В ячейке В9 появится среднее значение выборки − *125,571.*

3. Следующей по важности характеристикой выборки является мера разброса

элементов выборки от среднего значения. Такой мерой является среднее квадратичное или стандартное отклонение. Для определения стандартного отклонения в контрольной группе необходимо установить табличный курсор в свободную ячейку (А10). На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию СТАНДОТКЛОН, после чего нажмите кнопку ОК. Появившееся диалоговое окно СТАНДОТКЛОН за серое поле мышью отодвиньте вправо на 1 -2 см от данных (при нажатой левой кнопке). Указателем мыши введите диапазон данных контрольной группы для определения стандартного отклонения *(A2.A8).* Нажмите кнопку ОК. В ячейке А10 появится стандартное отклонение выборки − 12,298. Существует правило, согласно которому при отсутствии артефактов данные должны лежать в диапазоне $M \pm 3\sigma$ (в примере 145,7±36,9).

В качестве упражнения требуется в ячейке В10 определить стандартное отклонение числа проданных путевок до начала рекламной компании. Для этого установите табличный курсор в ячейку В10. На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне выберите категорию

Статистические и функцию СТАНДОТКЛОН, после чего нажмите кнопку ОК. Появившееся диалоговое окно СТАНДОТКЛОН за серое поле мышью отодвиньте вправо на 1−2см от данных (при нажатой левой кнопке). Указателем мыши введите диапазон данных для определения стандартного отклонения *(B2:B8).* Нажмите кнопку ОК. В ячейке В10 появится стандартное отклонение выборки − *10,277.*

Упражнения

1. Найдите среднее значение и стандартное отклонение результатов бега на дистанцию 100 м у группы студентов: 12,8; 13,2; 13,0; 12,9; 13,5; 13,1.
2. Найдите выборочные среднее, медиану, моду, дисперсию и стандартное отклонение для следующей выборки 26,35,29,27,33,35,30,33,31,29.
3. Определите верхнюю (0,75) и нижнюю (0,25) квартили для выборки результатов измерений роста группы студенток: 164, 160, 157, 166, 162, 160, 161, 159, 160, 163, 170, 171.
4. Определите выборочные асимметрию и эксцесс для данных измерений роста из упражнения 1.

Использование инструментов Пакета анализа

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

Для установки раздела Анализ данных в пакете «MS Excel» сделайте следующее:

О в меню Сервис выберите команду «Надстройки»;

О в появившемся списке установите флажок «Пакет анализа».

Ввод данных. Исследуемые данные следует представить в виде таблицы, где столбцами являются соответствующие показатели. При создании таблицы «MS Excel» информация вводится в отдельные ячейки. Совокупность ячеек, содержащих анализируемые данные, называется входным диапазоном.

Последовательность обработки данных. Для использования статистического

пакета анализа данных необходимо:

О указать курсором мыши на пункт меню «Сервис» и щелкнуть левой кнопкой мыши;

О в раскрывающемся списке выбрать команду «Анализ данных» (если команда Анализ данных отсутствует в меню «Сервис», то необходимо установить в «MS Excel» пакет анализа данных входя в «Сервис» → «Надстройки…» → «Пакет анализа»);

О выбрать необходимую строку в появившемся списке «Инструменты анализа»;

О ввести входной и выходной диапазоны и выбрать необходимые параметры.

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. Для выполнения процедуры необходимо:

- выполнить команду Сервис →Анализ данных;
- в появившемся списке Инструменты анализа выбрать строку Описательная статистика и нажать кнопку ОК (рис. 1.6);

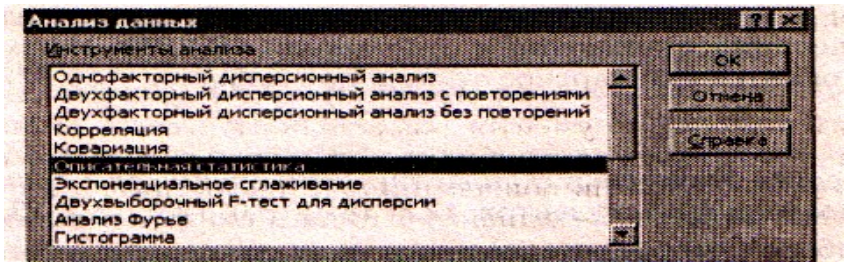


Рис. 1.6 Окно выбора метода обработки данных

- в появившемся диалоговом окне указать входной диапазон, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Для этого следует навести указатель мыши на левую верхнюю ячейку данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к правой нижней ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые

будут выведены результаты анализа. Для этого следует поставить переключатель в положение Выходной диапазон (навести указатель мыши и щелкнуть левой клавишей), далее навести указатель мыши в поле ввода Выходной диапазон и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши;

- в разделе «Группировка» переключатель установить в положение по столбцам;
- установить флажок в поле «Итоговая статистика»;
- нажать кнопку ОК.

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

Пример 1.4. Рассматривается зарплата основных групп работников гостиницы: администрации, обслуживающего персонала и работников ресторана. Были получены следующие данные:

Администрация	Персонал	Ресторан
4500	2100	3200
4000	2100	3000
3700	2000	2500
3000	2000	2000
2500	2000	1900
	1900	1800
	1800	
	1800	

Необходимо определить основные статистические характеристики в группах данных.

Решение

1. Для использования инструментов анализа исследуемые данные следует представить в виде таблицы, где столбцами являются соответствующие показатели. Значения зарплат сотрудников администрации введите в диапазон A1:A5, обслуживающего персонала − в диапазон B1:B8 и т. д. В результате получится таблица, представленная на рис. 1.7.

	A	B	C
1	450	210	320
2	400	210	300
3	370	200	250
4	300	200	200
5	250	200	190
6		190	180
7		180	
8		180	

Рис. 1.7. Таблица из примера 1.4

Далее необходимо провести элементарную статистическую обработку. Для этого, указав курсором мыши на пункт меню Сервис, выберите команду Анализ данных. Затем в появившемся списке Инструменты анализа выберите строку Описательная статистика.

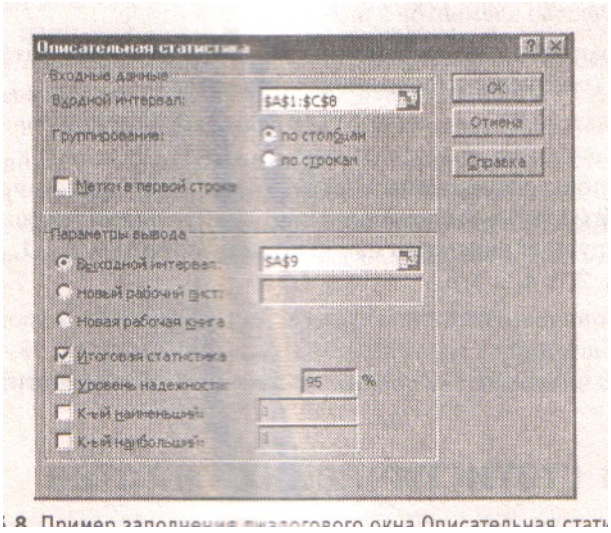


Рис. 1.8. Пример заполнения диалогового окна Описательная статистика

3. В появившемся диалоговом окне (рис. 1.8) в рабочем поле Входной интервал укажите входной диапазон — *A 1:C8.* Активировав переключателем рабочее поле Выходной интервал, укажите выходной диапазон − ячейку *A9.* В разделе Группировка

переключатель установите в положение по столбцам. Установите флажок в поле Итоговая статистика и нажмите кнопку ОК.

В результате анализа (рис. 1.9) в указанном выходном диапазоне для каждого столбца данных получим соответствующие результаты.

Столбец1		Столбец2		Столбец3	
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. 1.9. Результаты работы инструмента «Описательная статистика»

Все полученные характеристики были рассмотрены ранее в разделе «Выборочные характеристики» данной главы, за исключением последних четырех:

О минимум − значение минимального элемента выборки;

О максимум − значение максимального элемента выборки;

О сумма − сумма значений всех элементов выборки;

О счет − количество элементов в выборке.

Среди этих характеристик наиболее важными являются показатели Среднее, Стандартная ошибка (среднего) и Стандартное отклонение.

Упражнения

8. Найдите наиболее популярный туристический маршрут из четырех реализуемых фирмой (моду), если за неделю последовательно были реализованы следу-

ющие маршруты (приводятся номера маршрутов); 1,3,3,2,1, 1,1, 4,4,2,4,1,3,2, 4,1,4,4,3,1,2,3,4,1,1,3.

9. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг/м^3): 12, 16, 15, 14, 10,20,16, 14, 18, 14, 15, 17, 23, 11. Необходимо определить основные выборочные характеристики.

Проверка статистических гипотез

Помимо описательной статистики, важной областью является также аналитическая статистика. Как уже указывалось в разделе «Понятие математической статистики», аналитическая статистика или теория статистических выводов ориентирована на обработку данных, полученных в ходе эксперимента, с целью формулировки выводов, имеющих прикладное значение. Здесь решается вопрос, отражают ли наблюдаемые данные объективно существующую реальность. Указанный вопрос решается проверкой соответствующих статистических гипотез. При этом могут выявляться достоверности различий между выборками, взаимосвязи между выборками, влияющие факторы и т. п.

Принятие статистических решений

Статистическая гипотеза − это предположение о виде или отдельных параметрах распределения вероятностей, которое подлежит проверке на имеющихся данных.

Проверка статистических гипотез − это процесс формирования решения о возможности принять или отвергнуть утверждение (гипотезу), основанный на информации, полученной из анализа выборки. Методы проверки гипотез называются критериями.

В большинстве случаев рассматривают так называемую нулевую гипотезу (нуль-гипотезу H_0), состоящую в том, что все события произошли случайно, естественным образом. Альтернативная гипотеза (H_1) состоит в том, что события случайным образом произойти не могли, и имело место воздействие некого фактора.

Обычно нулевая гипотеза формулируется таким образом, чтобы на основании эксперимента или наблюдений ее можно было отвергнуть с заранее заданной веро-

ятностью ошибки а. Эта, заранее заданная вероятность ошибки, называется уровнем значимости.

Уровень значимости − максимальное значение вероятности появления события, при котором событие считается практически невозможным. В статистике наибольшее распространение получил уровень значимости, равный $\alpha = 0{,}05$. Поэтому если вероятность, с которой интересующее событие может произойти случайным образом $p < 0{,}05$, то принято считать это событие маловероятным, и если оно все же произошло, то это не было случайным. В наиболее ответственных случаях, когда требуется особая уверенность в достоверности полученных результатов, надежности выводов, уровень значимости принимают равным а = 0,01 или даже а = 0,001.

Величину *P,* равную 1 - α, называют доверительной вероятностью (уровнем надежности), 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.

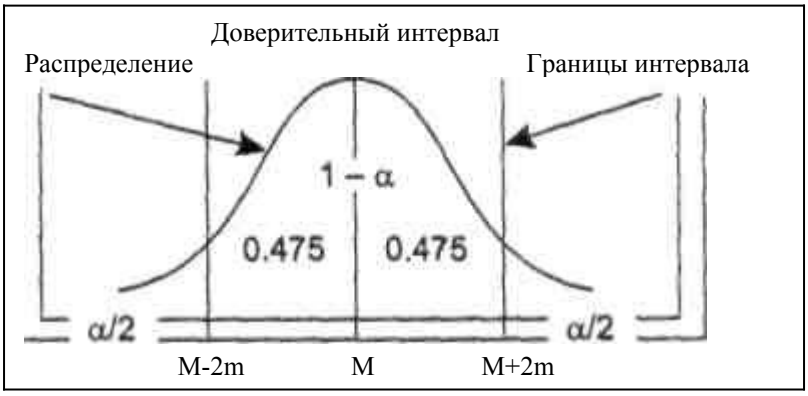


Рис. 1.10. 95%-ный доверительный интервал для среднего значения

Интервал, в котором с заданной доверительной вероятностью $P = 1 - \alpha$ находится оцениваемый параметр, называется доверительным интервалом. В соответствии с доверительными вероятностями на практике используются 95%-, 99%-, 99,9%-ные доверительные интервалы. Граничные точки доверительного интервала называют доверительными пределами (рис. 1.10).

Выбор того или иного уровня значимости, выше которого результаты отвергаются как статистически не подтвержденные, или, соответственно, доверительной вероятности, в общем случае является произвольным. Окончательное

решение зависит от исследователя, традиций и накопленного практического опыта в данной области исследований.

Анализ одной выборки

Анализ однородности выборки. Одним из важных вопросов, возникающих при анализе выборки, является вопрос: относится та или иная варианта к данной статистической совокупности? Решение вопроса не представляет сложности, если распределение в этой совокупности является нормальным. Для этого достаточно использовать правило трех сигм. Согласно этому правилу, в пределах $M \pm 3\sigma$ находится 99,7% всех вариант. Поэтому, если варианта попадает в этот интервал, то она считается принадлежащей к данной совокупности. Если не попадает, то она может быть отброшена. Хотя этот метод и предполагает нормальность исходного распределения, на практике он успешно работает и может быть использован в большинстве других случаев.

При числе элементов в выборке $n < 30$ способ более точного определения границ доверительного интервала по формуле

$$[M - t_{n,p}\, s; \quad M + t_{n,p}\, s] \tag{1.1}$$

будет показан ниже в примере 1.5. В формуле (1.1) M − среднее значение, s − стандартное отклонение, $t_{n,p}$ − табличное значение распределения Стьюдента с числом степеней свободы n и доверительной вероятностью p.

Построение доверительных интервалов для среднего. Еще одной 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, значимо ли отличие выборочного среднего значения от среднего значения генеральной совокупности, из которой предположительно взята выборка, или наблюдаемое различие является случайным.

Действительно, средние значения, получаемые по выборочным данным, обычно

не совпадают с генеральным средним (математическим ожиданием). 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 характер нормального распределения. Это значит, что в случае средних арифметических значения выборочных средних относительно генерального среднего распределяются по нормальному закону. То есть так же, как относительные отклонения нормально распределенных вариант от среднего арифметического выборки.

Отсюда, в частности, следует, что 68,3% всех выборочных средних находятся в пределах $\Delta = M \pm т$, где Δ − предельная ошибка выборки, M − среднее выборочное, т − стандартная ошибка среднего значения. Иными словами, имеется вероятность 0,683, что выборочное среднее отличается от генерального не более, чем на ±т. Здесь 0,683 − доверительная вероятность, 1 - 0,683 = 0,317 − уровень значимости а, $\Delta = M \pm т$ − 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 ±2т (см. рис. 1.8). При доверительной вероятности 0,99, доверительный интервал составит примерно ±3т. Для более точного определения границ доверительного интервала можно воспользоваться формулой

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

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.

В MS «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(алъфа; станд-откл;размер) определяет полуширину доверительного интервала и содержит следующие параметры:

О Альфа − уровень значимости, используемый для вычисления доверительной вероятности. Доверительная вероятность равняется 100*(1 - альфа)% процентам, или, другими словами, альфа, равное 0,05, означает 95%-ный уровень доверительной вероятности;

О Станд-откл − стандартное отклонение генеральной совокупности для интервала данных, предполагается известным;

О Размер − это размер выборки.

Пример 1.5. Найти границы 95%-ного доверительного интервала для среднего значения, если у 25 телефонных аккумуляторов среднее время разряда в режиме ожидания составило 140 часов, а стандартное отклонение − 2,5 часа.

Решение

1. Откройте новую рабочую таблицу. Установите табличный курсор в ячейку A1.

2. Для определения границ доверительного интервала необходимо на панели инструментов Стандартная нажать кнопку Вставка функции (fx). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ДОВЕРИТ, после чего нажмите кнопку ОК.

3. В рабочие поля появившегося диалогового окна ДОВЕРИТ с клавиатуры введите условия задачи; Альфа − *0,05;* Станд-откл −*2,5;* Размер −*25* (рис. 1.11). Нажмите кнопку ОК.

4. В ячейке A1 появится полуширина 95%-ного доверительного интервала для среднего значения выборки — 0,979981. Другими словами, с 95%-ным уровнем надежности можно утверждать, что средняя продолжительность разряда аккумулятора составляет 140 ± 0,979981 часа или от 139,02 до 140,98 часа.

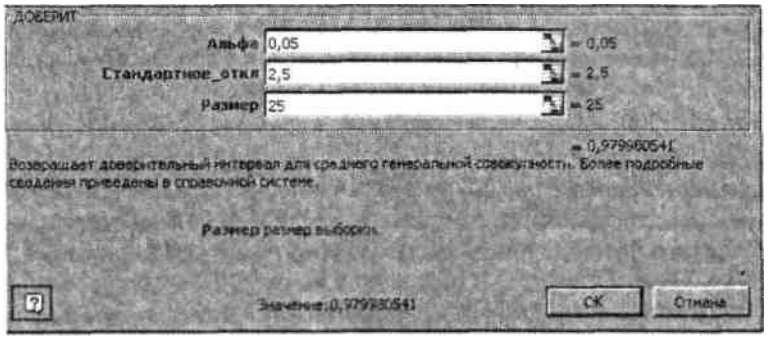


Рис. 1.1. Пример заполнения диалогового окна «ДОВЕРИТ»

Пример 1.1. Пусть имеется выборка, содержащая числовые значения: 13, 15, 17, 19,22,25,19. Необходимо определить границы 95%-ного доверительного интервала для среднего значения и для нахождения «выскакивающей» варианты.

Решение

1. В диапазон A1:A7 введите исходный ряд чисел.

2. Далее вызовите процедуру Описательная статистика. Для этого, указав курсором мыши на пункт меню Сервис, выберите команду Анализ данных. Затем в появившемся списке Инструменты анализа выберите строку Описательная статистика.

3. В появившемся диалоговом окне в рабочем поле Входной интервал: укажите входной диапазон − *A 1:A7.* Переключателем активизируйте Выходной интервал и укажите выходной диапазон — ячейку *B1.* В разделе Группировка переключатель установите в положение по столбцам. Установите флажок в левое поле Уровень надежности: и в правом поле (%) −*95.* Затем нажмите кнопку ОК.

4. В результате анализа в указанном выходном диапазоне для доверительной вероятности 0,95 получаем значения доверительного интервала (рис. 1.12).

A	B	
1 3	*Столбец 1*	
1 5		
1 6 1 7	Уровень надежности (95.0%)	3,77027204 6

Рис. 1.12. Исходная выборка (A1:A7) и результат вычислений (C3) из примера 1.6

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 ± 3,77. Здесь 18,571 − выборочное среднее M для рассматриваемого примера, которое находится обычно процедурой Описательная статистика одновременно с доверительным интервалом.

5. Для нахождения доверительных границ для «выскакивающей» варианты необходимо полученный выше доверительный интервал умножить на $\sqrt{n}$ (в примере $\sqrt{7}$, то есть 3,77* $\sqrt{7}$ =9,975). В «MS Excel» это можно выполнить следующим образом. Табличный курсор установите в свободную ячейку C4; введите с клавиатуры знак =; мышью укажите на ячейку C3 (в которой находится результат вычислений); введите с клавиатуры знак *; с панели инструментов Стандартная вызовите Мастер функций (кнопка fx); выберите категорию Математические, тип функции Корень; нажмите ОК; введите с клавиатуры число 7 и нажмите ОК. В результате получим в ячейке C4 значение доверительного интервала − 9,975.

Таким образом, варианта, попадающая в интервал 18,571 ± 9,975, считается принадлежащей данной совокупности с вероятностью 0,95. Выходящая за эти границы может быть отброшена с уровнем значимости $\alpha = 0{,}05$.

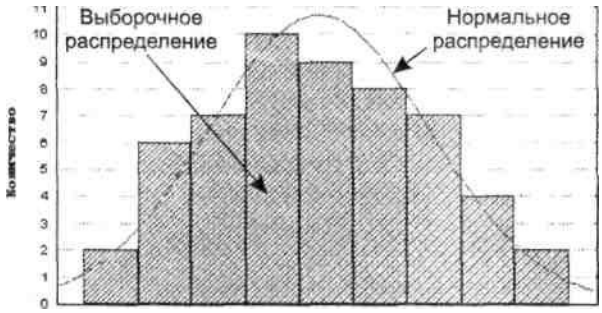
Проверка соответствия теоретическому распределению. Следующей задачей, возникающей при анализе одной выборки, является оценка меры соответствия (расхождения) полученных эмпирических данных и каких-либо теоретических распределений. Это связано с тем, что в большинстве случаев при решении реальных задач 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, поскольку большинство статистических процедур ориентировано на выборки, полученные из нормально распределенной 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 (рис. 1.13).



56 57 58

59 60 61 62 63 64 65

Вес

Рис. 1.13. Сопоставление выборочного распределения веса студенток и кривой нормального распределения

При большом числе наблюдений (n > 100) неплохие результаты дает вычисление выборочных параметров формы распределения: эксцесса и асимметрии (см. разделы «Использование специальных функций» и «Использование инструментов Пакета анализа»). Принято говорить, что предположение о нормальности 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.

Наиболее убедительные результаты дает использование критериев согласия. Критериями согласия 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 (хи-квадрат). Он основан на сравнении эмпирических частот интервалов группировки с теоретическими (ожидаемыми) частотами, рассчитанными по формулам нормального распределения.

Отметим, что сколько-нибудь уверенно о нормальности закона распределения можно судить, если имеется не менее 50 результатов наблюдений. В случаях меньшего числа данных можно говорить только о том, что данные не противоречат нормальному закону, и в этом случае обычно используют графические методы оценки соответствия. При большем числе наблюдений целесообразно совместное использование графических и статистических (например, тест хи-квадрат или аналогичные) методов оценки, естественно дополняющих друг друга.

Использование критерия согласия хи-квадрат. Для применения критерия желательно, чтобы объем выборки $n > 40$, выборочные данные были сгруппированы в интервальный ряд с числом интервалов не менее 7, а в каждом интервале находилось не менее 5 наблюдений (частот).

Отметим, что сравниваться должны именно абсолютные частоты, а не относительные (частоты). При этом, как и любой другой статистический критерий, критерий хи-квадрат не доказывает справедливость нулевой гипотезы (соответствие эмпирического распределения нормальному), а лишь может позволить ее отвергнуть с определенной вероятностью (уровнем значимости).

В MS «MS Excel» критерий хи-квадрат реализован в функции ХИ2ТЕСТ. Функция ХИ2-ТЕСТ вычисляет вероятность совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) 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.

Функция имеет следующие параметры:

ХИ2ТЕСТ (фактический_интервал;ожидаемый_интервал). Здесь:

О *фактический_интервал* − это интервал данных, которые содержат наблюдения, подлежащие сравнению с ожидаемыми значениями;

О *ожидаемый интервал* − это интервал данных, который содержит теоретические (ожидаемые) значения для соответствующих наблюдаемых.

Пример 1.7. Проверить соответствие выборочных данных из примера 1.1. (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) нормальному закону распределения.

Решение

1. Повторите пункты 1-7 решения примера 1.1. В результате получится таблица (см. рис. 1.2).

2. Найдите теоретические частоты нормального распределения. Для этого предварительно необходимо найти среднее значение и стандартное отклонение выборки.

В ячейке I13 с помощью функции СРЗНАЧ найдите среднее значение для данных из диапазона A2:E12 (60,855). В ячейке J13 с помощью функции СТАНД0ТКЛ0Н найдите стандартное отклонение для этих же данных (2,05). В ячейки К1 и К2 введите название столбца − *Теоретические частости.* Затем с помощью функции НОРМ-РАСП найдите теоретические частости. Установите курсор в ячейку К4, вызовите указанную функцию и заполните ее рабочие поля: *x* − *G4;* Среднее − *$1$13;* Стан-дартное_откл − *J13;* Интегральный − *0.* Получим в ячейке К4 *0,033.* Далее протягиванием скопируйте

содержимое ячейки К4 в диапазон ячеек К5:К12. Затем в ячейки L1 и L2 введите название нового столбца − *Теоретические частоты.* Установите курсор в ячейку L4 и введите формулу *=H$13*K4.* Далее протягиванием скопируйте содержимое ячейки L4 в диапазон ячеек L5:L12. Результаты вычислений представлены на рис. 1.14.

G	H	I	J	K	L
Вес кг	Абсолютные частоты	Относи-тельные частоты	Накопленные частоты	Теоретические частости	Теоретические частоты
57	2	0,036364	0,036364	0,033205828	1,82632055
58	6	0,109091	0,145454909	0,073795567	4,058756212
59	7	0,127273	0,272727636	0,129258576	7,109221655
60	10	0,181818	0,454545818	0,178443849	9,814411704
61	9	0,163636	0,618182182	0,194158732	10,67873029
62	8	0,145455	0,763636727	0,16650428	9,157735407
63	7	0,127273	0,890909455	0,112540024	6,189701326
64	4	0,072727	0,963636727	0,059951732	3,297345259
65	2	0,036364	1,000000364	0,025171529	1,384434082

Рис. 1.14. Результаты вычисления теоретических частостей и частот из примера 1.7

С помощью функции ХИ2ТЕСТ определите соответствие данных нормальному закону распределения. Для этого установите табличный курсор в свободную ячейку L13. На панели инструментов Стандартная нажмите кнопку Вставка функции (f_x). В

появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ХИ2ТЕСТ, после чего нажмите кнопку ОК. Появившееся диалоговое окно ХИ2ТЕСТ отодвиньте вправо на 1-2 см от данных. Указателем мыши в рабочие поля введите фактический H4:H12 и ожидаемые L4:L12 диапазоны частот (рис. 1.15). Нажмите кнопку ОК. В ячейке L13 появится значение вероятности того, что выборочные данные соответствуют нормальному закону распределения − 0,9842.

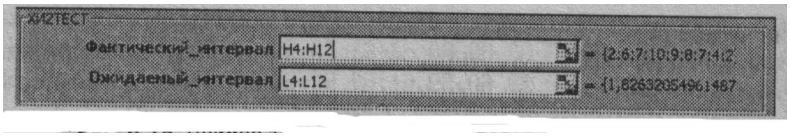


Рис. 1.15. Пример заполнения рабочих полей функции ХИ2ТЕСТ

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

Упражнения

10. 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.

11. Определите с уровнем значимости α = 0,05 максимальное отклонение среднего значения генеральной совокупности от среднего выборки 3, 4, 4, 2, 5, 3, 4, 3, 5, 4, 3, 5, 1.

12. Найдите соответствие экспериментальных данных нормальному закону распределения для следующей выборки весов детей (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, 25, 25, 25, 21, 21, 21, 21, 21, 21, 21, 21, 21, 26, 26,

26, 26, 26, 27, 27.

Анализ двух выборок

Выявление достоверности различий. Следующей задачей статистического анализа, решаемой после определения основных выборочных характеристик и анализа одной выборки, является совместный анализ нескольких выборок. Важнейшим вопросом, возникающим при анализе двух выборок, является вопрос о наличии различий между этими выборками. Обычно для этого проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве генеральных средних. В рассмотренном ранее примере 1.3. такие различия выявляются путем сравнения данных реализации турфирмой путевок за периоды до и после начала активной рекламной компании. Если сопоставить средние значения числа реализованных за месяц путевок до (125,6) и после (145,7) начала рекламной компании, видно, что они различаются. Можно ли по этим данным сделать вывод об эффективности рекламной компании?

Для решения задач такого типа используются так называемые критерии различия. Для проверки одной и той же гипотезы могут быть использованы разные статистические критерии. Правильный выбор критерия определяется как спецификой данных и проверяемых гипотез, так и уровнем статистической подготовки исследователя.

Статистические критерии различия подразделяются на параметрические и непараметрические критерии. Параметрические критерии служат для проверки гипотез о параметрах определенных распределений генеральной совокупности (чаще всего нормального распределения). Непараметрические критерии для проверки гипотез не используют предположений о законе распределения генеральной совокупности и не требуют знания параметров распределения.

Параметрические критерии. Параметрические критерии служат для проверки гипотез о положении и рассеивании. Из параметрических критериев наибольшей популярностью при проверке гипотез о равенстве генеральных средних (математических ожиданий) пользуется f-критерий **Стьюдента** (t-критерий различия).

Критерий Стьюдента (t) наиболее часто используется для проверки гипотезы: «Средние двух выборок относятся к одной и той же совокупности». Критерий позволяет найти вероятность того, что оба средних относятся к одной и той же совокупности. Если эта вероятность p ниже уровня значимости ($p < 0,05$), то принято считать, что выборки относятся к двум разным совокупностям.

При использовании t-критерия можно выделить два случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть контрольная группа и опытная группа, состоящие, например, из разных пациентов, количество которых в группах может быть различно.

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних, используется так называемый парный t-критерий. Выборки при этом называют зависимыми, связанными. Например, измеряется содержание лейкоцитов у здоровых животных, а затем у тех же самых животных после облучения определенной дозой излучения.

В обоих случаях в принципе должно выполняться требование нормальности распределения исследуемого признака в каждой из сравниваемых групп и равенства дисперсий в сравниваемых совокупностях. Однако на практике по большому счету корректное применение t-критерия Стьюдента для двух групп часто бывает затруднительно, поскольку достоверно проверить эти условия удается далеко не всегда.

Для оценки достоверности отличий по критерию Стьюдента принимается нулевая гипотеза, что средние выборок равны между собой. Затем вычисляется значение вероятности того, что изучаемые события (например, количества реализованных путевок в обеих выборках) произошли случайным образом.

В MS «MS Excel» для оценки достоверности отличий по критерию Стьюдента используются специальная функция «ТТЕСТ» и процедуры пакета анализа (см. раздел «Использование Пакета анализа для выявления различий» ниже).

Все перечисленные инструменты вычисляют вероятность, соответствующую критерию Стьюдента, и используются, чтобы определить, насколько вероятно, что

две выборки взяты из генеральных совокупностей, которые имеют одно и то же среднее.

Функция «ТТЕСТ» использует следующие параметры: ТТЕСТ (массив1; массив2; хвосты; -тип). Здесь:

О *массив 1* − это первое множество данных;

О *массш2* − это второе множество данных;

О *хвосты* − число хвостов распределения. Обычно число хвостов равно 2;

О *тип* − это вид исполняемого *t*-теста. Возможны 3 варианта выбора: 1 − парный тест, 2 − двухвыборочный тест с равными дисперсиями, 3 − двухвыборочный тест с неравными дисперсиями.

Пример 1.8. Выявить, достоверны ли отличия при сравнении данных реализации турфирмой путевок за периоды до и после начала активной рекламной компании (см. пример 1.3).

Решение

1. Введите данные (как в пункте 1 примера 1.3).

1. Для выявления достоверности отличий табличный курсор установите в свободную ячейку (А11). На панели инструментов необходимо нажать кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ТТЕСТ, после чего нажмите кнопку ОК. Появившееся диалоговое окно ТТЕСТ за серое поле мышью отодвиньте вправо на 1-2 см от данных (при нажатой левой кнопке). Указателем мыши введите диапазон данных контрольной группы в поле Массив 1 *(A2:A8).* В поле Массив 2 введите диапазон данных исследуемой группы *(B2:B8).* В поле Хвосты всегда вводится с клавиатуры цифра *2* (без кавычек), а в поле Тип с клавиатуры введите цифру *3*. Нажмите кнопку ОК. В ячейке А11 появится значение вероятности - *0,006295*.

2. Поскольку величина вероятности случайного появления анализируемых выборок (0,006295) меньше уровня значимости (α = 0,05), то нулевая гипотеза отвергается. Следовательно, различия между выборками не случайные и средние

выборок считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия Стьюдента можно сделать вывод о большей эффективности реализации путевок после начала рекламной компании *(p* < 0,05).

Как указывалось выше, при использовании t-критерия выделяют два основных случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть две различных выборки, количество элементов в которых может быть также различно. При заполнении диалогового окна TTECT при этом указывается Тип *3.*

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних, используется так называемый парный t-критерий. Выборки при этом называют зависимыми, связанными (при заполнении диалогового окна TTECT указывается Тип 1*).* Например, сравнивается реализация путевок двумя фирмами в соответствующие месяцы. В качестве упражнения рассмотрим пример.

Пример 1.9. Сравнивается количество наличных денег у двух групп студентов (в рублях);

A	**B**
30	10
30	20
40	30
50	40
60	50

Необходимо определить достоверность различия между группами при двух вариантах постановки задачи:

О группы состоят из различных студентов (тип 3);

О группы состоят из одних и тех же студентов, но первая − до посещения буфета,

а вторая − после (тип 1).

Решение. В ячейки C1:C5 введите количество денег у студентов первой группы. В ячейки D1:D5 введите количество денег у студентов второй группы,

1. Табличный курсор установите в свободную ячейку (C6). На панели инструментов необходимо нажать кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ТТЕСТ, после чего нажмите кнопку ОК. Появившееся диалоговое окно ТТЕСТ за серое поле мышью отодвиньте вправо на 1-2 см от данных (при нажатой левой кнопке). Указателем мыши ввести диапазон данных первой группы в поле Массив 1 *(C1:C5).* В поле Массив 2 введите диапазон данных второй группы *(D1:D5).* В поле Хвосты всегда вводится цифра *2* (без кавычек), а в поле Тип введите цифру *3.* Нажмите кнопку ОК. В ячейке C6 появится значение вероятности - *0,228053.*

Поскольку величина вероятности случайного появления анализируемых выборок (0,228053) больше уровня значимости (α = 0,05), то нулевая гипотеза не может быть отвергнута (принимается). Следовательно, различия между выборками могут быть случайными и средние выборок не считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия Стьюдента нельзя сделать вывод о достоверности отличий двух групп студентов по количеству карманных денег, имеющихся у них *(p* > 0,05).

3. Табличный курсор установите в свободную ячейку (D6). На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ТТЕСТ, после чего нажмите кнопку ОК. Появившееся диалоговое окно ТТЕСТ за серое поле мышью отодвиньте вправо на 1-2 см от данных (при нажатой левой кнопке). Указателем мыши введите диапазон данных первой группы в поле Массив 1 *(C1:C5).* В поле Массив 2 введите диапазон данных второй группы *(D1:D5).*В поле Хвосты всегда вводится цифра *2* (без кавычек), а в поле Тип введите цифру 1. Нажмите кнопку ОК. В ячейке D6 появится значение вероятности *−0,003883.*

Поскольку величина вероятности случайного появления анализируемых

выборок (0,003883) меньше уровня значимости (α = 0,05), то нулевая гипотеза отвергается. Следовательно, различия между выборками не могут быть случайными и средние выборок считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия Стьюдента можно сделать вывод о том, что в двух группах студентов выявлены достоверные отличия по количеству карманных денег ($p < 0,05$), что явилось результатом посещения буфета.

Таким образом, ясно, что применение различных типов критерия Стьюдента может приводить к различным результатам на основании одних и тех же исходных данных. Можно предложить следующий приблизительный способ выбора типа критерия: если не ясно, какой тип критерия выбирать, выбирается тип 3; если очевидно, что выборки зависимы, связаны (например, это одни и те же студенты), то следует выбирать тип 1.

Критерий Фишера. Критерий Фишера используют для проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности и, следовательно, их равенстве. При этом предполагается, что данные независимы и распределены по нормальному закону. Гипотеза о равенстве дисперсий принимается, если отношение большей дисперсии к меньшей меньше критического значения распределения Фишера.

$$F=s_1^2/s_2^2, \quad F<F_{крит}$$

где Fкрит зависит от уровня значимости и числа степеней свободы для дисперсий в числителе и знаменателе.

В MS «MS Excel» для расчета уровня вероятности выполнения гипотезы о равенстве дисперсий могут быть использованы функция ФТЕСТ(массив1; массив2) и процедура пакета анализа Двухвыборочный F-тест для дисперсий.

Непараметрические критерии. Непараметрические критерии используются в тех случаях, когда закон распределения данных отличается от нормального или неизвестен. Из большого числа непараметрических критериев рассмотрим критерий хи-квадрат.

Критерий согласия χ^2 - Бывают ситуации, когда необходимо сравнить две относительные или выраженные в процентах величины (доли). Примером может слу-

жить случай проверки успешности трудоустройства молодых специалистов, когда известен процент трудоустроившихся выпускников двух институтов. Для проверки достоверности различий здесь критерий Стьюдента применить не удастся. В таких задачах обычно используют критерий χ^2 (хи-квадрат). Критерий хи-квадрат относится к непараметрическим критериям.

Здесь, как и в случае с критерием Стьюдента, принимается нулевая гипотеза о том, что выборки принадлежат к одной генеральной совокупности. Кроме того, определяется ожидаемое значение результата. Обычно это среднее значение между выборками рассматриваемого показателя. Затем оценивается вероятность того, что ожидаемые значения и наблюдаемые принадлежат к одной генеральной совокупности.

В MS «MS Excel» критерий хи-квадрат реализован в функции ХИ2ТЕСТ. Функция ХИ2-ТЕСТ вычисляет вероятность совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости (0,05), то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют теоретическим (ожидаемым) значениям.

Функция имеет следующие параметры: *ХИ2ТЕСТ(фактический_интервал; ожидаемый_интервал)*. Здесь:

О *фактический^интервал* − это интервал данных, которые содержат наблюдения, подлежащие сравнению с ожидаемыми значениями;

О *ожидаемый_интервал* − это интервал данных, который содержит теоретические (ожидаемые) значения для соответствующих наблюдаемых.

Пример 1.10. Пусть после окончания двух институтов экономического профиля трудоустроилось по специальности из первого института 90 человек, а из второго 60 (обе группы молодых специалистов включали по 100 человек).

Решение

1. Принимается нулевая гипотеза, что выборки принадлежат к одной генеральной совокупности.

2. Определяется ожидаемое значение результата (среднее значение между выборками): (60 + 90)/2 = 75, то есть мы ожидали, что разницы между группами нет, и

в обоих случаях должно было трудоустроиться по 75 человек.

3. Затем вычисляется значение вероятности того, что изучаемые события (трудоустройство в обеих выборках) произошли случайным образом. Для этого введите данные в рабочую таблицу: 90 − в ячейку E1, 60 − в F1, 75 − в E2,F2. Табличный курсор установите в свободную ячейку (E3). На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию ХИ2ТЕСТ, после чего нажмите кнопку ОК. Появившееся диалоговое окно ХИ2ТЕСТ за серое поле мышью отодвиньте вправо на 1-2 см от данных (при нажатой левой кнопке). Указателем мыши введите диапазон данных наблюдавшегося количества трудоустроившихся в поле Фактический интервал (E1:F1). В поле Ожидаемый интервал введите диапазон данных предполагаемого количества трудоустроившихся *(*E2:F2*)*. Нажмите кнопку ОК. В ячейке E3 появится значение вероятности − *0,014301*.

4. Поскольку величина вероятности случайного появления анализируемых выборок (0,0143) меньше уровня значимости (a = 0,05), то нулевая гипотеза отвергается. Следовательно, различия между выборками не могут быть случайными и выборки считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия хи-квадрат можно сделать вывод о том, что в двух группах выпускников выявлены достоверные отличия по успешности трудоустройства *(p* < 0,05), что, по-видимому, явилось результатом более высокой репутации выпускников первого института.

Упражнения

13. Даны результаты бега на дистанции 100 м в секундах в двух группах студентов. Студенты первой группы в течение года посещали факультативные занятия по физкультуре. Определите, достоверны ли отличия по результатам бега в этих группах.

Посещавшие факультатив	**Не посещавшие**

12,3	13,2
11,9	13,0
12,2	12,9
12,4	13,1
13,0	13,5
12,6	12,8

	Да	Нет	Не помню
Мужчины	58	11	10
Женщины	35	25	23

в

14. В ходе социологического опроса на вопрос о перенесенном детстве заболевании ответы распределились следующим образом:

Есть ли достоверные отличия в ответах женщин и мужчин?

15. Приведены данные ежемесячной результативности (количество голов) футбольной команды в двух сезонах:

Месяц	3	4	5	6	7	8	9	10	11
2000 г.	3	4	5	8	9	1	2	4	5
2001 г.	6	19	3	2	14	4	5	17	1

Определите, есть ли статистические различия в ежемесячной результативности команды в рассматриваемых сезонах?

10. Определите, имеют ли выборки {6; 7; 9; 15; 21} и {20; 28; 31; 38; 40} различные уровни разнородности (отличаются ли дисперсии)?

Использование инструмента Пакет анализа для выявления различий между

выборками

Для анализа двух выборок с помощью t-теста Стьюдента могут быть использованы следующие процедуры: Парный двухвыборочный t-тест для средних; Двухвыборочный t-тест с одинаковыми дисперсиями и Двухвыборочный t-тест с различными дисперсиями. Как указывалось в разделе «Анализ двух выборок», в общем случае необходимо воспользоваться процедурой Двухвыборочный t-тест с различными дисперсиями, так как процедуры Парный двухвыборочный t-тест для средних и Двухвыборочный t-тест с одинаковыми дисперсиями относятся к частным, специальным случаям.

Для выполнения процедуры анализа необходимо:

О выполнить команду «Сервис» → «Анализ данных»;

О в появившемся списке Инструменты анализа выбрать строку Двухвыборочный t-тест с различными дисперсиями, щелкнуть левой кнопкой мыши и нажать кнопку 0К;

О в появившемся диалоговом окне указать Интервал переменной 1, то есть ввести ссылку на первый диапазон анализируемых данных, содержащий один столбец данных. Для этого следует навести указатель мыши на верхнюю ячейку первого столбца данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к нижней ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши;

о указать Интервал переменной 2, то есть ввести ссылку на второй диапазон анализируемых данных, содержащий один столбец данных. Для этого следует навести указатель мыши в поле ввода Интервал переменной 2 и щелкнуть левой кнопкой мыши, затем навести указатель мыши на верхнюю ячейку второго столбца данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к нижней ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши;

О указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует поставить флажок в левое

поле Выходной диапазон (навести указатель мыши и щелкнуть левой кнопкой), далее навести указатель мыши на правое поле ввода Выходной диапазон и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные.

О нажать кнопку ОК.

Результаты анализа. В выходной диапазон будут выведены: средняя, дисперсия и число наблюдений для каждой переменной, гипотетическая разность средних, *df* (число степеней свободы), значение t-статистики, *P(T<= t)* одностороннее, *t* критическое одностороннее, P(T<= *t)* двухстороннее, *t* критическое двухстороннее.

Интерпретация результатов. Если величина вероятности случайного появления анализируемых выборок *(P(T<=t)* двухстороннее) меньше уровня значимости (α = 0,05), принято считать, что различия между выборками не случайные, то есть различия достоверные.

Пример 1.11. Рассматривается заработная плата обслуживающего персонала и работников ресторана (из примера 1.4).

Персонал	Ресторан
2100	3200
2100	3000
2000	2500
2000	2000
2000	1900
1900	1800
1800	

1800	

Можно ли по этим данным сделать вывод о большей зарплате работников ресторана?

Решение. Для решения задач такого типа используются так называемые критерии различия, в частности, t-критерий Стьюдента.

1. Введите данные: для персонала — в диапазон A1:A8; работников ресторана -в диапазон B1:B1.

2. Выбор процедуры осуществляется из трех вариантов t-теста. Поскольку данные не имеют попарного соответствия, число их различно и говорить о равенстве дисперсий затруднительно, выберите процедуру Двухвыборочный t-тест с различными дисперсиями.

Для реализации процедуры в пункте меню Сервис выберите строку Анализ данных и далее укажите курсором мыши на строку Двухвыборочный t-тест с различными дисперсиями.

3. В появившемся диалоговом окне задайте Интервал переменной 1. Для этого наведите указатель мыши на верхнюю ячейку столбца (A1), нажмите левую кнопку мыши и, не отпуская ее, протяните указатель мыши к нижней ячейке (A8) затем отпустите левую кнопку мыши.

4. Аналогично укажите Интервал переменной 2, то есть введите ссылку на диапазон второго столбца *B1:B1.*

5. Далее укажите выходной диапазон. Для этого поставьте переключатель в положение Выходной диапазон (наведите указатель мыши и щелкните левой кнопкой), затем наведите указатель мыши на правое поле ввода Выходной диапазон, щелкнув левой кнопкой мыши, указатель мыши наведите на левую верхнюю ячейку выходного диапазона (C1). Щелкните левой кнопкой мыши и нажмите кнопку ОК.

Результаты анализа. В выходном диапазоне C1:E13 появятся результаты процедуры Двухвыборочный t-тест с различными дисперсиями (рис. 1.16).

	A	B	C	D	F
1	2100	3200	Двухвыборочный t-тест с различными дисперсиями		
2	2100	3000			
3	2000	2500		*Переменная 1*	*Переменная 2*
4	2000	2000	Среднее	1962,5	2400
5	2000	1900	Дисперсия	14107,14286	356000
6	1900	1800	Наблюдения	8	6
7	1800		Гипотетическая разность средних	0	
8	1800		df	5	
9			t-статистика	-1,769982969	
10			P(T<=t) одностороннее	0,068475305	
11			t критическое одностороннее	2,015049176	
12			P(T<=t) двухстороннее	0,13695061	
13			t критическое двухстороннее	2,570577635	

Рис. 1.16. Исходные данные (A1:B8) и результаты анализа (C1:E13) из примера 1.11

Интерпретация результатов. Средние значения заработной платы (1962 руб персонала и 2400 руб. для работников ресторана) довольно сильно отличаются. Тем не менее нулевая гипотеза о том, что разницы между группами нет (то есть средние выборок равны между собой), отвергнута быть не может. Это следует из того, что вероятность реализации нулевой гипотезы достаточно велика (p = 0,1389, что больше чем уровень значимости 0,05, то есть p > 0,05) и величина вероятности случайного появления анализируемых выборок (P(T<=t) двухстороннее) больше уровня значимости (α = 0,05). А это позволяет говорить, что различия между выборками могут быть случайными, то есть различия недостоверные.

Таким образом, из полученных результатов исследования вытекает, что на основании приведенных данных нельзя сделать вывод о достоверно большей зарплате работников ресторана.

Упражнения

17. Определите, достоверны ли различия в количестве приобретаемых туристских путевок семейными парами и отдельными туристами.

Количество приобретаемых путевок

Месяцы	**1**	**2**	**3**	**4**	**5**	**6**
Пары	**67**	**75**	**58**	**89**	**96**	**94**
Одиночки	**43**	**56**	**78**	**87**	**85**	**90**

18. В таблице приведены результаты группы студентов по скоростному чтению до и после специального курса по быстрому чтению.

Студент	1	2	3	4	5	6	7	8	9	10	
До курса	86	83	86	70	66	90	70	85	77	86	
После	82	79	91		77	68	86	81	90	85	94

Произошли ли статистически значимые изменения скорости чтения у студентов?

Дисперсионный анализ

В случае необходимости оценить достоверность различия между несколькими группами наблюдений (выборками) используют методы дисперсионного анализа.

Дисперсионный анализ предназначен для исследования задачи о действии на измеряемую случайную величину (отклик) одного или нескольких независимых факторов, имеющих несколько градаций. Причем в однофакторном, двухфакторном и т. д. анализе влияющие на результат факторы считаются известными, и речь идет только о выяснении существенности или оценке этого влияния.

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

В дальнейшем ограничимся рассмотрением простейшего случая дисперсионного анализа − однофакторного анализа. При этом 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. Значимость различий проверяется по критерию Фишера.

Влияние случайной составляющей характеризуют внутригрупповая дисперсия 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:

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

Здесь М − общее среднее, М $M_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$, m − количество групп, n − количество элементов в группе.

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

Для проведения дисперсионного анализа необходимо:

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

О выполнить команду Сервис → Анализ данных;

О в появившемся диалоговом окне «Анализ данных» в списке «Инструменты анализа» выбрать процедуру «Однофакторный дисперсионный анализ», указав курсором мы; и щелкнув левой кнопкой мыши. Затем нажать кнопку ОК;

О в появившемся диалоговом окне задать Входной интервал, то есть ввести ссылку на диапазон анализируемых данных, содержащий все столбцы данных. Для этого следует навести указатель мыши на верхнюю левую ячейку диапазона данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши и нижней правой ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши (рис. 1.17);

О в разделе «Группировка» переключатель установить в положение по столбцам:

О указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует поставить переключатель в положение Выходной интервал (навести указатель мыши и щелкнуть левой кнопкой), далее навести указатель мыши на правое поле ввода Выходной интервал: щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши. Размер выходного диапазона

будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на и исходные данные.

О нажать кнопку ОК.

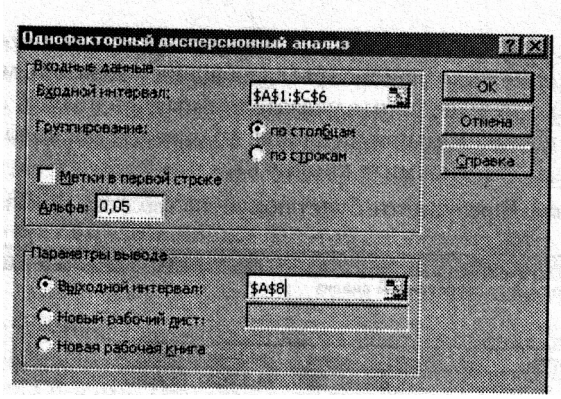


Рис. 1.17. Пример заполнения диалогового окна Однофакторный дисперсионный анализ

Результаты анализа. Выходной диапазон будет включать 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 Р-Значение. В случаях, когда Р-Значение < 0,05, критерий Фишера значим и влияние исследуемого фактора можно считать доказанным.

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

Пример 1.12. Необходимо выявить, влияет ли расстояние от центра города на степень заполняемости гостиниц. Пусть введены 3 уровня расстояний от центра города: 1) до 3 км, 2) от 3 до 5 км и 3) свыше 5 км. Данные заполняемости представлены в таблице.

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

1. Исследуемые данные введите в рабочую таблицу «MS Excel» по столбцам: в столбец А − заполняемость гостиниц в центре города, в столбец В − гостиниц, находящихся на расстоянии от 3 до 5 км и т. д. (диапазон A1:C6).

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

3. В появившемся диалоговом окне «Однофакторный дисперсионный анализ» в поле «Входной интервал» задайте A1:C1. Для этого наведите указатель мыши на ячейку A1 и протяните его к ячейке C6 при нажатой левой кнопке мыши.

4. В разделе «Группировка» переключатель установите в положение по столбцам.

5. Далее необходимо указать выходной диапазон. Для этого поставьте переключатель в положение Выходной интервал (наведите указатель мыши и щелкните левой кнопкой), затем щелкните указателем мыши в правом поле ввода Выходной интервал, и щелчком мыши на ячейке A8 укажите расположение выходного диапазона (рис. 1.17). Нажмите кнопку ОК.

Результаты анализа. В результате будет получена таблица, показанная на рис. 1.18.

Однофакторный дисперсионный анализ						
ИТОГИ						

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
Столбец 1	**6**	**560**	**93,33333**	**13,4667**		
Столбец 2	**6**	**516**	**86**	**14**		
Столбец 3	**6**	**475**	**79,1667**	**32,9667**		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	**602,3**	**2**	**301,166**	**14,95036**	**0,0002684**	**3,6823166**
Внутри групп	**302,1**	**15**	**20,144**			
Итого	**904,5**	**17**				

Рис. 1.18. Результат работы инструмента Однофакторный дисперсионный анализ

Интерпретация результатов. В таблице «Дисперсионный анализ» на пересечении строки Между группами и столбца P-Значение находится величина 0,0002684. Величина P-Значение < 0,05, следовательно, критерий Фишера значим и влияние фактора расстояния от центра города на эффективность заполнения гостиниц доказано статистически.

Упражнение

19. Определите, влияет ли фактор образования на уровень зарплаты в гостинице на основании следующих данных:

Образование	Зарплата сотрудника
высшее	3200 3000 2600 2000 1900 1900
среднее спец	2600 2000 2000 1900 1800 1800
среднее	2000 2000 1900 1800 1700 1700

Корреляционный анализ

Важным разделом статистического анализа является корреляционный анализ, служащий для выявления взаимосвязей между выборками.

Коэффициент корреляции

Выявление взаимосвязей. Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между некоторыми наблюдаемыми переменными. Знание взаимозависимостей отдельных признаков дает возможность решать одну из кардинальных задач любого научного исследования: возможность

предвидеть, прогнозировать развитие ситуации при изменении конкретных характеристик объекта исследования. Например, основное содержание любой экономической политики, в конечном счете, может быть сведено к регулированию экономических переменных, осуществляемому на базе выявленной тем или иным образом информации об их взаимовлиянии. Поэтому, проблема изучения взаимосвязей показателей различного рода является одной из важнейших в статистическом анализе.

Обычно взаимосвязь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей различают корреляцию и регрессию.

Регрессионный анализ (см. раздел «Регрессионный анализ») устанавливает формы зависимости между случайной величиной Y и значениями одной или нескольких переменных величин.

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

Для оценки степени взаимосвязи наибольшее распространение получил коэффициент линейной корреляции (Пирсона), предполагающий нормальный закон распределения наблюдений.

Коэффициент корреляции *(R, r)* − 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. Здесь под прямой зависимостью понимают зависимость, при которой увеличение или уменьшение значения одного признака ведет, соответственно, к увеличению или уменьшению второго. Например, при увеличении температуры возрастает давление

газа, а при уменьшении − снижается (при постоянном объеме). При обратной зависимости увеличение одного признака приводит к уменьшению второго и наоборот. Примером обратной корреляционной зависимости может служить связь между температурой воздуха на улице и количеством топлива, расходуемого на обогрев помещения.

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

$$r = \frac{\sum_i (x_i - M_x)(y_i - M_y)}{\sqrt{\sum_i (x_i - M_x)^2 (y_i - M_y)^2}}$$

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. 1.19). Для оценки степени взаимосвязи можно руководствоваться следующими эмпирическими правилами. Если коэффициент корреляции (r) по абсолютной величине (без учета знака) больше, чем 0,95, то принято считать, что между параметрами существует практически линейная зависимость, (прямая − при положительном r и обратная − при отрицательном r). 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 обычно считают, что линейная взаимосвязь между параметрами выявить не удалось.

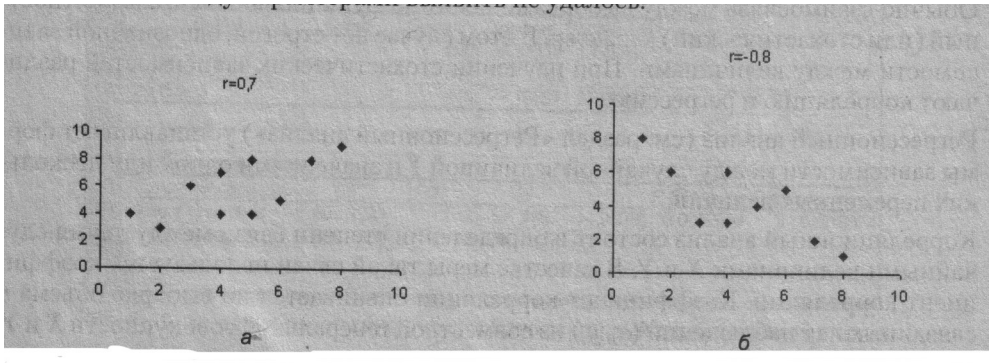


Рис. 1.19. Примеры прямой (r» 0,7, *а)* и обратной (r= -0,8, *б)* корреляционной зависимости

О В MS «MS Excel» для вычисления парных коэффициентов линейной корреляции используется специальная функция К0РРЕЛ. Параметрами функции являются К0РРЕЛ(*масив1;массив2),* где: *массив1* — это диапазон ячеек первой случайной величины;

О *массив2* − это второй интервал ячеек со значениями второй случайной величины.

Пример 1.13. Имеются результаты семимесячных наблюдений реализации путевок двух туристских маршрутов тура А и тура В.

Тур А	**Тур В**
120	20
121	15
105	18
92	16
113	19
90	16
80	15

Необходимо определить, имеется ли взаимосвязь между количеством продаж путевок обоих маршрутов.

Решение

Для выявления степени взаимосвязи, прежде всего, необходимо ввести данные в рабочую таблицу.

Откройте новую рабочую таблицу. Введите в ячейку А1 слова Тур А. Затем в ячейки А2:А8 − соответствующие значения числа продаж. В ячейки В1:В8 введите название и значения для тура В. Затем вычисляется значение коэффициента корреляции между выборками. Для этого табличный курсор установите в свободную ячейку (А9). На панели инструментов нажмите кнопку Вставка функции (f_x). В появившемся диалоговом окне Мастер функций выберите категорию Статистические и функцию КОРРЕЛ, после чего нажмите кнопку ОК. Появившееся диалоговое окно

КОРРЕЛ за серое поле мышью отодвиньте вправо на 1-2 см от данных (при нажатой левой клавише). Указателем мыши введите диапазон данных *Тур А* в поле Массив 1 *(А2:А8).* В поле Массив 2 введите диапазон данных *Тур В (В2:В8).* Нажмите кнопку ОК. В ячейке А9 появится значение коэффициента корреляции − *0,995493.* Значение коэффициента корреляции больше чем 0,95. Значит, можно говорить о том, что в течение периода наблюдения имелась высокая степень прямой линейной взаимосвязи между количествами проданных путевок обоих маршрутов (r= 0,557292).

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 строки и столбца находится коэффициент корреляции между соответствующими параметрами.

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

Для реализации процедуры необходимо:

О выполнить команду Сервис → Анализ данных;

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

О в появившемся диалоговом окне указать Входной интервал, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Для этого следует навести указатель мыши на левую верхнюю ячейку данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к правой нижней ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши. Входной интервал должен содержать не менее двух столбцов.

О в разделе Группировка переключатель установить в соответствии с введенными

данными;

О указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует поставить флажок в левое поле Выходной интервал (навести указатель мыши и щелкнуть левой кнопкой), далее навести указатель мыши на правое поле ввода Выходной интервал и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные (рис. 1.20).

О нажать кнопку ОК.

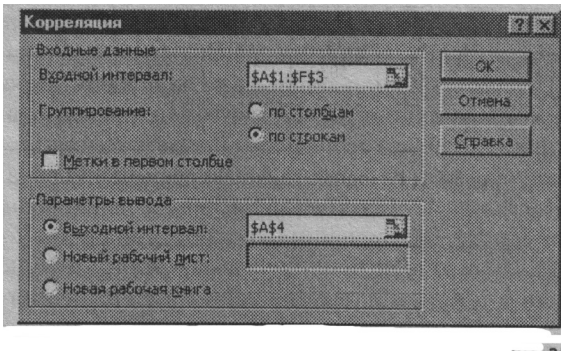


Рис. 1.20. Пример установки параметров корреляционного анализа

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

Интерпретация результатов. Рассматривается отдельно каждый коэффициент корреляции между соответствующими параметрами. Его числовое значение оценивается по эмпирическим правилам, изложенным в разделе «Коэффициент корреляции». Отметим, что хотя в результате будет получена треугольная матрица, корреляционная матрица симметрична, и коэффициенты корреляции

$r_{ij}=r_{ji}$.

Пример 1.14. Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью музеев и парков.

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

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

Решение. Для выполнения корреляционного анализа введите в диапазон A1:C3 исходные данные (рис. 1.21).

Затем в меню Сервис выберите пункт Анализ данных и далее укажите строку Корреляция. В появившемся диалоговом окне укажите Входной интервал *B1:G3.* Укажите, что данные рассматриваются по строкам. Укажите выходной диапазон. Для этого поставьте флажок в левое поле Выходной интервал и в правое поле ввода Выходной интервал введите *A4* (рис. 1.20). Нажмите кнопку ОК.

	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. 1.21. Исходные данные из примера 1.14

	Строка 1	Строка 2	Строка 3
Строка 1	1		
Строка 2	-0,921	1	
Строка 3	0,974	-0,919	1

Рис. 1.22. Результаты вычисления корреляционной матрицы из примера 1.14

Результаты анализа. В выходном диапазоне получаем корреляционную матрицу (рис. 1.22).

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 выявлены зависимости: сильная степень обратной линейной взаимосвязи между посещаемостью музея и количеством солнечных дней (r =-0,92) и 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).

Подразумевается, что в пустых клетках в правой верхней половине таблицы находятся те же коэффициенты корреляции, что и в нижней левой (симметрично расположенные относительно диагонали).

Упражнения

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

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

21. Определите, имеется ли взаимосвязь между годовым уровнем инфляции (%), ставкой рефинансирования (%) и курсом доллара (руб./$), по следующим данным ежегодных наблюдений:

Уровень инфляции	Ставка рефинансирования	Курс $
84	85	6,3
45	55	14
56	65	20
34	40	28
23	28	29

Регрессионный анализ

При исследовании взаимосвязей между выборками помимо корреляции различают также и регрессию. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных. Соответственно, наряду с корреляционным анализом еще одним

инструментом ˙изучения стохастических зависимостей является регрессионный анализ.

Регрессионный анализ устанавливает формы зависимости между случайной величиной Y (зависимой) и значениями одной или нескольких переменных величин (независимых), причем значения последних считаются точно заданными. Такая зависимость обычно определяется некоторой математической моделью (уравнением регрессии), содержащей несколько неизвестных параметров. В ходе регрессионного анализа на основании выборочных данных находят оценки этих параметров, определяются статистические ошибки оценок или границы доверитель интервалов и проверяется соответствие (адекватность) 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), …,$(X_n\ Y_n)$ построить (подобрать) прямую линию, называемую линией регрессии, которая «наилучшим образом» приближает наблюдаемые значения. Yравнение этой линии $Y = aX + b$ является регрессионным уравнением. С помощью_ регрессионного уравнения можно предсказать ожидаемое значение зависимости личины Y_0, соответствующее заданному значению независимой переменной X_0.

Таким образом, можно сказать, что линейный регрессионный анализ заключало в подборе графика и его уравнения для набора наблюдений. В регрессионном анализе все признаки (переменные), входящие в уравнение, должны иметь непрерывную, а не дискретную природу.

В случае, когда рассматривается зависимость между одной зависимой переменной

Y и несколькими независимыми $X_u\ X_2, ..., X_n$, говорят о множественной линейной регрессии. В этом случае регрессионное уравнение имеет вид

$$Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n$$

где a_1, a_2, ..., a_n − требующие определения коэффициенты при независимых

переменных X_1, X_2, ...,X_n a_0 − константа.Мерой эффективности регрессионной модели является коэффициент детерминации R^2 (R-квадрат). Коэффициент детерминации (R-квадрат) определяет, с какой степенью точности полученное регрессионное уравнение описывает (аппроксимирует) исходные данные.

Исследуется также значимость регрессионной модели 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, a_n от нуля проверяется с помощью критерия Стьюдента. В случаях, когда $p > 0,05$, коэффициент может считаться нулевым, а это означает, что влияние соответствующей независимой переменной на зависимую переменную недостоверно, и эта независимая переменная может быть исключена из уравнения.

В MS «MS Excel» экспериментальные данные аппроксимируются линейным уравнением до 16 порядка: $У = a_0 + a_1X_1 + a_2X_2 + ... + a_nX_n$,

где Y− зависимая переменная, X1, ...,Xn − независимые переменные, a0, a1, ..., an − искомые коэффициенты регрессии.

Для получения коэффициентов регрессии используется процедура Регрессия из пакета анализа. Кроме того, могут быть использованы функция Л ИНЕЙН для получения параметров регрессионного уравнения и функция ТЕНДЕНЦИЯ для получения предсказанных значений Y в требуемых точках (см. раздел «Несколько независимых переменных» главы 3).

Для реализации процедуры Регрессия необходимо:

О выполнить команду «Сервис» → «Анализ данных»;

О в появившемся диалоговом окне Анализ данных в списке Инструменты анализа выбрать строку Регрессия, указав курсором мыши и щелкнув левой кнопкой мыши. Затем нажать кнопку 0К;

О в появившемся диалоговом окне задать Входной интервал Y, то есть ввести ссылку на диапазон анализируемых зависимых данных, содержащий один столбец данных. Для этого следует навести указатель мыши на верхнюю ячейку

столбца зависимых данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к нижней ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши;

О указать *Входной интервал X,* то есть ввести ссылку на диапазон независимых данных, содержащий до 16 столбцов анализируемых данных. Для этого следует навести указатель мыши на поле ввода Входной интервал X и щелкнуть левой кнопкой мыши, затем навести указатель мыши на верхнюю левую ячейку диапазона независимых данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к нижней правой ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши;

О указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует поставить переключатель в положение Выходной интервал (навести указатель мыши и щелкнуть левой кнопкой), далее навести указатель мыши на правое поле ввода Выходной интервал и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши (рис. 1.23). Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные;

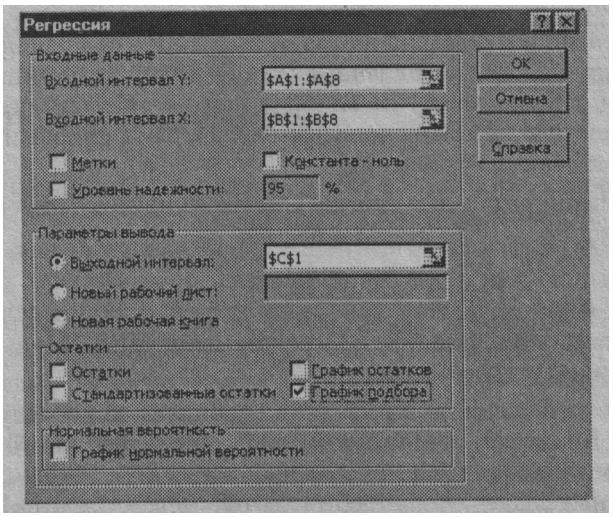


Рис. 1.23. Пример заполнения диалогового окна Регрессия

О если необходимо визуально проверить отличие экспериментальных точек от

предсказанных по регрессионной модели, следует установить флажок в поле График подбора;

О 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.

Интерпретация результатов. Значения коэффициентов регрессии находятся в столбце Коэффициенты и соответствуют:

О Y-пересечение − a_0

О переменная X_1 − a_1

О переменная X_2 −a_2 и т. д.

В столбце P-Значение приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда $P > 0{,}05$, коэффициент может считаться нулевым, что означает, что соответствующая независимая переменная практически не влияет на зависимую переменную.

Приводимое значение R-квадрат (коэффициент детерминации) определяет, с какой степенью точности полученное регрессионное уравнение аппроксимирует исходные данные. Если R-квадрат > 0,95, говорят о высокой точности аппроксимации (модель хорошо описывает явление). Если R-квадрат лежит в диапазоне от 0,8 до 0,95, говорят об удовлетворительной аппроксимации (модель в целом 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 1.15. В отделе снабжения гостиницы имеется информация об изменении стоимости стирального порошка за длительный период времени. Сопоставляя его с изменениями курса доллара за этот же период времени, можно построить регрессионное уравнение. Ниже приведены стоимость пачки стирального порошка (в руб.) и соответствующий курс доллара (руб./USD).

N	Порошок	Курс
1	5	6,3
2	7	9
3	9	12
4	12	15
5	15	19
6	16	21
7	20	25
8	25	29,3

Необходимо на основании этих данных построить регрессионное уравнение, позволяющее по курсу доллара определять предполагаемую стоимость пачки стирального порошка.

Решение

1. Введите данные в рабочую таблицу: стоимость пачки порошка − в диапазон A1:A8; курс доллара в диапазон B1:B8 (заметим, что знаку запятой, отделяющей целую часть от дробной, соответствует «запятая»).

2. В пункте меню Сервис выберите строку Анализ данных и далее укажите курсором мыши на строку Регрессия.

3. В появившемся диалоговом окне (рис. 1.23) задайте *Входной интервал Y*. Для этого наведите указатель мыши на верхнюю ячейку столбца зависимых данных (A1), нажмите левую кнопку мыши и, не отпуская ее, протяните указатель мыши к нижней ячейке (A8), затем отпустите левую кнопку мыши. (Обратите внимание, что зависимые данные − это те данные, которые предполагается вычислять.)

4. Так же укажите *Входной интервал X,* то есть введите ссылку на диапазон независимых данных B1:B8. (Независимые данные − это те данные, которые будут измеряться или наблюдаться.)

5. Установите флажок в поле График подбора.

6. Далее укажите выходной диапазон. Для этого поставьте переключатель в положение Выходной интервал (наведите указатель мыши и щелкните левой кнопкой), затем наведите указатель мыши на правое поле ввода Выходной интервал и, щелкнув левой кнопкой мыши, указатель мыши наведите, на левую верхнюю ячейку выходного диапазона (C1). Щелкните левой кнопкой мыши (рис. 1.1) Нажмите кнопку ОК.

Результаты анализа. В выходном диапазоне появятся следующие результаты и (рис. 1.24).

Регрессионная статистика	
Множественный R	0,9960
R-квадрат	0,9920
Нормированный R-квадрат	0,9900
Стандартная ошибка	0,6518
Наблюдения	8,0000

	df	SS	MS	F	Значимость F
Регрессия	1,000	317,33	317,33	748,5832	1,575E-07
Остаток	6,000	2,5434	0,4239		
Итого	7,000	319,875			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%
Y-пересечение	-0,8309	0,5763	-1,4417	0,1994	-2,2411
Переменная X 1	0,8466	0,0309	27,3602	1,58E-07	0,77089

	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	0,5793	-2,2411	0,5793
Переменная X 1	0,9223	0,7708	0,9223

Рис. 1.24. Результаты анализа и график соответствия экспериментальных точек и предсказанных по регрессионной модели из примера 1.15

Интерпретация результатов. В таблице Дисперсионный анализ оценивается общее качество полученной модели: ее достоверность по уровню значимости критерия Фишера − *p,* который должен быть меньше, чем 0,05 (строка Регрессия, столба Значимость F, в примере – 1,575E-07 (0,0000001575), то есть p = 0,0000001575 и модель значима) и степень точности описания моделью процесса − *R*-квадрат (вторая строка сверху в таблице Регрессионная статистика, в примере *R* -квадрат = 0,992) Поскольку *R* -квадрат > 0,95, можно говорить о высокой точности аппроксимации (модель хорошо описывает явление (рис. 1.24)).

Далее необходимо определить значения коэффициентов модели. Они 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 стоимости пачки порошка в рублях будет иметь следующий вид: *-0,83 + 0,847*(Курс доллара, руб./USD).*

Полученная модель с высокой точностью позволяет определять стоимость пачки стирального порошка *(R^2* = 99,2%).

Воспользовавшись полученным уравнением, можно рассчитать ожидаемую стоимость пачки стирального порошка при изменениях курса доллара. Например, для расчета при курсе доллара 35 руб./USD необходимо поставить табличный курсор в любую свободную ячейку (A10); ввести с клавиатуры знак =, щелкнуть указателем мыши по ячейке D17, ввести с клавиатуры знак +, щелкнуть по ячейке D18, ввести с клавиатуры знак * и число *35.* В результате в ячейке A10 будет получена ожидаемая стоимость пачки порошка − *28,8* руб.

Пример 1.16. Построить регрессионную модель для предсказания изменений уровня заболеваемости органов дыхания (Y) в зависимости от содержания в воздухе двуокиси углерода (X_1,) и степени запыленности *(X_2).* В таблице приведены данные наблюдений в течение 29 месяцев.

X1	X2	Y
1,0	1,3	1160
1,0	1,3	1155
1,1	1,4	1158
1,1	1,4	1157
1,1	1,5	1160
1,1	1,5	1161
1,0	1,4	1157
1,0	1,5	1159
1,2	1,6	1256

1,2	1,7	1260
0,6	1,0	1040
0,6	1,0	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	1140
0,8	1,4	1138
0,78	1,5	1240
0,80	1,5	1239
0,78	1,5	1241
0,78	1,6	1240
0,80	1,7	1239
0,80	1,8	1239
0,75	1,8	1240
0,78	1,9	1238
0,75	1,9	1238

Решение

1. Введите данные наблюдений в диапазон A1:C30 рабочей таблицы «MS Excel».

1. В пункте меню «Сервис» выберите строку «Анализ данных» и далее укажите курсором мыши на строку «Регрессия». Нажмите кнопку ОК.

2. В появившемся диалоговом окне задаем *Входной интервал Y*. Для этого наведите указатель мыши на верхнюю ячейку столбца зависимых данных (C2) и жмите левую кнопку мыши и, не отпуская ее, протяните указатель мыши к нижней ячейке (C30), затем отпустите левую кнопку мыши. (Обратите внимание, что зависимые данные − это те данные, которые предполагается вычислять).

3. Так же укажите *Входной интервал X,* то есть введите ссылку на диапазон независимых данных A2:B30. (Независимые данные − это те данные, которые будут измеряться или наблюдаться).

2. Установите флажок в поле График подбора.

4. Далее укажите выходной диапазон. Для этого поставьте переключатель в положение Выходной интервал (наведите указатель мыши и щелкните левой кнопкой), затем наведите указатель мыши на правое поле ввода Выходной интервал и, щелкнув левой кнопкой мыши, указатель мыши наведите на левую верхнюю ячейку выходного диапазона (D1). Щелкните левой кнопкой мыши. Нажмите кнопку ОК.

5. В выходном диапазоне появятся результаты регрессионного анализа и графики предсказанных точек (рис. 1.25).

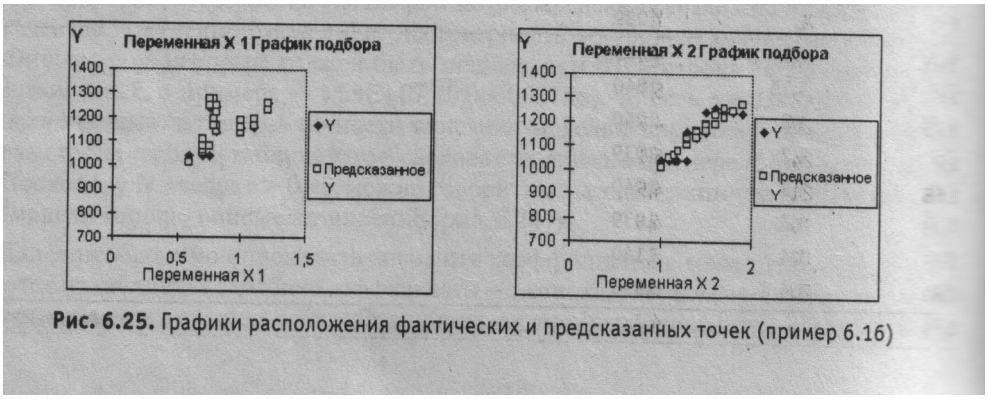


Рис. 1.25. Графики расположения фактических и предсказанных точек (пример 1.16)

Интерпретация результатов. В таблице «Дисперсионный анализ» оценивается достоверность полученной модели по уровню значимости критерия Фишера (строка Регрессия, столбец Значимость F, в примере − *1,4E-09 (1,4*10^{-9}),* то есть p « 0,05 и модель значима) и степень описания моделью процесса − R-квадрат (вторая строка сверху в таблице Регрессионная статистика, в примере R-квадрат = 0,89). Поскольку R-квадрат > 0,8, можно говорить о довольно высокой точности аппроксимации (модель хорошо описывает зависимость заболеваемости от содержания углекислого газа и запыленности воздуха (рис. 1.25)).

Далее необходимо определить значения коэффициентов модели. Они 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. Все коэффициенты значимы, то есть p < 0,05, и коэффициенты могут считаться не равными нулю.

Поэтому выражение для определения уровня заболеваемости органов дыхания в зависимости от содержания углекислого газа и пыли в воздухе будет иметь вид:

$У=682+91*X_1+275*X_2$.

Упражнения

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

Возраст	**Зарплата**
20	800
50	2500
45	2500
40	2000
25	1200
30	1800

23 Постройте зависимость жизненной емкости легких в литрах (Y) от роста в метрах (X_1) и возраста в годах (X_2) для группы мужчин:

X1	X2	Y
1,85	18	5,4
1,80	25	5,7
1,75	20	4,8
1,70	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

24. Определите должное значение жизненной емкости легких для мужчины возраста 22 лет и роста 183 см из регрессионного уравнения, полученного в предыдущем упражнении.

25. Имеются данные о цене на нефть *x* (ден. ед.) и индексе акций нефтяных компаний Y (усл. ед.):

X	Y
17,28	537
17,05	534
18,30	550
18,80	555
19,20	560
18,50	552

Постройте зависимость индекса акций нефтяных компаний от цены на нефть.