

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
"Оренбургский государственный университет"

Кафедра математических методов и моделей в экономике

Н.П. ФОТ, А.Г. ГАНСКАЯ, О.Н. ЯРКОВА

МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ С ПРИМЕНЕНИЕМ ЭЛЕКТРОННОЙ ТАБЛИЦЫ EXCEL

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНОМУ ПРАКТИКУМУ И
САМОСТОТЕЛЬНОЙ РАБОТЕ СТУДЕНТОВ

Рекомендовано к изданию Редакционно-издательским советом
государственного образовательного учреждения
высшего профессионального образования
"Оренбургский государственный университет"

Оренбург 2006

УДК 519.23 (076.5)
ББК 22.172 я 73
Ф 81

Рецензент
кандидат экономических наук, доцент С.В. Дьяконова

Фот Н.П.

Ф 81 Методы математической статистики с применением электронной таблицы Excel [Текст]: методические указания к лабораторному практикуму и самостоятельной работе студентов/ Н.П.Фот, А.Г.-Ганская, О.Н. Яркова– Оренбург: ГОУ ОГУ, 2006. – 29 с.

Методические указания предназначены для выполнения лабораторного практикума и самостоятельной работы по дисциплине «Теория вероятностей и математическая статистика» для студентов экономических специальностей.

ББК 22.172 я 73

© Фот Н.П., 2006
Ганская А.Г.
Яркова О.Н.

© ГОУ ОГУ, 2006

Содержание

Введение.....	5
1 Постановка задачи.....	6
2 Построение гистограммы и статистической функции распределения в среде пакета Excel.....	6
3 Оценка основных выборочных характеристик: среднего значения, дисперсии, среднеквадратического отклонения. Нахождение их посредством электронной таблицы Excel.....	12
4 Проверка гипотезы о законе распределения генеральной совокупности с помощью критерия согласия	14
5 Построение доверительных интервалов для основных характеристик генеральной совокупности.....	17
6 Оценка парного коэффициента корреляции (нахождение его в таблице Excel). Проверка его значимости и построение доверительного интервала.....	19
7 Оценка уравнения регрессии. Проверка значимости и построение доверительных интервалов для значимых параметров регрессии.....	22
8 Вопросы к защите.....	25
Список использованных источников.....	26
Приложение А.....	28

Введение

Математическая статистика изучает различные методы обработки и осмысления результатов многократно повторяемых случайных событий. Обработка данных и получения на ее основе каких-либо рекомендаций относительно принятия того или иного управленческого решения – процесс многоэтапный.

Для выявления закономерностей, присущих экономическим и социально-экономическим показателями, выборочные данные подвергают первичной обработке (построение гистограммы и функции распределения), производят оценку числовых характеристик показателей (математического ожидания и дисперсии) и выявляют степень тесноты линейной статистической связи (расчет коэффициента корреляции).

Целью данных методических указаний является:

- 1) наработка навыков получения точечных и интервальных оценок с помощью электронной таблицы Excel;
- 2) проверка статистических гипотез по выборочным данным;
- 3) выявление взаимосвязей между признаками посредством корреляционно-регрессионного анализа.

1 Постановка задачи

Предприятия машиностроительной отрасли характеризуются двумя основными показателями производственно – хозяйственной деятельности.

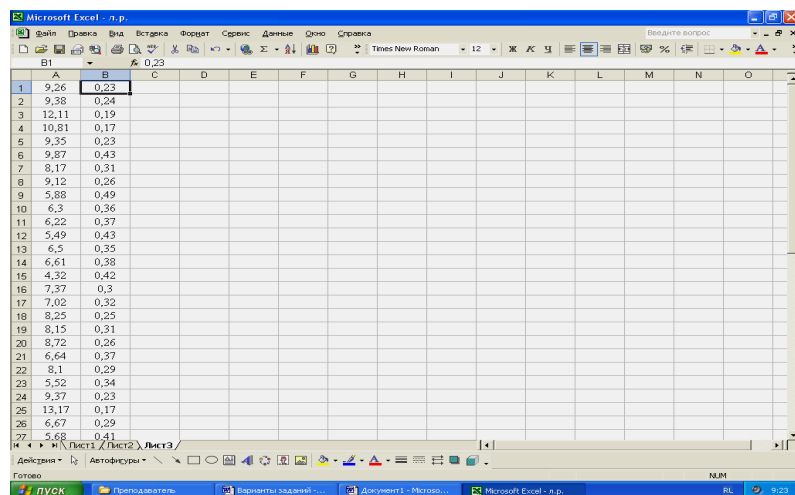
На основании выборочных данных из генеральной совокупности по двум, указанным в варианте (приложение А), признакам:

- 1) по выборочным данным построить гистограмму и график эмпирической функции распределения;
- 2) определить основные числовые характеристики: среднее значение, дисперсию, среднеквадратическое отклонение;
- 3) на основании выборочных данных X и Y с помощью критерия согласия при уровне значимости $\alpha=0,05$, проверить гипотезу о законе распределения генеральной совокупности;
- 4) построить 95%-ые доверительные интервалы для основных числовых характеристик генеральной совокупности;
- 5) найти коэффициент корреляции между признаками X и Y , проверить его значимость, для значимого параметра связи с $\alpha=0,05$ построить доверительный интервал, сделать выводы;
- 6) оценить уравнение регрессии, проверить значимость коэффициентов регрессии, для значимых параметров построить доверительные интервалы, сделать выводы ($\alpha=0,05$).

2 Построение гистограммы и статистической функции распределения в среде пакета Excel

Пример приведен с использованием электронной таблицы Excel и рассмотрен для 0 варианта приложения А.

Данные вводятся в среду электронного пакета Excel, где переменной А – соответствуют значения X_1 , переменной В– значения X_4 (фрагмент таблицы представлен на рисунке 1).



	A	B
1	9,26	0,23
2	9,38	0,24
3	12,11	0,19
4	10,81	0,17
5	9,35	0,23
6	9,87	0,43
7	8,17	0,31
8	9,12	0,26
9	5,88	0,49
10	6,3	0,36
11	6,22	0,37
12	5,49	0,43
13	6,5	0,35
14	6,61	0,39
15	4,32	0,42
16	7,37	0,3
17	7,02	0,32
18	8,25	0,25
19	8,15	0,31
20	8,72	0,26
21	6,64	0,37
22	8,1	0,29
23	5,52	0,34
24	9,37	0,23
25	13,17	0,17
26	6,67	0,29
27	5,68	0,41

Рисунок 1 – Ввод исходных данных

Для построения гистограммы на панели инструментов, в меню «Сервис» выбирают функцию «Анализ данных», окно которого представлено на рисунке 2 (при отсутствии данной категории - в меню «Сервис»: в модуле «Настройка» активизируйте пункт «Пакет анализа», после чего появляется запрашиваемая функция).

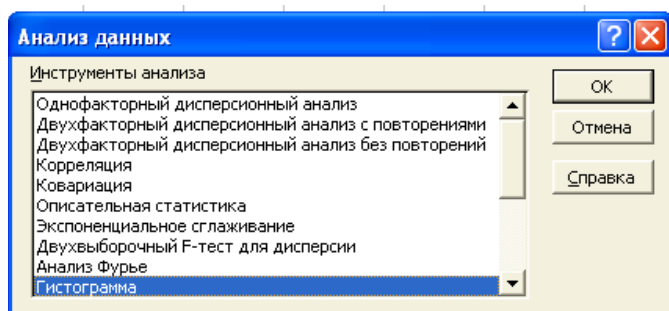


Рисунок 2 – Выбор меню «Анализ данных»

На следующем этапе производится выбор категории «Гистограмма» - в диалоговом окне указывают основные параметры для построения гистограммы по исходным данным (рисунок 3).

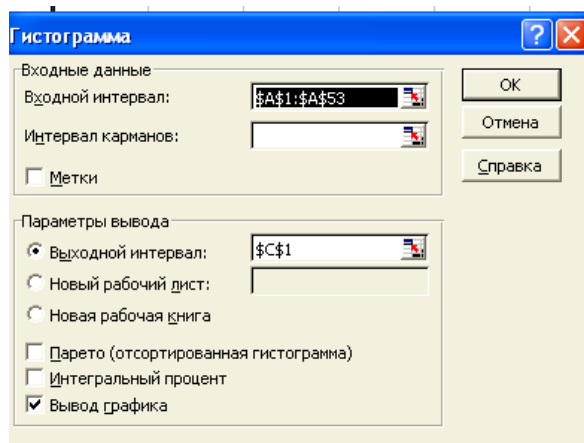


Рисунок 3 – Окна ввода данных

К вводимым параметрам относятся:

- i. входной диапазон - ссылка на диапазон, содержащий анализируемые данные;
- ii. интервал карманов (необязательный) - вводится диапазон ячеек, определяющих отрезки (карманы). Эти значения должны быть введены в возрастающем порядке. В Microsoft Excel вычисляется число попаданий данных между текущим началом отрезка и соседним большим по порядку, если такой есть. При этом включаются значения на нижней границе отрезка и не включаются значения на верхней границе. Если диапазон карманов не введен, то набор отрезков, равномерно распре-

- ленных между минимальным и максимальным значениями данных, будет создан автоматически;
- iii. выходной диапазон - вводят ссылку на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные;
 - iv. новый лист - устанавливают переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, вводят имя нового листа в поле, расположенном напротив соответствующего положения переключателя;
 - v. новая книга - устанавливают переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге;
 - vi. Парето (отсортированная диаграмма) - устанавливают флажок, чтобы представить данные в порядке убывания частоты. Если флажок снят, то данные в выходном диапазоне будут представлены в порядке возрастания отрезков, а трех самых правых столбцов с отсортированными данными не будет;
 - vii. интегральный процент - устанавливают флажок для генерации интегральных процентных отношений включения в гистограмму графика интегральных процентов. Чтобы не вычислять интегральные процентные соотношения флажок снимают;
 - viii. вывод графика - устанавливают флажок для автоматического создания встроенной диаграммы на листе, содержащем выходной диапазон.

Для построения гистограммы переменной X1 в окне ввода данных указываем входной интервал - «A1:A53», выходной интервал – ячейку «C1» и ставим флажок для вывода графика. Далее нажимаем на кнопку «ОК». Вывод результатов представлен на рисунке 4.

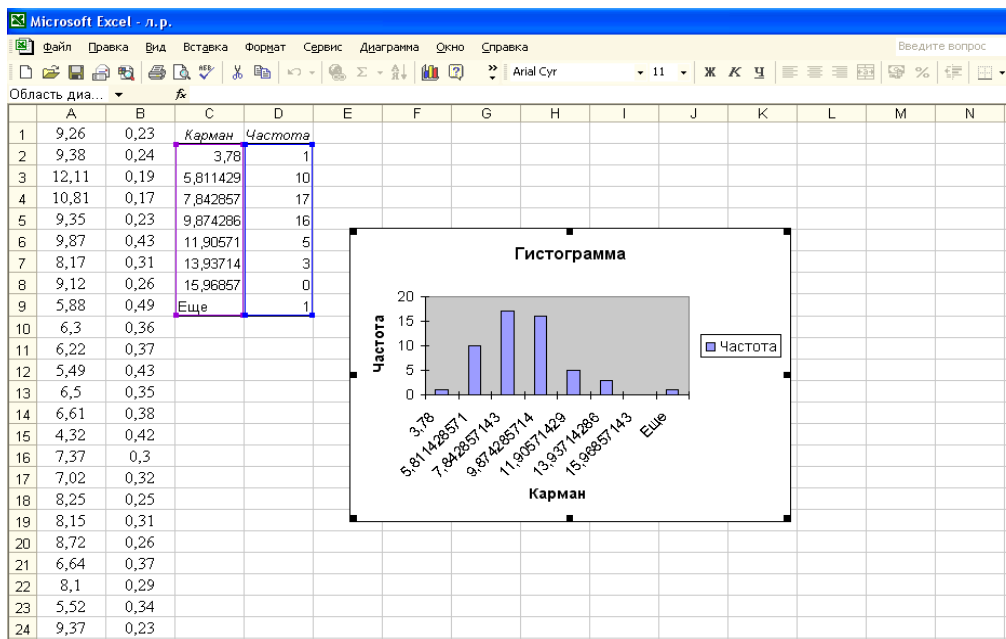


Рисунок 4 – Вывод результатов

Вследствие того, что не был указан параметр – «Интервал карманов», все исходные значения автоматически разделены на равные промежутки (в выводе результатов – столбец «Карманы»), и для каждого из них посчитано количество заданных значений, попавших в соответствующий интервал (столбец – «Частота»). Графическим представлением полученных выводов является график «Гистограмма».

Построение графика эмпирической функции распределения не предусмотрено статистическим модулем «Анализ данных» и поэтому на первом этапе построения проделывают несколько дополнительных вычислений.

Геометрическим представлением эмпирической функции распределения называют кумулятой или кумулятивной прямой, где по оси x – расположены границы интервалов (карманов), а по оси y – накопленная частота v^H_k , равная сумме частот всех предшествующих интервалов.

Найдем накопленные частоты для рассматриваемого примера, путем последовательного сложения ячеек столбца «Частоты» (рисунок 5). Таким образом, получаем столбец, состоящий из найденных накопленных частот. Отметим, что для ввода формулы используем в меню «Вставка» категорию «Функция», в которой выбираем операцию сложения. Все формулы в среде электронной таблицы Excel вводятся в свободную ячейку и начинаются со знака «=».

	A	B	C	D	E	F	G
1	9,26	0,23	Карман	Частота	Накопленные частоты	F(x)	
2	9,38	0,24	3,78	1		1	
3	12,11	0,19	5,811429	10		11	
4	10,81	0,17	7,842857	17		28	
5	9,35	0,23	9,874286	16	=E4+D5		
6	9,87	0,43	11,90571	5			
7	8,17	0,31	13,93714	3			
8	9,12	0,26	15,96857	0			
9	5,88	0,49	Еще	1			
10	6,3	0,36					
11	6,22	0,37					
12	5,49	0,43					
13	6,5	0,35					

Рисунок 5 – Вычисление накопленных частот

На следующем этапе для каждого интервала вычисляем эмпирическую функцию распределения по формуле:


$$F(x) = \frac{\nu_k^H}{n}, \quad (1)$$

принимая значение равное нулю при $\nu_k^H = 0$ и значение единицы при $\nu_k^H = 53$ (n – объем выборки, равен 53 для каждого интервала).

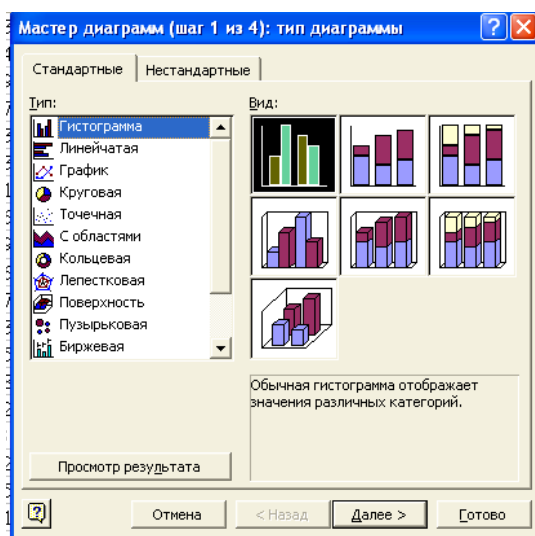
Для этого в свободную ячейку вводим формулу (1) и составляем столбец значений функции распределения, последнее значение должно быть =1 (рисунок 6).

	A	B	C	D	E	F	G
1	9,26	0,23	Карман	Частота	Накопленные частоты	F(x)	
2	9,38	0,24	3,78	1	1	0,01886792	
3	12,11	0,19	5,811429	10	11	=E3/53	
4	10,81	0,17	7,842857	17	28		
5	9,35	0,23	9,874286	16	44		
6	9,87	0,43	11,90571	5	49		
7	8,17	0,31	13,93714	3	52		
8	9,12	0,26	15,96857	0	52		
9	5,88	0,49	Еще	1	53		
10	6,3	0,36					
11	6,22	0,37					
12	5,49	0,43					
13	6,5	0,35					

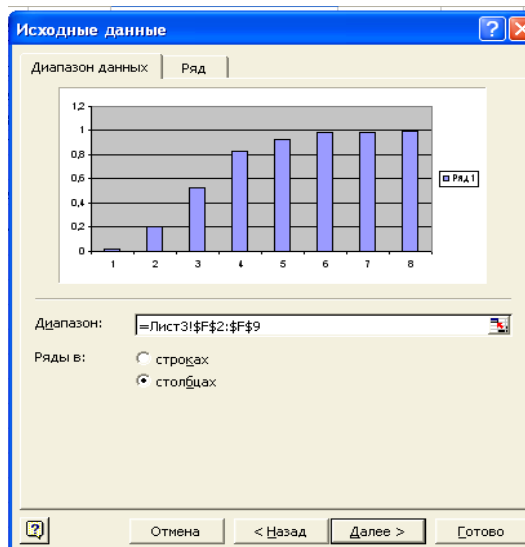
Рисунок 6 - Вычисление значений функции распределения

Для построения графика на панели инструментов щелкаем на значке  и в меню «Мастер диаграмм» последовательно указываем необходимые параметры, после каждого шага нажимая кнопку «Далее».

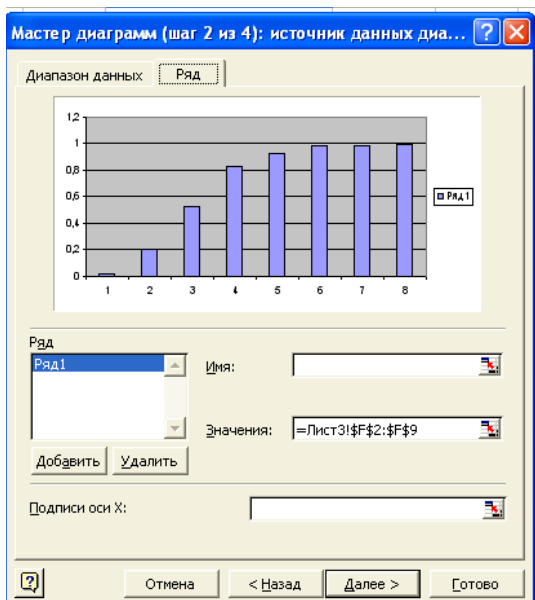
На первом шаге предлагается выбор типа и вида диаграммы (рисунок 7а). На втором шаге указываем диапазон данных для построения путем выделения их левой кнопкой мыши (рисунок 7б). Для указания дополнительных параметров: подписей по оси X и имени построенного ряда переходят в окно «Ряд» (рисунок 7в). На третьем шаге предлагается указать следующие параметры диаграммы: заголовки, оси, линии сетки, легенды, подписи данных и таблицы данных (рисунок 7г).



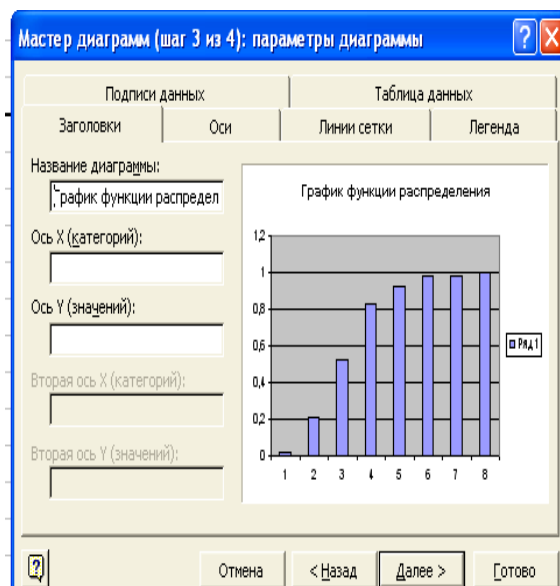
а



б



в



г

Рисунок 7 – Построение графика функции распределения

На завершающем, четвертом шаге указываем размещение диаграммы, после нажатия на кнопку «Готово» выводится график (рисунок 8).

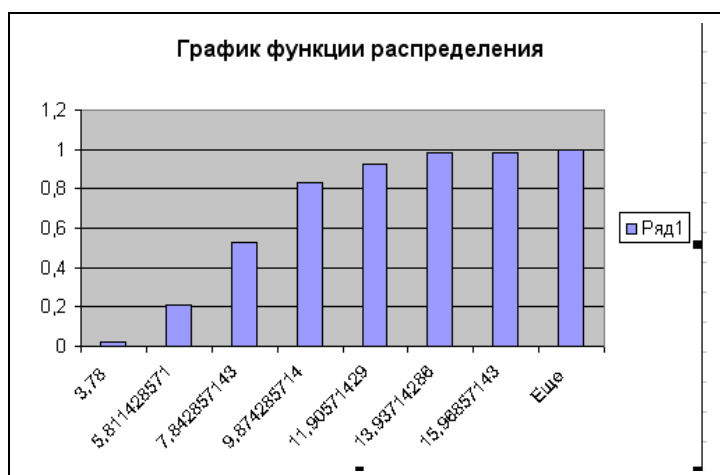


Рисунок 8 – График функции распределения

Построение гистограммы и функции распределения для переменной X4 (переменная 2) аналогично построению для величины X1.

3 Оценка основных выборочных характеристик: среднего значения, дисперсии, среднеквадратического отклонения. Нахождение их посредством электронной таблицы Excel.

Нахождение основных выборочных характеристик генеральной совокупности также проводится с использованием меню «Анализ данных» (рисунок 2) и категории «Описательная статистика», в диалоговом окне которого помимо параметров рассмотренных в п.1 (входной и выходной интервалы, новый рабочий лист, новая рабочая книга) указывают:

- i. группирование - устанавливают переключатель в положение «По столбцам» или «По строкам» в зависимости от расположения данных во входном диапазоне;
- ii. метки в первой строке/Метки в первом столбце - если первая строка исходного диапазона содержит названия столбцов, устанавливают переключатель в положение «Метки» в первой строке. Если названия строк находятся в первом столбце входного диапазона, устанавливают переключатель в положение «Метки» в первом столбце. Если входной диапазон не со-

- держит меток, то необходимые заголовки в выходном диапазоне будут созданы автоматически;
- iii. уровень надежности - устанавливают флажок, если в выходную таблицу необходимо включить строку для уровня надежности. В поле вводят требуемое значение. Например, значение 95% вычисляет уровень надежности среднего со значимостью 0,05;
 - iv. К-ый наибольший - устанавливают флажок, если в выходную таблицу необходимо включить строку для k-го наибольшего значения для каждого диапазона данных. В соответствующем окне вводят число k. Если k равно 1, эта строка будет содержать максимум из набора данных;
 - v. К-ый наименьший - устанавливают флажок, если в выходную таблицу необходимо включить строку для k-го наименьшего значения для каждого диапазона данных. В соответствующем окне вводят число k. Если k равно 1, эта строка будет содержать минимум из набора данных;
 - vi. итоговая статистика - устанавливают флажок, если в выходном диапазоне необходимо получить по одному полю для каждого из следующих видов статистических данных: Среднее, Стандартная ошибка (среднего), Медиана, Мода, Стандартное отклонение, Дисперсия выборки, Эксцесс, Асимметричность, Интервал, Минимум, Максимум, Сумма, Счет, Наибольшее (#), Наименьшее (#), Уровень надежности.

Для анализируемых данных укажем входной интервал «A1:B53», выходной интервал – свободную ячейку «C1» и выбираем пункты «Итоговая статистика» и «уровень надежности =95%» (рисунок 9).

Щелчок на кнопке ОК приведет к выводу на экран окна итогов по столбцам (рисунок 10).

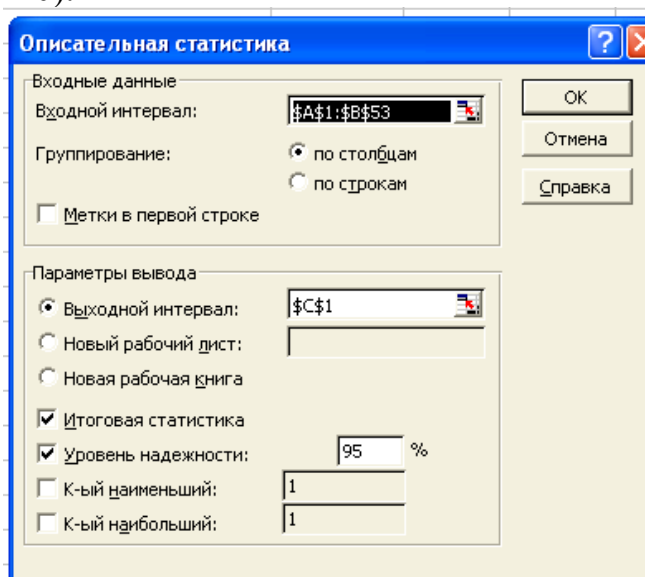


Рисунок 9 – Окно «Описательная статистика»

Столбец1		Столбец2	
Среднее	7,970377	Среднее	0,30245283
Стандартн	0,358542	Стандартная ошибка	0,01448914
Медиана	7,37	Медиана	0,31
Мода	5,22	Мода	0,31
Стандартн	2,610225	Стандартное отклонение	0,10548251
Дисперси	6,813273	Дисперсия выборки	0,01112656
Эксцесс	3,026449	Эксцесс	0,68504366
Асимметр	1,296765	Асимметричность	-0,40075604
Интервал	14,22	Интервал	0,5
Минимум	3,78	Минимум	0,01
Максимум	18	Максимум	0,51
Сумма	422,43	Сумма	16,03
Счет	53	Счет	53
Уровень н	0,719467	Уровень надежности(95,0%)	0,02907456

Рисунок 10 – Окно вывода итогов меню «Описательная статистика»

4 Проверка гипотезы о законе распределения генеральной совокупности с помощью критерия согласия

Для проверки гипотезы о законе распределения выдвигают нулевую и альтернативную гипотезы

$$H_0 : F_{\xi}(x) \in [F_{\text{mod}}(x; \theta)] \quad H_1 : F_{\xi}(x) \notin [F_{\text{mod}}(x; \theta)].$$

Если нулевая гипотеза истинна, то распределение критической статистики:

$$\chi^2_{\text{набл}} = \sum_{j=1}^s \frac{(v_j - np_j)^2}{np_j} \quad (2)$$

сходится (при $n \rightarrow \infty$) к $\chi^2(l - k - 1)$ - распределению,

где l – общее число интервалов группирования;

k - число неизвестных параметров, оцененных по выборке.

Далее по заданному уровню значимости критерия α и числом степеней свободы $l - k - 1$ из таблиц χ^2 - распределения находят $100(1 - \alpha/2)\%$ и $100\alpha/2\%$ -ые точки $\chi^2_{1 - \alpha/2}(l - k - 1)$ и $\chi^2_{\alpha/2}(l - k - 1)$ и если

$$\chi^2_{1 - \alpha/2}(l - k - 1) < \chi^2_{\text{набл}} < \chi^2_{\alpha/2}(l - k - 1), \quad (3)$$

то нулевая гипотеза принимается - выборочные данные распределены по нормальному закону.

Проверка гипотезы подразумевает выполнение следующих действий: вычисление точечных характеристик выборочных данных, группировка их, вычисление значения вероятности гипотетического распределения.

Для определения значений функции нормального распределения воспользуемся найденными ранее в пунктах «Построение гистограммы» и «Определение основных выборочных характеристик» точечными характеристиками и группировкой исходных данных.

Для выполнения данного раздела в Excel существует специализированная функция «НОРМРАСП» в меню «Функция», в диалоговом окне ввода данных которой задаются (рисунок 11):

- 1) значение, для которого вычисляется функция - в примере первое значение находящееся в столбце «Карман»;
- 2) среднее значение – для величины $X_1 = 7,97$;
- 3) среднеквадратическое отклонение - для величины $X_1 = 2,61$;
- 4) значение интегральной – либо ИСТИНА либо ЛОЖЬ.

Для каждого значение КАРМАНА определим значение функции и найдем вероятность попадания в соответствующий интервал группирования - разность найденных значений функций и ее же значений, сдвинутых на один шаг.

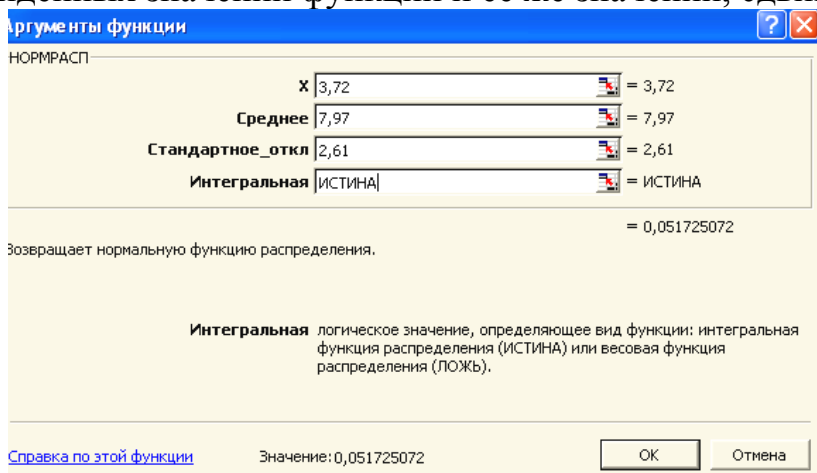


Рисунок 11 – Меню функции «НОРМРАСП»

Для удобства вычисления $\chi^2_{набл}$ составим таблицу в Excel - рисунок 12.

В первом столбце приводятся значения кармана, во втором - частоты, в третьем столбце расположены найденные значения эмпирической функции распределения для каждого кармана, в четвертом p_j - разность между текущими и предыдущими значениями столбца 3, в пятом - произведение числа выборочных n данных и соответствующего значения p_j , и в шестом столбце находятся значения, найденные по формуле (2) для каждого кармана.

C	D	E	F	G	H
<i>Карман</i>	<i>Частота</i>	$\Phi(X)$	$p_j = \Phi(x) - \Phi(x-1)$	np_j	$(v - np_j)^2 / np_j$
3,78	1	0,054206859	0	0	0
5,8114286	10	0,20410746	0,149900601	7,9447318	0,531689083
7,8428571	17	0,480573647	0,276466187	14,652708	0,376024705
9,8742857	16	0,767186193	0,286612547	15,190465	0,043141994
11,905714	5	0,934214586	0,167028392	8,8525048	1,676564269
13,937143	3	0,988880553	0,054665967	2,8972963	0,003640655
15,968571	0	0,998910076	0,010029523	0,5315647	0,531564728
			Σ		3,162625435

Рисунок 12– Данные для проверки гипотезы о законе распределения

Таким образом, $\chi^2_{набл} = 3,1626$. Для нахождения $100(1-\alpha/2)\%$ и $100\alpha/2\%$ -ых точек «хи-квадрат» распределения воспользуемся встроенной функцией «ХИ2ОБР» меню «Функция». Аргументами данного модуля являются вероятность и степени свободы (рисунок 13). При $\alpha = 0,05$, количестве интервалов группирования $l = 7$, числе оцененных параметров по выборке $k=2$ значения критических точек будут:

$$\chi^2_{1-0,05/2}(7 - 2 - 1) = 0,484 \quad \text{и} \quad \chi^2_{0,05/2}(7 - 2 - 1) = 11,14.$$

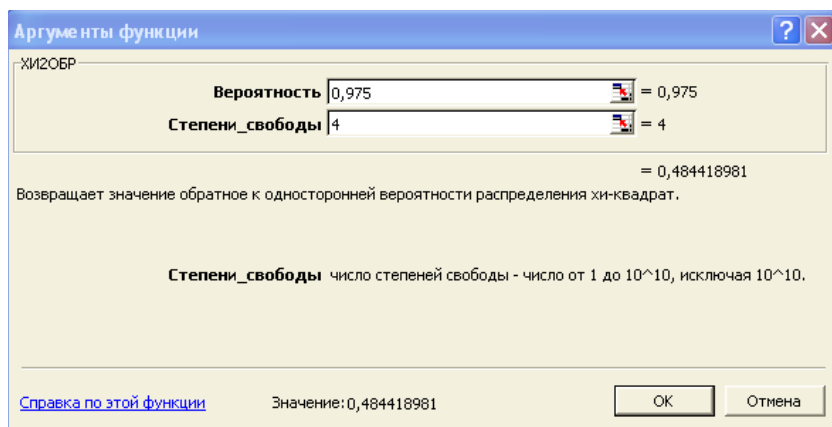


Рисунок 13 – Вычисление критических значений «хи-квадрат» распределения

Так как наблюдаемое значение попадает в интервал (3), то делаем вывод о том, что нулевая гипотеза принимается и генеральная совокупность, представленная выборкой X_1 подчинена нормальному закону распределения.

Аналогично проверяется гипотеза о законе распределения генеральной совокупности по переменной X_4 .

5 Построение доверительных интервалов для основных характеристик генеральной совокупности

Построение интервальной оценки для математического ожидания основано на статистике:

$$\frac{\bar{X} - \mu}{S} \sqrt{n - 1},$$

которая при случайной выборке из генеральной совокупности с нормальным распределением имеет распределение Стьюдента с (n-1) степенями свободы. И поэтому можно утверждать, что доверительный интервал для математического ожидания примет следующий вид:

$$\bar{X} - t_{\alpha} \cdot \frac{S}{\sqrt{n - 1}} \leq \mu \leq \bar{X} + t_{\alpha} \cdot \frac{S}{\sqrt{n - 1}}, \quad (4)$$

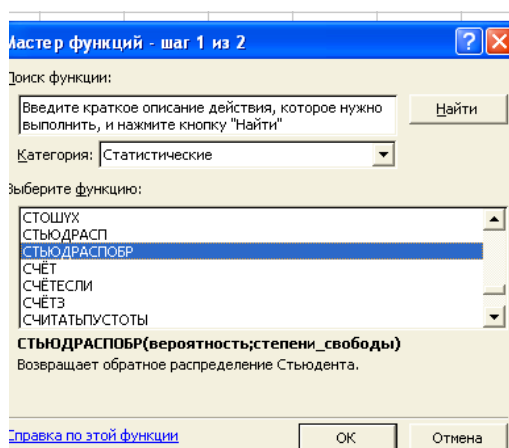
где \bar{X} – выборочное среднее значение;

S – выборочное среднеквадратическое отклонение;

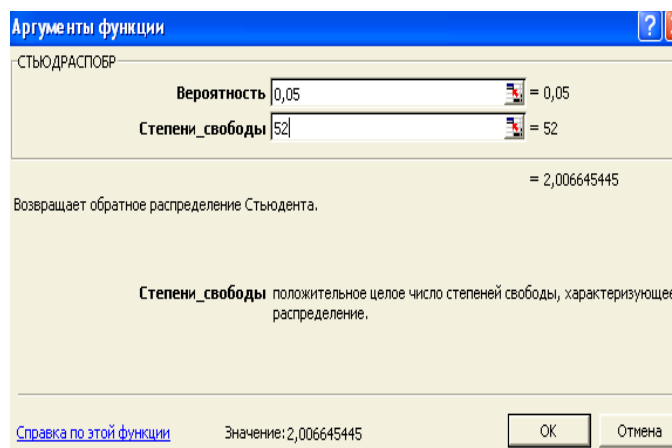
n – объем выборки;

t_{α} – статистика, имеющая распределение Стьюдента.

Значение статистики t_{α} можно определить с помощью электронной таблицы Excel. Для этого в меню «Вставка» выбирают пункт «Функция», в категории «Статистические» функцию «СТЮДОБР», нажимают кнопку «ОК» (рисунок 14а). Вычисление значения t-статистики производится после задания значения $\alpha = 1 - \gamma$ в следующем окне (рисунок 14б). При $\gamma = 0,95$, $\alpha = 1 - 0,95 = 0,05$, значение $t_{\alpha} = 2$.



а



б

Рисунок 14 – Определение статистики t_{α}

Подставляя найденные значения в формулу (4), можно утверждать, что с вероятностью 95% математическое ожидание генеральной совокупности X_1 лежит в интервале:

$$7,97 - 2 \cdot \frac{2,61}{\sqrt{53-1}} \leq \mu \leq 7,97 + 2 \cdot \frac{2,61}{\sqrt{53-1}}$$

$$7,25 \leq \mu \leq 8,69$$

Для совокупности X_4 :

$$0,302 - 2 \cdot \frac{0,1}{\sqrt{53-1}} \leq \mu \leq 0,302 + 2 \cdot \frac{0,1}{\sqrt{53-1}}$$

$$0,272 \leq \mu \leq 0,332$$

Построим доверительные интервалы для дисперсии и среднеквадратического отклонения.

Доверительный интервал для дисперсии имеет вид:

$$\frac{n \cdot S^2}{\chi_{q^2(n-1)}} \leq D \leq \frac{n \cdot S^2}{\chi_{1-q^2(n-1)}}, \quad (5)$$

где $\chi_{q^2(n-1)}$ и $\chi_{1-q^2(n-1)}$ распределены по закону χ^2 и находятся по таблице χ^2 - распределения с числом степеней свободы $\nu = (n-1)$, $q = (1-\gamma)/2$ и $1-q$.

Значение данной статистики также можно вычислить посредством электронной таблицы Excel. Для этого в меню «Вставка» выбирают «Функцию», а в категории «Статистические» выбирают функцию «ХИ2ОБР», которая вычисляет значение статистики χ^2 в зависимости от числа степеней свободы и q .

Для нашего примера вычислим необходимые параметры: число степеней свободы $\nu = 53-1 = 52$; $q = (1-\gamma)/2 = (1-0,95)/2 = 0,025$ и $1-q = 1-0,025 = 0,975$. Посредством Excel найдем значения $\chi^2_{0,025}(52) = 73,8$ и $\chi^2_{0,975}(52) = 33,96$ (рисунок 15).

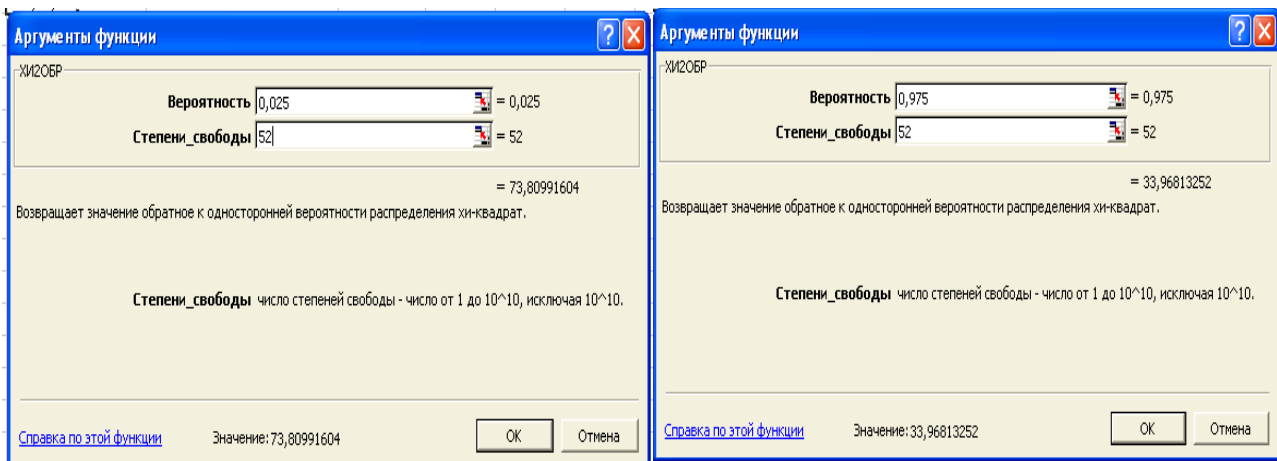


Рисунок 15 – Вычисление критических значений $\chi_q^2(n-1)$ и $\chi_{1-q}^2(n-1)$

Полученные значения подставим в формулу (5). Для величины X1:

$$\frac{53 \cdot 6,81}{73,8} \leq D \leq \frac{53 \cdot 6,81}{33,96}$$

$$4,89 \leq D \leq 10,62$$

Аналогично для величины X4:

$$\frac{53 \cdot 0,011}{73,8} \leq D \leq \frac{53 \cdot 0,011}{33,96}$$

$$0,008 \leq D \leq 0,017$$

Доверительные интервалы для среднеквадратических отклонений получают путем вывода из-под квадратного корня полученных граничных значений доверительных интервалов дисперсии. То есть доверительными интервалами для среднеквадратических отклонений величин X и Y соответственно будут:

$$2,21 \leq \sigma \leq 3,25$$

$$0,089 \leq \sigma \leq 0,13$$

6 Оценка парного коэффициента корреляции (нахождение его в таблице Excel). Проверка его значимости и построение доверительного интервала

Основная задача двумерного корреляционного анализа признаков (X, Y) состоит в оценке пяти параметров:

\bar{X} - оценка среднего значения величины X,

\bar{Y} - оценка среднего значения величины Y,

$\bar{X^2}$ - оценка среднего значения величины X^2 ,

$\overline{Y^2}$ - оценка среднего значения величины Y^2 ,

\overline{XY} - оценка среднего значения произведения величин X и Y .

Откуда:

$S_X^2 = \overline{X^2} - (\overline{X})^2$ - оценка дисперсии величины X ,

$S_Y^2 = \overline{Y^2} - (\overline{Y})^2$ - оценка дисперсии величины Y ,

$r = r_{XY} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{S_X \cdot S_Y}$ - оценка для парного коэффициента корреляции.

Парный коэффициент корреляции служит мерой линейной статистической зависимости между величинами и является одним из основных показателей взаимосвязи между ними, принимающий значения от -1 до 1.

Проверка значимости парного коэффициента корреляции.

С этой целью проверки гипотезы о значимости коэффициента корреляции выдвигают нулевую гипотезу $H_0: \rho = 0$, при альтернативной гипотезе $H_1: \rho \neq 0$. Для ее проверки используют статистику

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2} \quad (6)$$

Критическое значение t находят по таблицам распределения Стьюдента с $\nu = n - 2$ степенями свободы и уровнем значимости α . Затем сравнивают наблюдаемое и критические значения. Если $t > t_{кр}$, то гипотеза о незначимости коэффициента корреляции отвергается.

Интервальные оценки находят для значимых параметров связи. При определении с надежностью γ доверительного интервала для ρ используют Z – преобразование Фишера

$$Z_r - \frac{t_\gamma}{\sqrt{n-2}} \leq Z \leq Z_r + \frac{t_\gamma}{\sqrt{n-2}}, \quad (7)$$

где t_γ вычисляют по таблице нормального распределения с заданным γ ; значение Z_r определяют по таблице Z – преобразования по найденному значению r .

Обратный переход от Z к ρ осуществляют также по таблицам Z – преобразования Фишера, после чего получают интервальную оценку для ρ :

$$r_{\min} \leq \rho \leq r_{\max} \quad (8)$$

Вычисление парного коэффициента корреляции производится с помощью функции «Корреляция» меню «Анализ данных». В окне для ввода параметров указываем входной интервал: «A1:B53» и выходной интервал «C1», после чего активизируем кнопку «ОК». В итоговом окне (рисунок 16) выводится матрица парных коэффициентов корреляции, на главной диагонали которой расположе-

ны единицы, а на других позициях парные коэффициенты корреляции между соответствующими блоками данных.

20		Столбец 1	Столбец 2
21	Столбец 1	1	
22	Столбец 2	-0,4915619	1

Рисунок 16 – Матрица парных коэффициентов корреляции

Таким образом, коэффициент корреляции между выборками X1 и X4 $r = -0,491$, свидетельствует об обратной и средней статистической связи между ними.

Проверим значимость найденного коэффициента корреляции, используя вычисленные характеристики в пункте 2 и формулу (6).

Выдвигаем нулевую гипотезу $H_0: \rho = 0$, при альтернативной гипотезе $H_1: \rho \neq 0$. Вычислим статистику:

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2} = \frac{-0,49}{\sqrt{1 - (-0,49)^2}} \cdot \sqrt{53 - 2} = 4,02$$

Критическое значение статистики $t_{кр}$ определим с помощью электронной таблицы Excel. Для этого, в меню «Вставка» выбираем меню «Функция», а в категории «Статистические» - функцию «СТЪЮДРАСПОБР», в окне ввода аргументов указываем значение $\alpha=0,05$ и число степеней свободы $\nu = n - 2 = 53 - 2 = 51$ (рисунок 17). Критическое значение статистики $t=2,007$.

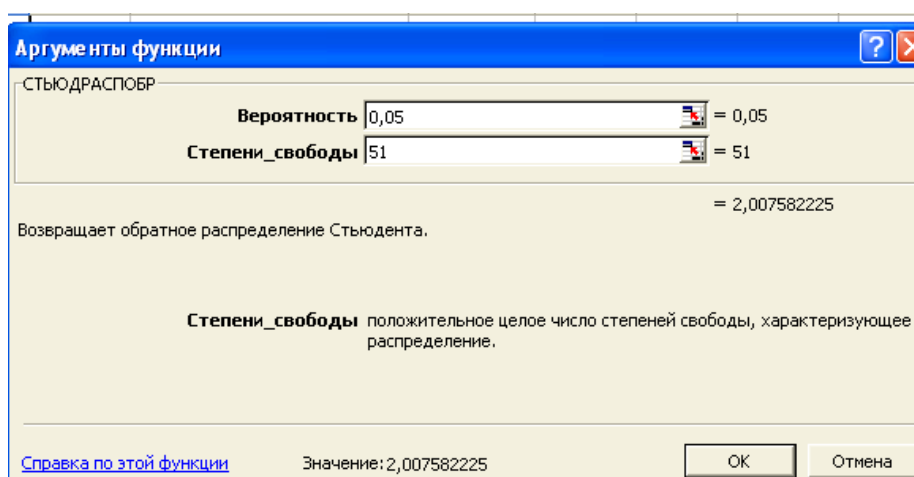


Рисунок 17 - Вычисление критического значения статистики t

Так как критическое значение статистики $t_{кр} <$ наблюдаемого значения t гипотеза о значимости коэффициента корреляции принимается и для него целесообразно построить доверительный интервал.

Для его построения на первом этапе вычислим величину Z_r по формуле:

$$Z_r = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln \frac{1+(-0,491)}{1-(-0,491)} = -0,53 \quad (9)$$

и подставляя в (7) получаем доверительный интервал для величины Z_r (значение статистики $t = 1,64$, рисунок 11):

$$-0,53 - \frac{1,64}{\sqrt{53-2}} \leq Z \leq -0,53 + \frac{1,64}{\sqrt{53-2}}$$

$$-0,76 \leq Z \leq -0,3$$

Обратный переход от Z_r к r совершим также с использованием формулы (9):

$$-0,76 = \frac{1}{2} \ln \frac{1+r}{1-r} \quad -0,3 = \frac{1}{2} \ln \frac{1+r}{1-r}$$

$$r_{\min} = -0,996 \quad r_{\max} = -0,31$$

Таким образом, доверительный интервал для коэффициента корреляции имеет вид:

$$-0,996 \leq \rho \leq -0,31$$

7 Оценка уравнения регрессии. Проверка значимости и построение доверительных интервалов для значимых параметров регрессии

Имея вычисленные ранее показатели, можно получить уравнения линий регрессии, которые показывают изменение условных математических ожиданий в зависимости от изменения соответствующих значений случайных переменных:

$$\text{уравнение регрессии } Y \text{ на } X: \hat{Y} = \bar{Y} + b_{YX}(x - \bar{X}) \quad (10)$$

$$\text{уравнение регрессии } X \text{ на } Y: \hat{X} = \bar{X} + b_{XY}(y - \bar{Y})$$

И тогда оценка коэффициентов уравнения регрессии:

$$b_{YX} = r \frac{S_Y}{S_X} \quad b_{XY} = r \frac{S_X}{S_Y} \quad (11)$$

В двумерном случае для наглядности наносят на график точки (x_i, y_j) - корреляционное поле; строят эмпирическую регрессию – ломанную, соединяющую точки (x_i, \bar{Y}_i) , где \bar{Y}_i - среднее арифметическое наблюдаемых значений Y ; оценки функции регрессии $\hat{Y} = \bar{Y} + b_{YX}(x - \bar{X})$ (рисунок 18).

Коэффициент уравнения регрессии b_{YX} показывает: насколько в среднем изменится значения показателя Y при изменении X на одну единицу своего измерения.

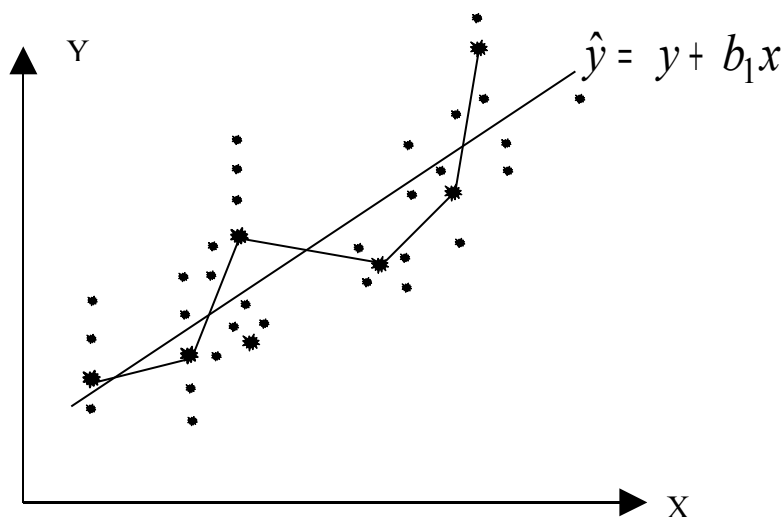


Рисунок 18 - График функции регрессии

Значимость коэффициентов регрессии проверяют с использованием статистики:

$$F_n = \frac{Q_R}{Q_{ocm} / (n - 2)}, \quad (12)$$

где

$$Q_R = \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2; \quad Q_{ocm} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (13)$$

которая при справедливости гипотезы $H_0 : \beta_1 = 0$ распределена по закону Фишера – Снедекора с заданным уровнем значимости α и числом степеней свободы $v_1 = v_2 = n - 1$

Интервальная оценка для коэффициентов регрессии имеет вид:

$$\beta_{YX} \in b_{YX} \pm t \left\{ \frac{S_Y^2 (1 - r^2)}{S_X^2 \cdot \sqrt{n - 2}} \right\}^{1/2} \quad (14)$$

Аналогично для коэффициента β_{XY} :

$$\beta_{XY} \in b_{XY} \pm t \left\{ \frac{S_X^2 (1 - r^2)}{S_Y^2 \cdot \sqrt{n - 2}} \right\}^{1/2} \quad (15)$$

Коэффициенты регрессии находятся в Excel с помощью меню «Сервис», функции «Анализ данных», и в предложенном списке методов анализа - метод «Регрессия», где заполняют «входной интервал Y», «входной интервал X» (для рассматриваемого примера соответственно X1 и X4), «выходной интервал», «Уровень надежности» и нажимают «ОК» (рисунок 19).

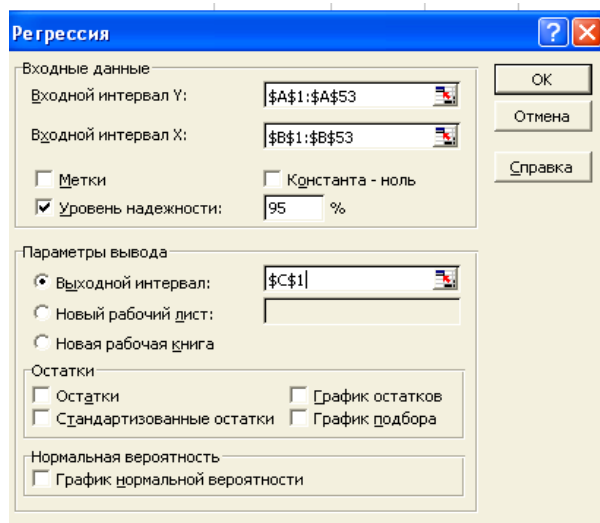


Рисунок 19 - Ввод данных для вычисления коэффициентов регрессии

Вывод итогов (рисунок 20) состоит из трех таблиц: в верхней указаны показатели качества построенного уравнения, в средней – дисперсионный анализ, в нижней - коэффициенты регрессии (y- пересечение – это свободный член уравнения, «переменная x» – значение коэффициента регрессии при переменной x). Границы доверительных интервалов находятся в нижней таблице – на пересечении столбцов «нижние 95%», «верхние 95%» и соответствующих строк – «y – пересечение» и «переменная x».

Вывод итогов									
C	D	E	F	G	H	I	J	K	L
Вывод итогов									
<i>Регрессионная статистика</i>									
Множественный R	0,491562								
R-квадрат	0,241633								
Нормированный R-квадрат	0,226763								
Стандартная ошибка	2,295272								
Наблюдения	53								
<i>Дисперсионный анализ</i>									
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>значимость F</i>				
Регрессия	1	85,60825198	85,608252	16,24977	0,000186				
Остаток	51	268,6819405	5,26827334						
Итого	52	354,2901925							
<i>Коэффициент</i> <i>Стандартная ошибка</i> <i>t-статистика</i> <i>P-значение</i> <i>Вероятность > F</i> <i>Вероятность < F</i> <i>Вероятность < F</i> <i>Вероятность < F</i> <i>Вероятность < F</i>									
Y-пересечение	11,64941	0,965583846	12,0646254	1,47E-16	9,710918	13,5879	9,710918	13,5879	
Переменная X 1	-12,164	3,017532919	-4,03110086	0,000186	-18,2219	-6,10603	-18,2219	-6,10603	

Рисунок 20 – Вывод итогов меню «Регрессия»

Уравнение регрессии для рассматриваемого варианта имеет вид:

$$X1 = 11,64 - 12,164 \cdot X4$$

Значимость построенного уравнения регрессии можно проверить посредством средней таблицы – дисперсионного анализа, в предпоследнем и последнем столбцах которой указываются значение статистики F и вероятность принятия нулевой гипотезы (столбец значимость F). Так как эта вероятность имеет значение значительно меньше уровня 0,05, делаем вывод, что построенное уравнение значимо.

На основании расчетных данных доверительный интервал для коэффициента регрессии имеет вид:

$$\beta_{YX} \in [-18,22; -6,1]$$

Аналогично для уравнения X4 на X1 (в меню «Регрессия» меняем входной интервал X и входной интервал Y местами):

$$X4 = 0,46 - 0,02 \cdot X1 \text{ - уравнение также является значимым}$$

$$\beta_{XY} \in [-0,03; -0,01]$$

Вывод: анализируя построенные уравнения регрессии, можно сказать, что: при изменении величины X4 на одну единицу аргумент X1 уменьшится на 12,164 единицы (для первого уравнения). И аналогично для второго: при изменении X1 на 1 единицу величина X4 уменьшится на 0,02.

8 Вопросы к защите

- 1) Статистическая совокупность, группировки; вариационный ряд частот, относительных частот; плотностей относительных частот.
- 2) Гистограмма, статистическая (кумулятивная) функция распределения.

- 3) Выборочная средняя, стандартные отклонения, мода, медиана.
- 4) Смещенные и несмещенные, состоятельные оценки параметров генеральной совокупности.
- 5) Интервальные оценки: доверительная вероятность, доверительный интервал.
- 6) Оценки законов распределения: критерии согласия.
- 7) Проверка гипотез: уровень значимости, мощность критерия, критические области и их нахождение.
- 8) Коэффициент корреляции.
- 9) Проверка значимости и интервальные оценки параметров линейной корреляционной зависимости.
- 10) Проверка значимости и интервальное оценивание коэффициентов и уравнения регрессии.

Список использованных источников

- 1 **Айвазян С.А.** Теория вероятностей и прикладная статистика [Текст]: учебник для вузов/ С.А.Айвазян, В.С.Мхитарян – М.: ЮНИТИ, 2001, том 1.-656 с.
- 2 **Шишкин Е.В.** Математические методы и модели в управлении [Текст]: учебное пособие/ Е.В. Шишкин, А.Г. Чхартишвили. – М.: Дело, 2000. – 440 с.
- 3 **Колемаев В.А.** Теория вероятностей и математическая статистика [Текст]: учебник / В.А.Колемаев, В.Н. Калинина – М.:ИНФРА-М, 1999. -302 с.

4 **Кремер Н.Ш.** Теория вероятностей и математическая статистика /Н.Ш. Кремер – М.: ЮНИТИ, 2002.

5 **Гмурман В.Е.** Теория вероятностей и математическая статистика/В.Е. Гмурман – М.: Высшая школа, 2001.

6 **Колемаев В.А.** Теория вероятностей и математическая статистика/ В.А. Колемаев, О.В.Староверов, В.Б.Турундаевский – М.: Высшая школа, 1990.

7 **Калинина В.Н.** Математическая статистика/ В.Н.Калинина, В.Н.Панкин – М.: Высшая школа, 2001.

8 **Мхитарян В.С.** Задачник по дисперсионному, корреляционному и регрессивному анализу/ В.С. Мхитарян, Л.И.Трошин – М.: МЭСИ, 1996.

Приложение А
(обязательное)

Исходные данные для анализа

Таблица А.1 -Выборочные данные

№ объекта	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	9,26	204,2	13,26	0,23	0,78	0,40	1,37	1,23	0,23	1,45
2	9,38	209,6	10,16	0,24	0,75	0,26	1,49	1,04	0,39	1,30
3	12,11	222,6	13,72	0,19	0,68	0,40	1,44	1,80	0,43	1,37
4	10,81	236,7	12,85	0,17	0,70	0,50	1,42	0,43	0,18	1,65
5	9,35	62,0	10,63	0,23	0,62	0,40	1,35	0,88	0,15	1,91
6	9,87	53,1	9,12	0,43	0,76	0,19	1,39	0,57	0,34	1,68
7	8,17	172,1	25,83	0,31	0,73	0,25	1,16	1,72	0,38	1,94
8	9,12	56,5	23,39	0,26	0,71	0,44	1,27	1,70	0,09	1,89
9	5,88	52,6	14,68	0,49	0,69	0,17	1,16	0,84	0,14	1,94
10	6,30	46,6	10,05	0,36	0,73	0,39	1,25	0,60	0,21	2,06
11	6,22	53,2	13,99	0,37	0,68	0,33	1,13	0,82	0,42	1,96
12	5,49	30,1	9,68	0,43	0,74	0,25	1,10	0,84	0,05	1,02
13	6,50	146,4	10,03	0,35	0,66	0,32	1,15	0,67	0,29	1,85
14	6,61	18,1	9,13	0,38	0,72	0,02	1,23	1,04	0,48	0,88
15	4,32	13,6	5,37	0,42	0,68	0,06	1,39	0,66	0,41	0,62

Продолжение таблицы А.1

1	2	3	4	5	6	7	8	9	10	11
16	7,37	89,8	9,86	0,30	0,77	0,15	1,38	0,86	0,62	1,09
17	7,02	62,5	12,62	0,32	0,78	0,08	1,35	0,79	0,56	1,60
18	8,25	46,3	5,02	0,25	0,78	0,20	1,42	0,34	1,76	1,53
19	8,15	103,5	21,18	0,31	0,81	0,20	1,37	1,60	1,31	1,40
20	8,72	73,3	25,17	0,26	0,79	0,30	1,41	1,46	0,45	2,22
21	6,64	76,6	19,40	0,37	0,77	0,24	1,35	1,27	0,50	1,32
22	8,10	73,01	21,0	0,29	0,78	0,10	1,48	1,58	0,77	1,48
23	5,52	32,3	6,57	0,34	0,72	0,11	1,24	0,68	1,20	0,68
24	9,37	199,6	14,19	0,23	0,79	0,47	1,40	0,86	0,21	2,30
25	13,17	598,1	15,81	0,17	0,77	0,53	1,45	1,98	0,25	1,37
26	6,67	71,2	5,23	0,29	0,80	0,34	1,40	0,33	0,15	1,51
27	5,68	90,8	7,99	0,41	0,71	0,20	1,28	0,45	0,66	1,43
28	5,22	82,1	17,50	0,41	0,79	0,24	1,33	0,74	0,74	1,82
29	10,02	76,2	17,16	0,22	0,76	0,54	1,22	0,03	0,32	2,62
30	8,16	119,5	14,54	0,29	0,78	0,40	1,28	0,99	0,89	1,75
31	3,78	21,9	6,24	0,51	0,62	0,20	1,47	0,24	0,23	1,54
32	6,48	48,4	12,08	0,36	0,75	0,64	1,27	0,57	0,32	2,25
33	10,44	173,5	9,49	0,23	0,71	0,42	1,51	1,22	0,54	1,07
34	7,65	74,1	9,28	0,26	0,74	0,27	1,46	0,68	0,75	1,44
35	8,77	68,6	11,42	0,27	0,65	0,37	1,27	1,00	0,16	1,40
36	7,00	60,8	10,31	0,29	0,66	0,38	1,43	0,81	0,24	1,31
37	11,06	355,6	8,65	0,01	0,84	0,35	1,50	1,27	0,59	1,12
38	9,02	264,8	10,94	0,02	0,74	0,42	1,35	1,14	0,56	1,16
39	13,28	526,6	9,87	0,18	0,75	0,32	1,41	1,89	0,63	0,88

Продолжение таблицы А.1

(1)	(2)	(3)	(4)	(5)	(6)	(7)		(9)	(10)	(11)
40	9,27	118,6	6,14	0,25	0,75	0,33	1,47	0,67	1,10	1,07
41	6,70	37,1	12,93	0,31	0,79	0,29	1,35	0,96	0,39	1,24
42	6,69	57,7	9,78	0,38	0,72	0,30	1,40	0,67	0,73	1,49
43	9,42	51,6	13,22	0,24	0,70	0,56	1,20	0,98	0,28	2,03
44	7,24	64,7	17,29	0,31	0,66	0,42	1,15	1,16	0,10	1,84
45	5,39	48,3	7,11	0,42	0,69	0,26	1,09	0,54	0,68	1,22
46	5,61	15,0	22,49	0,51	0,71	0,16	1,26	1,23	0,87	1,72
47	5,59	87,5	12,14	0,31	0,73	0,45	1,36	0,78	0,49	1,75
48	6,57	108,4	15,25	0,37	0,65	0,31	1,15	1,16	0,16	1,46
49	6,54	267,3	31,34	0,16	0,82	0,08	1,87	4,44	0,85	1,60
50	4,23	34,2	11,56	0,18	0,80	0,68	1,17	1,06	0,13	1,47
51	5,22	26,8	30,14	0,43	0,83	0,03	1,61	2,13	0,49	1,38
52	18,00	43,6	19,71	0,40	0,70	0,02	1,34	1,21	0,09	1,41
53	11,03	72,0	23,56	0,31	0,74	0,22	1,22	2,20	0,79	1,39

X1 – производительность труда;

X2 – индекс снижения себестоимости продукции;

X3 – рентабельность;

X4 – трудоемкость единицы продукции;

X5 – удельный вес рабочих в составе ППП;

X6 – удельный вес покупных изделий;

X7 – коэффициент сменности оборудования;

X8 – премии и вознаграждения на одного работника;

X9 – удельный вес потерь от брака;

X10 – фондоотдача.

Таблица А.2 – Варианты заданий

№ варианта	Номера признаков
0	X1, X4
1	X1, X6
2	X1, X8
3	X1, X10
4	X2, X3
5	X2, X5
6	X2, X7
7	X2, X9
8	X3, X1
9	X3, X4
10	X3, X5
11	X3, X7
12	X3, X8
13	X4, X6,
14	X4, X7,
15	X4, X10,
16	X5, X6
17	X5, X7
18	X5, X8
19	X5, X9
20	X6, X7
21	X6, X8
22	X6, X10
23	X7, X8
24	X7, X9
25	X8, X10

