БЕЛКООПСОЮЗ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКИЙ ТОРГОВО-ЭКОНОМИЧЕСКИЙ
УНИВЕРСИТЕТ ПОТРЕБИТЕЛЬСКОЙ КООПЕРАЦИИ»

Кафедра высшей математики

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Пособие
для студентов экономических специальностей

Гомель 2010

УДК 519.2
ББК 22.171
Т 33

Авторы-составители: Л. П. Авдашкова, канд. физ.-мат. наук, доцент кафедры высшей математики Белорусского торгово-экономического университета потребительской кооперации;
Л. Н. Марченко, канд. техн. наук, доцент кафедры экономической кибернетики и теории вероятностей Гомельского государственного университета им. Ф. Скорины

Рецензенты: Л. Л. Великович, канд. физ.-мат. наук, доцент кафедры высшей математики Гомельского государственного технического университета им. П. О. Сухого;
И. А. Кузменкова, канд. физ.-мат. наук, доцент кафедры высшей математики Белорусского торгово-экономического университета потребительской кооперации

Рекомендовано к изданию научно-методическим советом учреждения образования «Белорусский торгово-экономический университет потребительской кооперации». Протокол № 2 от 13 декабря 2005 г.

Т 33 **Теория** вероятностей и математическая статистика : пособие для студентов экономических специальностей / авт.-сост. : Л. П. Авдашкова, Л. Н. Марченко. – Гомель : учреждение образования «Белорусский торгово-экономический университет потребительской кооперации», 2010. – 124 с.
ISBN 978-985-461-725-1

УДК 519.2
ББК 22.171

ISBN 978-985-461-725-1

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Пособие «Теория вероятностей и математическая статистика» предназначено для студентов высших учебных заведений и ориентировано на использование персонального компьютера при изучении материала.

В теме 1 пособия приведены краткие теоретические сведения по работе с программой *MS Excel*. В темах 2–6 излагается материал по математической статистике. Темы содержат теоретические сведения, вопросы для самоконтроля, описание технологии работы с табличным процессором и порядка выполнения лабораторных работ с помощью *Excel*, анализ полученных результатов.

Для организации самостоятельной работы даны индивидуальные задания, которые позволяют закрепить изученный материал.

В пособии приводится список рекомендуемой литературы, который предлагается использовать при изучении дисциплины.

Данное пособие адресовано преподавателям и студентам для проведения практических занятий и организации самостоятельной учебной работы.

Тема 1. ОЗНАКОМЛЕНИЕ С ТАБЛИЧНЫМ ПРОЦЕССОРОМ *MICROSOFT EXCEL*

План

1. Ознакомление с работой в редакторе *MS Excel*.
2. Работа с формулами.

1. Ознакомление с работой в редакторе *MS Excel*

Табличные процессоры (*электронные таблицы*) представляют собой пакеты программ для решения задач, исходные данные которых и результаты вычислений заданы в виде таблиц и являются универсальным средством для автоматизации расчетов с применением больших объемов данных.

Для запуска программы *Excel* в меню *Пуск* выбирается команда *Программы*, вкладка *Microsoft Office*, а затем *Microsoft Excel*, или нужно дважды щелкнуть левой кнопкой мыши на ярлыке  *MS Excel* на *Рабочем столе*. На экране появится окно документа, который называется рабочей книгой и имеет расширение *.xls. У новой книги есть имя *Книга 1*. Внешний вид окна программы *Excel* изображен на рис. 1.

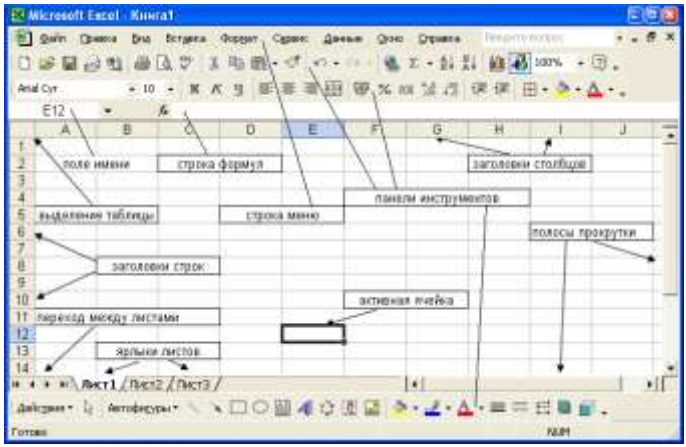


Рис. 1. **Окно программы *Excel***

Окно *Excel* содержит области, приведенные ниже.

Строка заголовка включает название документа.

Строка меню содержит меню *Файл*, *Правка*, *Вид*, *Вставка*, *Формат*, *Сервис*, *Данные*, *Окно*, *Справка*. Например, для того чтобы изменить название книги, необходимо в строке меню выбрать *Файл*, щелкнув левой кнопкой мыши, затем – *Сохранить как*, определить нужную папку (или диск) для записи, дать название книге и выбрать *Сохранить*. Открыть ранее сохраненную книгу можно, выбрав в строке меню *Файл*, затем – *Открыть* (или, подведя курсор в панели инструментов к кнопке  , открыть документ, щелкнуть левой кнопкой мыши), в появившемся списке дисков, папок, документов выбрать необходимый документ.

Панели *Стандартная* и *Форматирование* содержат наиболее часто используемые команды.

Рабочая область состоит из строки формулы и строки состояния. Строка формулы содержит поле имени с адресом активной ячейки, поле для ввода формулы. Кнопки  – очистка,  – закрепление,  – вызов мастера функций появляются при вводе информации в ячейку.

Рабочий лист – это отдельная таблица в рабочей книге, состоящая из 16384 строк и 256 столбцов. Каждому столбцу в качестве заголовка соответствует буква, которая стоит вне сетки листа сверху на сером поле. Каждой строке соответствует натуральное число, которое располагается слева от сетки листа на сером поле в качестве заголовка строки. На пересечении строки и столбца находится ячейка, положение которой определяется адресом ячейки (буквой столбца и номером строки). Выделенная ячейка называется активной. В ячейки листа можно вводить 2 типа данных (константы и формулы). Константы разделяются на следующие основные категории: числовые значения содержат цифры от 0 до 9 и специальные символы (+; –; *E*; *e*; (,) ; %; /). Текстовые значения включают любой набор символов, значения даты и времени, логические и ошибочные значения. Книга содержит 16 рабочих листов, каждый из которых может быть отнесен к одному из следующих типов: рабочие листы, листы диаграмм, модуль *Visual Basic*, листы диалога, листы макросов.

Под рабочей областью находятся кнопки перехода между листами, ярлыки листов и горизонтальная полоса прокрутки листа. Кнопки перехода позволяют просмотреть содержание книги. Чтобы активизировать лист, необходимо щелкнуть кнопкой мыши по его ярлыку. Изменение названия листа осуществляется с помощью двойного щелчка левой кнопкой мыши по ярлыку этого листа и введения нового названия. Просмотреть содержание листа позволяют горизонтальная и вертикальная полосы прокрутки, расположенные снизу и справа соответственно.

Панель инструментов *Рисование* содержит команды для изображения графических объектов и включается кнопкой *Рисование* на панели инструментов *Стандартная*.

В *строке состояния* указываются выполняемые действия.

2. Работа с формулами

Формулы в программе *Excel* позволяют выполнять различные действия над данными. Они могут содержать имена или заголовки, представляющие ячейки или диапазоны ячеек. Формула должна начинаться со знака равенства, за которым следует набор участвующих в вычислениях величин. Ячейка, содержащая формулу, является зависимой. Если изменить данные в ячейках, на которые ссылается формула, то по ней можно посчитать новое значение. *Оператор* формулы указывает, какую операцию над элементами следует провести. В *Microsoft Excel* различают следующие четыре вида операторов:

- *арифметические* – сложение (+), вычитание (–), умножение (*), деление (/), процент (%), возведение в степень (^);
- *сравнения* – равно (=), больше (>), меньше (<), больше или равно (>=), меньше или равно (<=), не равно (< >);
- *текстовые* – объединение последовательности символов в одну последовательность (&);
- *адресные*, когда оператор диапазона (:) ссылается на все ячейки между границами диапазона включительно, оператор объединения (,) – на объединение ячеек диапазонов, оператор пересечения (пробел) – на общие ячейки диапазонов.

Для упрощения сложных вычислений используются функции.

Функция – это стандартная формула *Microsoft Excel*, которая возвращает результат выполнения определенных действий над значениями, называемыми аргументами функции. Двойным щелчком по кнопке f_x или командой *Функция* меню *Вставка* вызывается *Мастер функций*, содержащий весь список имеющихся в *Excel* стандартных функций. В одноименном диалоговом окне (рис. 2) из верхнего списка (вызов списка осуществляется щелчком по кнопке ▼ в окне *Категория*) выбирается категория функции, а из нижнего – имя функции двойным щелчком кнопки мыши на выделенной функции или одним щелчком и кнопкой *ОК*.

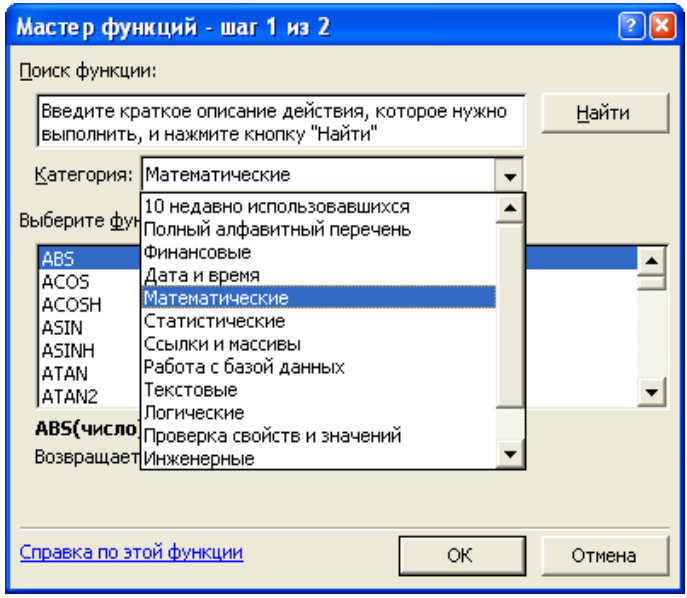


Рис. 2. **Диалоговое окно *Мастер функций***

Диалоговое окно функции содержит описание функции, аргументы, возвращаемое функцией значение и справку.

В соответствующие поля диалогового окна вводятся значения аргументов или адреса ячеек с этими значениями путем их выделения или записи с помощью операторов, затем нажимается клавиша *Enter* или *ОК*. В выбранной ячейке появится значение, полученное по запи-

санной в ней формуле.

Например, математическая функция КОРЕНЬ (число) возвращает значение квадратного корня из числа. В окно *Число* вводится число. Пример заполнения диалогового окна, вид строки формул и заполняемой ячейки приведены на рис. 3.

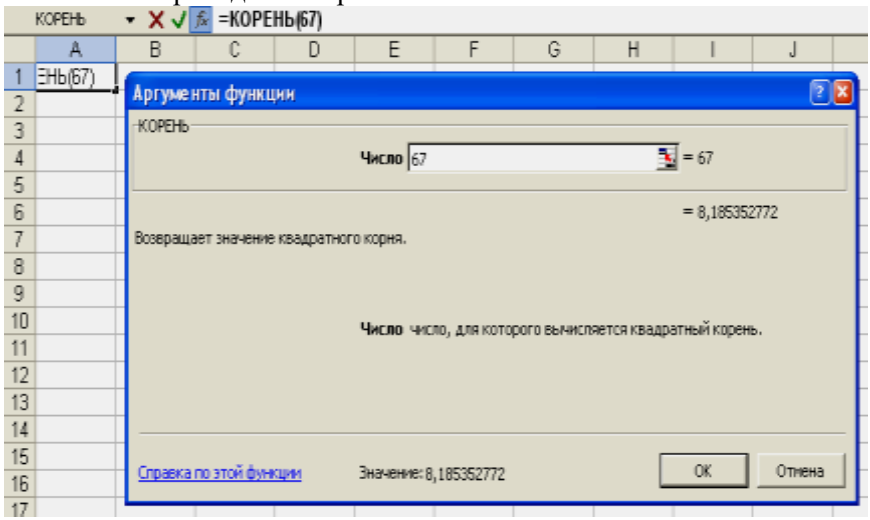


Рис. 3. **Диалоговое окно функции КОРЕНЬ, вид строки формул и заполняемой ячейки A1**

В строке формул отображаются имя функции и ее аргументы, указанные в определенной последовательности и имеющие соответствующий тип. Формула может содержать одну или более функций. Если выделить ячейку с вычисленным по формуле значением, то в строке формул появится запись этой формулы.

Для изменения формулы выполняются следующие действия:

- активизируется ячейка, содержащая формулу;
- в строке формул в необходимой части формулы устанавливается курсор щелчком левой кнопки мыши;
- производятся необходимые изменения;
- фиксируются изменения нажатием клавиши *Enter*.

Для выполнения вычислений формулами массива над массивом данных с получением массива такого же объема выполняются следующие действия:

• выделяются ячейки, в которые будет записан результат вычислений;

• нажимается клавиша *F2*;

• в строку формул вводится формула с использованием функции, аргументом которой является значение одной из ячеек исходного массива (например, применить формулу для ячеек C2:C34);

• нажимается сочетание клавиш *Ctrl+Shift+Enter* (при нажатии *OK* или *Enter* вычисляется значение только для первой ячейки выделенного массива; в этом случае снова нажимается *F2* и *Ctrl+Shift+Enter*).

В выделенных ячейках появится результат вычислений для всех ячеек исходного массива данных. Если выделить хотя бы одну из ячеек, то в строке формул появится запись формулы в фигурных скобках. Исправление формулы необходимо проводить для всех выделенных ячеек массива. В настоящем пособии формулы, применяемые для массива данных, заключаются в фигурные скобки.

Тема 2. ВЫБОРКА И ЕЕ АНАЛИЗ

План

1. Выборка.
2. Статистические ряды.
3. Эмпирическая функция распределения и кумулятивная кривая.
4. Лабораторная работа.

1. Выборка

Пусть случайная величина ξ описывает количественный или качественный признак некоторого физического или экономического процесса. В реальных условиях обычно бывает трудно или экономически нецелесообразно, а иногда и невозможно, исследовать всю совокупность значений случайной величины (*генеральную совокупность*). Поэтому на практике широко применяется выборочный метод, когда исследуется часть генеральной совокупности (*выборочная совокупность*, или *выборка*). При этом выборка считается *репрезентативной* (представительной), что в силу закона больших чисел достигается случайностью отбора. Основными типами выборок являются сле-

дующие:

1. *Собственно случайная:*

• *повторная* (элементы после выбора возвращаются обратно);

• *бесповторная* (выбранные элементы не возвращаются).

2. *Типическая*. Генеральная совокупность предварительно разбивается на группы типических элементов, и выборка осуществляется из каждой. Необходимо различать следующие виды выборок:

• *равномерную* (при равенстве объемов исходных групп в генеральной совокупности выбирается одинаковое количество элементов из каждой);

• *пропорциональную* (численность выборок формируют пропорционально численностям или средним квадратическим отклонениям групп генеральной совокупности);

• *комбинированную* (численность выборок пропорциональна и средним квадратическим отклонениям, и численностям групп генеральной совокупности).

3. *Механическая*. Отбор элементов проводится через определенный интервал.

4. *Серийная*. Отбор проводится не по одному элементу, а сериями для проведения сплошного обследования.

5. *Комбинированная*. Используются различные комбинации вышеуказанных методов (например, типическая выборка сочетается с механической и собственно случайной).

2. Статистические ряды

Пусть из генеральной совокупности значений случайной величины ξ извлекается выборка $X = \{x_1, x_2, \ldots, x_n\}$ объема n. *Вариационным рядом* называется таблица, в первой строке которой указываются *варианты* $x_{(1)} = x_{\min}$, $x_{(2)}$, …, $x_{(k)} = x_{\max}$ наблюдаемого признака ξ в порядке возрастания $x_{(1)} \le x_{(2)} \le \ldots \le x_{(k)}$, во второй строке – соответствующие им частоты n_i или относительные частоты $w_i\left(\xi = x_{(i)}\right) = \dfrac{n_i}{n}$ (табл. 1).

Здесь n_i – количество наблюдений значения x_i в выборке, $\sum_{i=1}^{k} n_i = n$.

Таблица 1. **Вариационный ряд**

Варианты ξ	$x_{(1)}$	$x_{(2)}$	…	$x_{(k)}$
Частота	$n_{(1)}$	$n_{(2)}$	…	$n_{(k)}$
Относительные частоты	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_k}{n}$

Полигоном относительных частот называется ломаная линия, состоящая из отрезков, соединяющих точки $\left(x_{(i)}; \frac{n_i}{n}\right)$, где $i = 1, 2, \ldots, k$. В полигоне частот соединяются точки $(x_{(i)}; n_i)$.

При большом объеме выборки ее элементы могут быть сгруппированы в *интервальный выборочный ряд* (табл. 2).

Таблица 2. **Интервальный вариационный ряд**

Варианты ξ	$[x_{(1)}; x_{(1)} + h)$	$[x_{(1)} + h; x_{(1)} + 2h)$	…	$[x_{(1)} + (m-1)h; x_{(k)})$
Частота	n_1^*	n_2^*	…	n_{m-1}^*
Относительные частоты	$\frac{n_1^*}{n}$	$\frac{n_2^*}{n}$	…	$\frac{n_{m-1}^*}{n}$

Для этого интервал изменения всех ее вариант разбивается на m непересекающихся полуинтервалов. Вычисления значительно упрощаются, если частичные интервалы имеют одинаковую длину $h = \frac{R}{m}$, где величина $R = x_{(k)} - x_{(1)}$ является *размахом* выборки. Количество интервалов – целая часть числа, определяемого по одной из формул: $m \approx \sqrt{n}$, $m \approx 5 \lg n$, $m \approx 1 + 3{,}322 \lg n$ (формула Стерджесса). Затем подсчитывается число вариант выборки, попавших в каждый из интервалов, вычисляются относительные частоты числа вариант в интервале. Интервальный статистический ряд представляет собой таблицу, в *первой строке* которой указываются полуинтервалы $[x_{(1)}; x_{(1)} + h)$, $[x_{(1)} + h; x_{(1)} + 2h)$, …, $[x_{(1)} + (m-1)h; x_{(k)})$, *во второй строке* – соответствующие им частоты n_i^* или относительные частоты $w_i = \frac{n_i^*}{n}$, $i = 1, 2, \ldots, m - 1$, где n_i^* – количество наблюдений в соответствую-

щем интервале, $\sum_{i=1}^{k} n_i^* = n$, $x_{(m)} = x_{(k)}$ – наибольшая варианта.

Если серединам $x_{(i)}^*$ каждого полуинтервала $[x_{(1)} + (i-1)h; ih)$, где $i = 1, 2, \ldots, m-1$, равной длины поставить в соответствие относительные частоты $w_i = \frac{n_i^*}{n}$, то получится *вариационный ряд с равноотстоящими вариантами* (табл. 3).

Таблица 3. **Вариационный ряд с равноотстоящими вариантами**

Варианты ξ	$x_{(1)}^*$	$x_{(2)}^*$	…	$x_{(m-1)}^*$
Относительные частоты	$\frac{n_1^*}{n}$	$\frac{n_2^*}{n}$	…	$\frac{n_{m-1}^*}{n}$

Аналогично, для полуинтервалов разной длины получается вариационный ряд с *неравноотстоящими вариантами*.

Гистограммой относительных частот называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат интервалы $[x_{(i)}; x_{(i+1)})$, где $i = 1, 2, \ldots, m-1$, длины h на оси Ox, а высоты равны отношению $\frac{n_i^*}{n \cdot h}$. Площадь каждого прямоугольника равна относительной частоте $\frac{n_i^*}{n}$. Сумма площадей всех прямоугольников равна единице. *Отсортированной гистограммой* называется ступенчатая фигура, состоящая из прямоугольников, построенных на интервалах $[x_{(i)}; x_{(i+1)})$, где $i = 1, 2, \ldots, m-1$, расположенных в порядке убывания относительных частот. *Кумулятивной кривой* (*кумулянтой*, или *кривой Лоренца*) называется ломаная линия, соединяющая точки $\left(x_{(i)}^*; \sum_{j=1}^{i} \frac{n_j^*}{n} \right)$, где $i = 1, 2, \ldots, m-1$.

Полигон относительных частот можно рассматривать в качестве статистического аналога плотности распределения, а кумулятивную кривую – в качестве статистического аналога функции распределения.

Пусть $p \in [0; 1]$. *Выборочной квантилью порядка p* называется абсцисса x_p точки, лежащей на кумулятивной кривой и имеющей ординату p (рис. 4).

Порядок квантили p определяет долю общего числа вариант в выборке, значения которых не превосходят x_p.

Процентным выражением кумулятивной кривой является интегральный процент, показывающий накопляемый процентный вклад каждого интервала значений величины ξ.

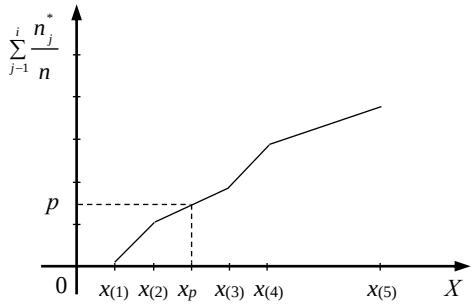


Рис. 4. **Выборочная квантиль порядка *p***

По виду полигона или гистограммы обычно выдвигают предположение о виде закона распределения исследуемой случайной величины, что позволяет придать определенную направленность исследованиям.

3. Эмпирическая функция распределения и кумулятивная кривая

Статистической (*эмпирической*) *функцией распределения* выборки называется функция $F^*(x)$ действительного аргумента x, определяющая относительную частоту события $\{\xi < x\}$: $F^*(x) = \frac{n_x}{n}$, где n_x – число вариант, меньших x.

Основными свойствами статистической функции распределения выборки являются следующие:

- $0 \le F^*(x) \le 1$;

• функция $F^*(x)$ неубывающая;

• $F^*(x) = 0$ при $x < x_1$ и $F^*(x) = 1$ при $x > x_k$, где x_1 и x_k – соответственно наименьшее и наибольшее значения выборки.

Если известен вариационный ряд выборки, то статистическая функция распределения определяется как $F^*(x) = \frac{n_x}{n} = \sum_{x_{(i)} < x} \frac{n_i}{n}$.

Для построения статистической функции распределения $F^*(x)$ в случае известного интервального ряда осуществляется переход к вариационному ряду с равноотстоящими или неравноотстоящими вариантами.

При работе в *Excel* с целью построения вариационного ряда и полигона частот используются соответственно инструменты *Сортировка по возрастанию*  и *Мастер диаграмм*.

Для нахождения интегрального ряда, кумулятивной кривой, гистограммы используется *Гистограмма* из пакета *Анализ данных* меню *Сервис*. Результат анализа появляется в виде таблицы. Формулы и соответствующие функции *Excel*, по которым выполняются расчеты в данном режиме, приводятся в табл. 4.

Таблица 4. **Интервальный ряд**

Карман (левые концы интервалов $[x_{(1)} + (i-1)h;\ x_{(1)} + ih)$)	*Частота*
$x_{(1)}$	n_1^* – число вариант от $x_{(1)}$ до $x_{(1)} + h$
$x_{(1)} + h$	n_2^* – число вариант от до $x_{(1)} + h$ до $x_{(1)} + 2h$
…	…
$x_{(1)} + (m-2)h$	n_{m-2}^* – число вариант от $x_{(1)} + (m-2)h$ до $x_{(1)} + (m-1)h$
Еще	n_{m-1}^* – число вариант от $x_{(1)} + (m-1)h$ до $x_{(k)}$

В табл. 5 указаны формулы, по которым проводятся расчеты в *Excel* для выполнения статистического анализа:

• *Парето* (графы 1, 2, 4, 5);

• *Интегральный процент* (графы 1–3);

• *Парето* и *Вывод графика* (графы 1, 2, 4, 5), при этом выводится гистограмма частот, расположенных в порядке убывания;

• *Парето*, *Интегральный процент* и *Вывод графика*, при этом появляются графы 1–6, гистограмма частот, расположенных в порядке убывания, и график кумулятивной кривой, соответствующей интегральному проценту.

Таблица 5. **Формулы для расчета результатов статистических анализов *Парето, Интегральный процент, Вывод графика***

Карман	*Частота*	*Интегральный %*	*Карман*	*Частота*	*Интегральный %*
1	2	3	4	5	6
$x_{(1)}$	n_1^* – число вариант от $x_{(1)}$ до $x_{(1)} + h$	$\frac{n_1^*}{n} \cdot 100\%$		l_1 – наибольшая частота	$\frac{l_1}{n} \cdot 100\%$

Окончание табл. 5

1	2	3	4	5	6
$x_{(1)} + h$	n_2^* – число вариант от $x_1 + h$ до $x_{(1)} + 2h$	$\frac{n_1^* + n_2^*}{n} \cdot 100\%$		l_2	$\frac{l_1 + l_2}{n} \cdot 100\%$
…	…	…		…	…
$x_{(1)} + (m - 2)h$	n_{m-2}^* – число вариант от $x_1 + (m-2)h$ до $x_1 + (m-1)h$	$\frac{n_1^* + \ldots + n_{m-2}^*}{n} \times$ $\times 100\%$		l_{m-2}	$\frac{l_1 + l_2 + \ldots + l_{m-2}}{n} \times$ $\times 100\%$
Еще	n_{m-1}^* – число вариант от $x_1 + (m-1)h$ до x_k	$\frac{n_1^* + \ldots + n_{m-1}^*}{n} \times$ $\times 100\% = 100\%$		l_{m-1} – наименьшая частота	$\frac{l_1 + l_2 + \ldots + l_{m-1}}{n} \times$ $\times 100\% = 100\%$

Примечание. В табл. 5 графы 1–6 обозначают следующее:

1 – левый конец каждого интервала;

2 – число вариант, содержащихся в соответствующем интервале;

3 – процентное выражение накопленных частот для интервалов;

4 – правые концы интервалов, расположенные в порядке убывания частот;

5 – частоты, расположенные в порядке убывания;

6 – процентное выражение накопленных частот, расположенных в порядке убывания.

Вопросы для самоконтроля

1. Что называется вариационным рядом?
2. Как строится интервальный ряд?
3. Какой ряд называется интервальным вариационным рядом с равноотстоящими вариантами?
4. Как строится гистограмма?
5. Что называется кумулятивной кривой?
6. Что понимают под статистической функцией распределения?
7. Какими свойствами обладает статистическая функция распределения?

4. Лабораторная работа

Задание

В результате наблюдений получена выборка X из генеральной совокупности случайной величины ξ, характеризующей расход сырья при производстве продукции по данной технологии:

114	112	132	124	119	124	119	116	129	116
124	119	119	114	129	116	124	129	116	119
110	124	140	119	124	129	119	124	124	124
116	129	119	124	110	124	112	114	129	116
119	116	129	116	119	114	132	119	124	112

Требуется:

1) построить вариационный ряд и полигон относительных частот;
2) построить интервальный ряд, гистограмму, кумулятивную кривую.

Порядок выполнения работы

Работа выполняется в следующем порядке:

1. Вариационный ряд, полигон частот. В ячейку A1 вводится название «Выборка X», а в ячейки A2:A51 – значения выборки X. Для построения вариационного ряда выполняются следующие действия:

• Выделяются ячейки A2:A51, копируются, сохраняются в ячейках B2:B51. Активизируется на панели инструментов кнопка *Сортировка*

по возрастанию  (отсортированная выборка располагается в тех же ячейках столбца B). В ячейку B1 вводится название «Выборка».

• В ячейки столбца C копируются неповторяющиеся варианты выборки (C2:C10), в соответствующих ячейках D2:D10 столбца D вводятся частоты вариант. В ячейку C1 вводится название «Частоты».

• В ячейки E2:E10 вводится формула массива (см. тему 1) {=D2:D10/50}. В ячейку D1 вводится название «Относительные частоты вариант».

Вариационный ряд представляет собой таблицу, расположенную в столбцах C, D, E рис. 5.

Для построения полигона частот выполняются следующие действия:

• На панели инструментов активизируется кнопка *Мастер диаграмм* (*шаг 1 из 4*), в одноименном диалоговом окне среди стандартных типов выбирается *График* и верхний левый вид диаграммы, нажимается кнопка *Далее>*.

• Открывается диалоговое окно *Мастер диаграмм* (*шаг 2 из 4*), в котором во вкладке *Диапазон данных* в поле *Диапазон* вводится ссылка на диапазон частот D2:D10. Во вкладке *Ряд* в поле *Подписи оси X* вводится ссылка на массив вариант C2:C10, в поле *Имя* – название «Полигон частот». Затем нажимается кнопка *Далее>*.

• Открывается диалоговое окно *Мастер диаграмм* (*шаг 3 из 4*), в котором во вкладке *Заголовки* в поле *Ось X*(*категорий*) вводится название «Варианты», в поле *Ось Y*(*значений*) – название «Частоты». Во вкладке *Легенда* снимается флажок *Добавить легенду* и нажимается кнопка *Далее>*.

• Открывается диалоговое окно *Мастер диаграмм* (*шаг 4 из 4*), в поле *имеющемся* устанавливается флажок.

Построенный полигон частот изображен на рис. 5.

Вывод. По виду полигона частот можно предположить, что случайная величина ξ расхода сырья при производстве продукции по данной технологии имеет нормальное распределение.

Замечание. На графике полигона частот по оси вариант не выдержан масштаб, что не мешает правильно оценить характер распределения исследуемого признака.

2. Интервальный ряд, гистограмма, кумулятивная кривая. Для построения интервального ряда выборки *X* используется анализ *Гистограмма* из *Анализа данных* меню *Сервис*. Значения параметров в од-

ноименном диалоговом окне устанавливаются следующим образом (рис. 6):

- *Входной интервал* (вводятся ссылки на ячейки A1:A51, в которых находятся значения выборки *X*);
- *Интервал карманов* (не заполняется);
- *Метки* (устанавливается флажок);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист Гистограмма*).

	A	B	C	D	E	F	G	H
1	Выборка X	Упорядоченная выборка	Варианты $x_{(i)}$	Частота n_i	Относиттельная частота n_i / n			
2	110	110	110	2	0,04			
3	110	110	112	3	0,06			
4	112	112	114	4	0,08			
5	112	112	116	8	0,16			
6	112	112	119	11	0,22			
7	114	114	124	12	0,24			
8	114	114	129	7	0,14			
9	114	114	132	2	0,04			
10	114	114	140	1	0,02			
11	116	116						
12	116	116						
13	116	116						
14	116	116						
15	116	116						
16	116	116						
17	116	116						
18	116	116						
19	119	119						
20	119	119						
21	119	119						
22	119	119						
23	119	119						
24	119	119						
25	119	119						
26	119	119						
27	119	119						
28	119	119						
29	119	119						
30	124	124						
31	124	124						
32	124	124						
33	124	124						
34	124	124						
35	124	124						
36	124	124						
37	124	124						
38	124	124						
39	124	124						
40	124	124						
41	124	124						
42	129	129						
43	129	129						
44	129	129						
45	129	129						
46	129	129						
47	129	129						
48	129	129						
49	132	132						
50	132	132						
51	140	140						

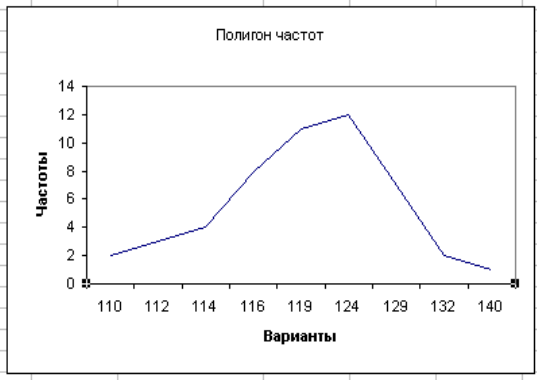


Рис. 5. **Вариационный ряд и полигон частот**

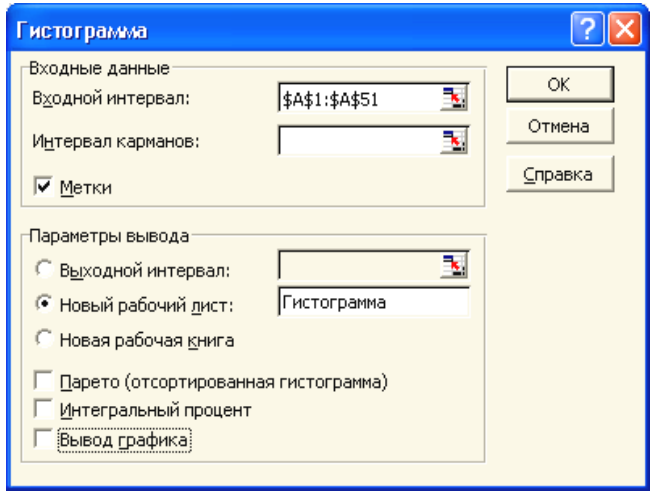


Рис. 6. **Заполнение диалогового окна *Гистограмма***

Результаты вычислений появляются в виде таблицы (рис. 7), которая является интервальным рядом.

C	D
Карман	*Частота*
110	2
114,2857	7
118,5714	8
122,8571	11
127,1429	12
131,4286	7
135,7143	2
Еще	1

Рис. 7. **Интервальный ряд**

В режиме *Гистограмма* выполняются статистические анализы *Парето, Интегральный процент, Вывод графика.*

Если устанавливается флажок *Парето* (отсортированная гистограмма), то в выходном диапазоне появятся гистограмма и отсортированная гистограмма (рис. 8).

Если устанавливается флажок *Интегральный процент,* то в выходном диапазоне появляется накопляемый процентный вклад каждого интервала значений вариант, представляющий собой процентное

выражение эмпирической функции распределения (рис. 9).

Карман	*Частота*	*Карман*	*Частота*
110	2	127,1429	12
114,2857	7	122,8571	11
118,5714	8	118,5714	8
122,8571	11	114,2857	7
127,1429	12	131,4286	7
131,4286	7	110	2
135,7143	2	135,7143	2
Еще	1	Еще	1

Рис. 8. **Результат анализа *Парето***

Карман	*Частота*	*Интегральный %*
110	2	4,00%
114,2857	7	18,00%
118,5714	8	34,00%
122,8571	11	56,00%
127,1429	12	80,00%
131,4286	7	94,00%
135,7143	2	98,00%
Еще	1	100,00%

Рис. 9. **Результат анализа *Интегральный процент***

Если устанавливается флажок *Вывод графика*, то выводятся интервальный ряд и гистограмма (рис. 10).

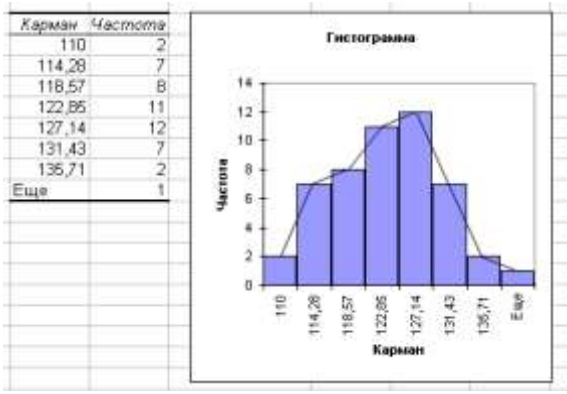


Рис. 10. **Результат анализа *Вывод графика***

Замечания:

1. В ячейках *Карман* можно изменить формат чисел, оставив два знака после запятой (выделяются ячейки, затем в командной строке выбирается *Формат – Ячейки… – Число – Числовой – Число десятичных знаков* 2).

2. Для того чтобы гистограмма приняла стандартный вид (столбики гистограммы изображались в виде смежных прямоугольных областей), необходимо выбрать *Область диаграммы – Ряд Частота –*

Формат рядов данных – Параметры – Ширина зазора 0, командами *Область построения диаграммы – Вид – Заливка – Прозрачная* изменяется фон области построения диаграммы.

3. При необходимости гистограмма интервального ряда может быть преобразована в полигон. Для этого нужно соединить середины верхних сторон прямоугольников прямыми линиями (ломаная линия на рис. 10 получена с помощью команды *Рисование – Линия* на панели инструментов *Рисование*).

Если установить флажки *Парето*, *Интегральный процент*, *Вывод графика*, результаты этих анализов получаются в виде таблицы, гистограммы и кумулятивной кривой (рис. 11).

Карман	Частота	Интегральный %	Карман	Частота	Интегральный %
110,00	2	4,00%	127,14	12	24,00%
114,29	7	18,00%	122,86	11	46,00%
118,57	8	34,00%	118,57	8	62,00%
122,86	11	56,00%	114,29	7	76,00%
127,14	12	80,00%	131,43	7	90,00%
131,43	7	94,00%	110,00	2	94,00%
135,71	2	98,00%	135,71	2	98,00%
Еще	1	100,00%	Еще	1	100,00%

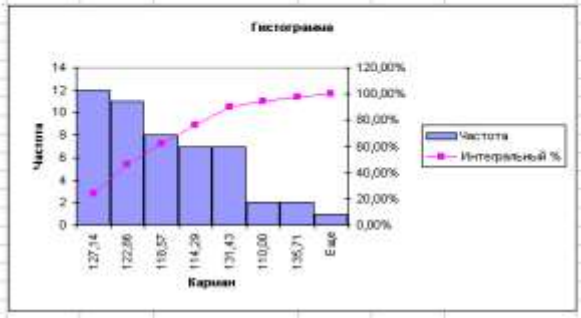


Рис. 11. **Результат анализов *Парето, Интегральный процент, Вывод графика***

Вывод. По виду гистограммы, изображенной на рис. 10, можно предположить, что случайная величина ξ расхода сырья при производстве продукции по данной технологии имеет асимптотически нормальное распределение.

Тема 3. ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

План

1. Числовые характеристики выборки.
2. Точечные оценки параметров распределения.
3. Интервальные оценки параметров распределения.
4. Лабораторная работа.

1. Числовые характеристики выборки

Пусть случайный эксперимент описывается случайной величиной ξ, распределение которой зависит от одного или нескольких *параметров*. К ним, в частности, относятся среднее, мода, медиана, среднее квадратичное отклонение, дисперсия, коэффициенты эксцесса и асимметрии, размах вариации, называемые *параметрами генеральной совокупности*. При исследовании случайной величины ξ из генеральной совокупности ее возможных значений извлекается выборка $X = \{x_1, x_2, \ldots, x_n\}$ объема n. По данной выборке можно приближенно вычислить значения каждого из изучаемых параметров, которые в статистике называются *числовыми оценками параметров*, или просто *оценками*.

Данные характеристики условно разбиваются на следующие четыре группы:

- показатели положения вариант на числовой оси;
- показатели разброса вариант относительно своего центра, определяющие кучность данных около центра;
- показатели асимметрии распределения вариант около своего центра;
- показатели, описывающие закон распределения.

Пусть выборка $X = \{x_1, x_2, \ldots, x_n\}$ задана вариационным рядом (табл. 6).

Таблица 6. **Вариационный ряд**

Варианты	x_1	x_2	…	x_k
Относительные частоты	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_k}{n}$

Выборочным средним значений выборки называется число, определяемое по формуле $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$, где x_i – варианта с частотой n_i, n – объем выборки, $\sum_{i=1}^{k} n_i = n$. Если частоты n_i равны единице, то $k = n$.

При достаточно больших n используют формулу $\bar{x} = \sum_{i=1}^{k} x_i w_i$, где k – число различных значений вариационного ряда, w_i – относительная частота варианты x_i. Если известен интервальный вариационный ряд, то для вычисления выборочного среднего используется формула $\bar{x} = \sum_{i=1}^{m} \tilde{x}_i w_i$, где m – число интервалов, $\tilde{x}_i = \frac{x_i + x_{i-1}}{2}$ – середина i-го интервала, w_i – относительная частота.

Выборочной медианой (*Me*) называется значение признака, находящегося в середине вариационного ряда. Если число вариант нечетно, т. е. $n = 2m + 1$, то медианой является $(m + 1)$-я варианта ($Me = x_{m+1}$); если же число вариант четно, то медиана равна среднему арифметическому двух значений в середине ряда $Me = \frac{x_m + x_{m+1}}{2}$. Если известен интервальный вариационный ряд, то медианой является выборочная квантиль $x_{0,5}$ (см. тему 2). Оценка медианы по интервальному ряду с одинаковой длиной интервалов h может быть вычислена по формуле $Me = x_{Me} + h \cdot \frac{\frac{n}{2} - \left(n_1^* + n_2^* + \ldots + n_{Me-1}^*\right)}{n_{Me}^*}$, где x_{Me} – нижняя граница интервала, к которому принадлежит медиана, n_{Me}^* – число элементов выборки в этом интервале, $\left(n_1^* + n_2^* + \ldots + n_{Me-1}^*\right)$ – число элементов выборки в интервалах, лежащих слева от интервала, содержащего медиану.

Выборочной модой (*Mo*) называется варианта выборки, имеющая наибольшую частоту. Если несколько соседних значений имеют наибольшую частоту, то модой является их среднее арифметическое

$Mo = \dfrac{x_i + x_{i+1}}{2}$, где для вариант x_i, x_{i+1} частоты $n_i = n_{i+1} = n_{\max}$. Если две или более несмежных вариант имеют разные наибольшие частоты, то ряд называется полимодальным. Если же все варианты встречаются одинаково часто, то ряд моды не имеет.

Если длина интервалов интервального ряда одинакова и равна h, то мода вычисляется по формуле $Mo = x_{Mo} + h \cdot \dfrac{n^*_{Mo} - n^*_{Mo-1}}{2n^*_{Mo} - n^*_{Mo-1} - n^*_{Mo+1}}$, где x_{Mo} – нижняя граница интервала, содержащего наибольшее число элементов выборки в этом интервале, n^*_{Mo-1}, n^*_{Mo+1} – число элементов выборки в соседних интервалах.

Мода и медиана используются в качестве характеристики среднего значения в случае, если границы ряда нечеткие или ряд не симметричен.

Выборочная дисперсия и выборочное среднее квадратическое отклонение используются для описания рассеивания значений случайной величины относительно выборочного среднего.

Выборочной дисперсией значений выборки $X = \{x_1, x_2, \ldots, x_n\}$ называется число, определяемое по формуле

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

или по формуле, полученной после преобразований последней:

$$\sigma_в^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}\right)^2 .$$

Если задан вариационный ряд, то используется формула $\sigma_в^2 = \sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 w_i$, где k – количество вариант, w_i – относительная частота варианты x_i. Если задан интервальный вариационный ряд, то $\sigma_в^2 = \sum_{i=1}^{m}\left(\widetilde{x}_i - \bar{x}\right)^2 w_i$, где m – количество интервалов, $\widetilde{x}_i$ – середина i-го интервала, w_i – соответствующая относительная частота.

Выборочным средним квадратическим отклонением называется

число, которое находится по формуле $\sigma_в = \sqrt{\sigma_в^2}$.

Коэффициент вариации $\upsilon = \frac{\sigma_в}{\bar{x}} \cdot 100\,\%$ – безразмерная величина, которая служит для сравнения величин рассеивания по отношению к выборочной средней вариационных рядов, варианты которых имеют различную размерность. Он характеризует относительное значение среднего квадратического отклонения.

Обобщающими характеристиками выборочных распределений являются моменты вариационного ряда.

Начальным выборочным моментом m-го порядка (m = 0, 1, 2, …) называется величина $\nu_m^* = \frac{1}{n}\sum_{i=1}^{k} x_i^m n_i$, где x_i – варианта с частотой n_k, n – объем выборки, $\sum_{i=1}^{k} n_i = n$. Очевидно, что $\nu_0^* = 1$, $\nu_1^* = \bar{x}$, $\nu_2^* = \overline{x^2}$ и т. д.

Центральным выборочным моментом m-го порядка (m = 0, 1, 2, …) называется величина $\mu_m^* = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^m n_i$, где x_i – варианта с частотой n_i, n – объем выборки, $\sum_{i=1}^{k} n_i = n$, $\bar{x}$ – выборочное среднее. Очевидно, что $\mu_0 = 1$, $\mu_1^* = 0$, $\mu_2^* = \sigma_в^2$, $\mu_3^* = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^3 n_i$, $\mu_4^* = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^4 n_i$.

Важную роль при исследовании статистических совокупностей играют асимметрия и эксцесс распределения признака, которые вычисляются соответственно по следующим формулам: $A_s = \frac{\mu_3^*}{\sigma_в^3}$, $E = \frac{\mu_4^*}{\sigma_в^4} - 3$.

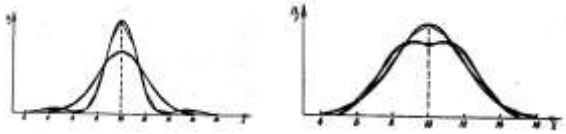
Если кривая распределения симметрична относительно прямой $x = \bar{x}$, то распределение симметрично. Тогда $A_s = 0$ ($\mu_3^* = 0$). При асимметричном распределении вершина кривой сдвинута относительно

ординаты выборочной средней. Если $A_s > 0$, то асимметрия является правосторонней (положительной), если $A_s < 0$, то – левосторонней (отрицательной) (рис. 12).

а б

Рис. 12. **Правосторонняя (а) и левосторонняя (б) асимметрии распределения**

Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение, отрицательный эксцесс – относительно сглаженное распределение (рис. 13).



а б

Рис. 13. **Положительный (а) и отрицательный (б) эксцессы распределения**

При работе в *Excel* с целью вычисления числовых характеристик выборки используется режим *Описательная статистика* из *Анализа данных* меню *Сервис*. Результат анализа представлен в виде таблицы. Формулы и функции *Excel*, по которым рассчитываются соответствующие числовые характеристики, приводятся в табл. 7.

Таблица 7. **Описательная статистика**

	Название выборки
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ СРЗНАЧ
Стандартная ошибка[1]	$\varepsilon = \frac{S}{\sqrt{n}}$

Медиана	$Me = \begin{cases} x_m & \text{при} \quad n = 2m-1, \\ \frac{1}{2}\left(x_m + x_{m+1}\right) & \text{при} \quad n = 2m \end{cases}$ МЕДИАНА
Мода	$Mo = x_i$, если $n_i = n_{\max}$ Если выборка не содержит одинаковых вариант, то возвращается значение ошибки #Н/Д МОДА
Стандартное отклонение[2]	$S = \sqrt{S^2}$ СТАНДОТК
Дисперсия выборки	Несмещенная выборочная дисперсия[2] $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ ДИСП

Окончание табл. 7

Название выборки	
Эксцесс	$E = \frac{1}{S^4 \cdot n}\cdot\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 - 3$ ЭКСЦЕСС
Асимметричность	$A_s = \frac{1}{S^3 \cdot n}\cdot\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3$ СКОС
Интервал	Размах вариаций $R = x_{\max} - x_{\min}$
Минимум	Наименьшее значение выборки $x_{\min}$ МИН
Максимум	Наибольшее значение выборки $x_{\max}$ МАКС
Сумма	$\sum_{i=1}^{n} x_i$ СУММ
Счет	Объем выборки n СЧЕТ
Наибольший (k)	Наибольшее значение во множестве данных, начиная с $x_{\max}$ НАИБОЛЬШИЙ
Наименьший (k)	Наименьшее значение во множестве данных, начиная с $x_{\min}$ НАИМЕНЬШИЙ

Уровень надежности[3] ((1 – α) · 100%)	Предельная ошибка средней выборочной $\Delta_{\bar{x}} = t\frac{\sigma}{\sqrt{n}}$ ДОВЕРИТ

Примечание. Параметры 1 и 2 определены в пункте 2, а 3 – в пункте 3 данной темы.

2. Точечные оценки параметров распределения

Пусть случайный эксперимент описывается случайной величиной ξ. Повторяя случайный эксперимент n раз, получаем последовательность значений $X = \{x_1, x_2, \ldots, x_n\}$ случайной величины ξ. Множество X называется выборкой объема n из генеральной совокупности случайной величины ξ.

Статистической оценкой $\tilde{\theta}(x_1, x_2, \ldots, x_n)$ параметра θ называется приближенное значение параметра, полученное на основе статистических (выборочных) данных. Статистическая оценка является выборочной характеристикой, вычисляемой по результатам n наблюдений величины ξ, используется в качестве оценки θ-характеристики генеральной совокупности. В качестве θ могут выступать параметры $M(\xi)$, $D(\xi)$, параметр распределения и т. д.

По статистическим данным нельзя получить точную оценку неизвестного параметра θ, но можно найти приближенную оценку. Более того, каждая выборка объема n из генеральной совокупности дает свою оценку одного и того же неизвестного параметра θ, т. е. для θ можно получить бесконечное множество его оценок. Поэтому оценку $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$ можно считать случайной величиной, а ее значение $\tilde{\theta}$, вычисленное по одной данной выборке, можно рассматривать как одно из ее возможных значений.

Различают точечные и интервальные оценки.

Точечная оценка параметра θ определяется одним числом $\tilde{\theta}$. Качество оценки $\tilde{\theta}$ устанавливается по следующим трем свойствам:

- *Несмещенность.* Оценка $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$ неизвестного параметра θ генеральной совокупности называется *несмещенной*, если

для фиксированного числа наблюдений n выполняется равенство $M(\tilde{\theta}) = \theta$.

• *Состоятельность.* Оценка $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$, найденная по выборке объема n, называется *состоятельной*, если для любого $\varepsilon > 0$ выполняется равенство $\lim\limits_{n\to\infty} P\left(\left|\tilde{\theta} - \theta\right| < \varepsilon\right) = 1$, которое означает, что при увеличении объема n выборки значение $\tilde{\theta}$ сходится по вероятности к неизвестному параметру θ.

• *Эффективность.* Несмещенная оценка $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$ неизвестного параметра θ называется *эффективной*, если среди всех подобных оценок того же параметра она имеет наименьшую дисперсию $D\left(\tilde{\theta}\right) \to \min$.

Оценкой математического ожидания случайной величины ξ является ее выборочная средняя $\bar{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$. Она является несмещенной, состоятельной и эффективной.

Выборочная дисперсия $\sigma_в^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$ является *смещенной* оценкой для дисперсии случайной величины ξ.

Величина $S^2 = \dfrac{n}{n-1}\sigma_в^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$ является несмещенной, эффективной и состоятельной оценкой дисперсии случайной величины ξ и называется *несмещенной выборочной дисперсией*, а смещенная дисперсия – *генеральной дисперсией*, или *дисперсией генеральной совокупности*.

Величина $S = \sqrt{S^2}$ называется *стандартным отклонением*.

Стандартная ошибка выборки определяется по формуле $\varepsilon = \dfrac{S}{\sqrt{n}}$.

Наиболее распространенными методами получения точечных оценок являются метод моментов и метод максимального правдоподобия.

Пусть распределение генеральной совокупности зависит от m параметров. Согласно *методу моментов* m теоретических моментов приравниваются к соответствующим выборочным моментам. При этом по-

лучается система из m уравнений, решением которой являются оценки каждого из m параметров. Достоинством данного метода является его простая вычислительная реализация, недостаток состоит в том, что получаемые оценки являются смещенными и малоэффективными.

Метод максимального правдоподобия заключается в том, что для получения оценки неизвестного параметра θ нужно найти такое значение $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$, при котором вероятность реализации выборки $x_1, x_2, \ldots, x_n$ была бы максимальной. С этой целью строится функция правдоподобия $L(x_1, x_2, \ldots, x_n, \theta)$, определяющая вероятность получения выборки $x_1, x_2, \ldots, x_n$, и находится точка максимума этой функции, которая является оценкой неизвестного параметра θ. Функция правдоподобия *непрерывной* случайной величины ξ с плотностью вероятности $p(x, \theta)$ имеет следующий вид:

$$L(x_1, x_2, \ldots, x_n, \theta) = p(x_1, \theta) \cdot p(x_2, \theta) \ldots p(x_n, \theta).$$

Функция правдоподобия *дискретной* случайной величины ξ, для которой распределение вероятности $P(\xi = x, \theta)$ зависит от параметра θ, имеет следующий вид:

$$L(x_1, x_2, \ldots, x_n, \theta) = P(\xi = x_1, \theta) \cdot P(\xi = x_2, \theta) \ldots P(\xi = x_n, \theta).$$

Если оцениваемых параметров несколько $\theta_1, \theta_2, \ldots, \theta_m$, то строится и исследуется на максимум функция правдоподобия вида $L(x_1, x_2, \ldots, x_n, \theta_1, \theta_2 \ldots, \theta_m)$.

Поскольку функции $L(x_1, x_2, \ldots, x_n, \theta)$ и $\ln L(x_1, x_2, \ldots, x_n, \theta)$ достигают экстремумы при одних и тех же значениях $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$, то для упрощения расчетов иногда пользуются логарифмической функцией правдоподобия.

Метод наибольшего правдоподобия дает состоятельные оценки. Недостаток метода заключается в том, что иногда оценки наибольшего правдоподобия являются смещенными.

3. Интервальные оценки параметров распределения

Более полный и надежный способ оценивания параметров распределений заключается в определении не единственного точечного зна-

чения, а интервала, который с заданной вероятностью накрывает истинное значение оцениваемого параметра. Для заранее выбранного уровня значимости α, $0 < \alpha < 1$, по выборке определяются два числа $\tilde{\theta}_1$ и $\tilde{\theta}_2$, $\tilde{\theta}_1 < \tilde{\theta}_2$, между которыми с вероятностью $1 - \alpha$ находится неизвестный параметр θ: $P(\tilde{\theta}_1 < \theta < \tilde{\theta}_2) = 1 - \alpha$.

Число $p = 1 - \alpha$ называется *доверительной вероятностью* (*надежностью*), $\tilde{\theta}_1$, $\tilde{\theta}_2$ – *доверительными* нижней и верхней *границами*. Величины $\tilde{\theta}_1, \tilde{\theta}_2$ определяются по результатам выборки и, следовательно, являются случайными.

Если $\tilde{\theta}$ является точечной оценкой неизвестного параметра θ, то $\tilde{\theta}_1 = \tilde{\theta} - \Delta$, $\tilde{\theta}_2 = \tilde{\theta} + \Delta$, где $\Delta(\Delta > 0)$ – *предельная ошибка* (*уровень надежности*) выборки, которая либо задается заранее, либо вычисляется.

На практике часто используются односторонние доверительные интервалы, которые определяются из условий $P(\theta > \tilde{\theta}_1) = 1 - \alpha$ или $P(\theta < \tilde{\theta}_2) = 1 - \alpha$ и называются *правосторонними* и *левосторонними* соответственно.

Длина доверительного интервала, характеризующая точность интервальной оценки, зависит от объема выборки n и надежности $p = 1 - \alpha$. При увеличении n длина доверительного интервала уменьшается, а с приближением надежности к 1 – увеличивается. В качестве $p = 1 - \alpha$ принимают значения 0,9; 0,95; 0,99, которые определяют 90-, 95-, 99%-ные доверительные интервалы соответственно.

Задача определения доверительного интервала может быть решена только тогда, когда удается найти закон распределения случайной величины, используемой в качестве оценки, т. е. плотность вероятности $p(\tilde{\theta})$. В общем случае этот закон зависит от самого неизвестного параметра. Однако иногда удается перейти от оценки $\tilde{\theta}$ к таким функциям выборочных значений, закон распределения которых зависит только от объема выборки n и закона распределения случайной величины ξ и не зависит от неизвестных параметров.

Пусть выборка $X = \{x_1, x_2, \ldots, x_n\}$ произведена из генеральной совокупности значений нормально распределенной случайной величи-

ны ξ с параметрами $M(\xi) = a$ и $D(\xi) = \sigma^2$, т. е. $\xi \sim N(a; \sigma^2)$.

Доверительный интервал для математического ожидания $M(\xi)$ *при известной дисперсии* $D(\xi) = \sigma^2$ с доверительной вероятностью $p = 1 - \alpha$ имеет следующий вид:

$$\left(\bar{x} - t\frac{\sigma}{\sqrt{n}}, \bar{x} + t\frac{\sigma}{\sqrt{n}} \right).$$

Здесь $t = t\left(\frac{1-\alpha}{2}\right)$ – квантиль порядка $\frac{1-\alpha}{2}$ нормального распределения, $\Delta_{\bar{x}} = t\frac{\sigma}{\sqrt{n}}$ – точность оценки. Из соотношения $\Delta_{\bar{x}} = t\frac{\sigma}{\sqrt{n}}$ находится минимальный объем n выборки, который обеспечивает заданную точность $\Delta_{\bar{x}}$: $n = \left(\frac{t \cdot \sigma}{\Delta_{\bar{x}}}\right)^2$.

Число σ, как правило, неизвестно, поэтому его заменяют приближенным значением $\sigma \approx S$.

При работе в *Excel* с целью построения доверительного интервала математического ожидания при известной дисперсии используется статистическая функция ДОВЕРИТ(альфа;станд_откл;размер), которая рассчитывает предельную точность оценки $\Delta_{\bar{x}} = t\frac{\sigma}{\sqrt{n}}$. Параметры этой функции могут быть взяты из результата анализа *Описательная статистика*.

Доверительный интервал для математического ожидания $M(\xi)$ *при неизвестной дисперсии* $D(\xi) = \sigma^2$ с доверительной вероятностью $p = 1 - \alpha$ имеет следующий вид:

$$\left(\bar{x} - t \cdot \frac{S}{\sqrt{n}}, \ \bar{x} + t \cdot \frac{S}{\sqrt{n}} \right),$$

где t – квантиль распределения Стьюдента порядка α с $\nu = n - 1$ степенями свободы при $n \le 30$, а при $n > 30$ – это квантиль порядка $\frac{1-\alpha}{2}$ нормального стандартного распределения;

S – стандартное отклонение;

n – объем выборки. Величина $\Delta_{\bar{x}} = t \cdot \dfrac{S}{\sqrt{n}}$ называется предельной ошибкой выборочной средней.

При работе в *Excel* с целью построения доверительного интервала математического ожидания при неизвестной дисперсии используется статистическая функция СТЬЮДРАСПОБР(вероятность;степени_свободы), которая рассчитывает значение $t(\alpha;\ n-1)$, или функция НОРМРАСПОБР(вероятность), которая рассчитывает значение $t\left(\dfrac{1-\alpha}{2}\right)$.

Доверительный интервал для дисперсии $D(\xi) = \sigma^2$ *при известном математическом ожидании* $M(\xi)$ имеет вид

$$\left(\frac{n\sigma_в^2}{\chi^2\left(\frac{\alpha}{2}; n\right)},\ \frac{n\sigma_в^2}{\chi^2\left(1-\frac{\alpha}{2}; n\right)} \right),$$

где $\chi^2\left(\dfrac{\alpha}{2}; n\right)$, $\chi^2\left(1-\dfrac{\alpha}{2}; n\right)$ – квантили χ^2 распределения порядков $\dfrac{\alpha}{2}$, $1-\dfrac{\alpha}{2}$ с n степенями свободы.

Доверительный интервал для дисперсии $D(\xi) = \sigma^2$ с неизвестным математическим ожиданием $M(\xi)$ имеет вид

$$\left(\frac{(n-1)S^2}{\chi^2\left(\frac{\alpha}{2}; n-1\right)},\ \frac{(n-1)S^2}{\chi^2\left(1-\frac{\alpha}{2}; n-1\right)} \right),$$

где $\chi^2\left(\dfrac{\alpha}{2}; n-1\right)$, $\chi^2\left(1-\dfrac{\alpha}{2}; n-1\right)$ – квантили χ^2 распределения поряд-

ков $\frac{\alpha}{2}$, $1-\frac{\alpha}{2}$ с $n-1$ степенями свободы.

При работе в *Excel* с целью построения доверительных интервалов для дисперсии используется статистическая функция

ХИ2ОБР(вероятность;степени_свободы).

Вопросы для самоконтроля

1. Какие числовые характеристики выборки относятся к показателям положения?
2. Какие числовые характеристики выборки относятся к показателям разброса?
3. Какой показатель характеризует симметрию распределения?
4. Что называется выборочным средним, модой, медианой, выборочной дисперсией?
5. Каким условиям должны удовлетворять точечные оценки?
6. В чем суть метода моментов?
7. В чем суть метода максимального правдоподобия?
8. Какая функция называется функцией правдоподобия?
9. Какие оценки называются интервальными?
10. Что называется доверительной вероятностью?

4. Лабораторная работа

Задание

В результате наблюдений получена выборка X из генеральной совокупности случайной величины ξ, характеризующей расход сырья при производстве продукции по данной технологии:

114	112	132	124	119	124	119	116	129	116
124	119	119	114	129	116	124	129	116	119
110	124	140	119	124	129	119	124	124	124
116	129	119	124	110	124	112	114	129	116
119	116	129	116	119	114	132	119	124	112

Требуется:

1) определить числовые характеристики выборки;

2) предполагая, что случайная величина ξ имеет нормальное распределение $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, методами моментов и максимального правдоподобия оценить неизвестные параметры a и σ;

3) найти доверительные интервалы для математического ожидания и дисперсии при уровне надежности 95%, считая, что ξ имеет нормальное распределение.

Порядок выполнения работы

В ячейку A1 вводится название «Расход сырья», в ячейки A2:A51 – значения вариант выборки *X*. Название листа – *Исходные данные*. Работа выполняется в следующем порядке:

1. Числовые характеристики выборки. Для расчета используется анализ *Описательная статистика* из *Анализа данных* меню *Сервис*. Значения параметров в одноименном диалоговом окне устанавливаются следующим образом:

- *Входной интервал* – вводятся ссылки на ячейки A1:A51;
- *Группирование* (*по столбцам, по строкам*) – по столбцам;
- *Метки в первой строке* (*метки в первом столбце*) – проставляется флажок;
- *Параметры вывода* – вводится ссылка на *Новый рабочий лист Описательная статистика*;
- устанавливаются флажки *Итоговая статистика*, *Уровень надежности* (95%), *К-й наибольший, К-й наименьший*.

В указанном выходном диапазоне получится результат анализа *Описательная статистика*, приведенный на рис. 14.

	A	B
1	*Выборка X*	
2		
3	Среднее	120,88
4	Стандартная ошибка	0,92212709
5	Медиана	119
6	Мода	124
7	Стандартное отклонение	6,52042317
8	Дисперсия выборки	42,5159184
9	Эксцесс	0,07552499
10	Асимметричность	0,52602494
11	Интервал	30
12	Минимум	110
13	Максимум	140
14	Сумма	6044
15	Счет	50
16	Наибольший(1)	140
17	Наименьший(1)	110
18	Уровень надежности(95,0%)	1,85308264

Рис. 14. **Результат анализа *Описательная статистика***

Вывод. Найденные характеристики выборки определяют расположение и разброс данных. Значения эксцесса $E = 0{,}075 > 0$ и асимметрии $A_s = 0{,}5268 > 0$ позволяют выдвинуть предположение о «более» острой вершине и правосторонней асимметрии плотности вероятностей случайной величины ξ относительно нормального распределения с теми же параметрами $a = 120{,}88$, $\sigma = 6{,}52$.

2. Точечные оценки параметров α и σ нормального распределения. Согласно *Методу моментов* находятся начальные теоретические моменты первого и второго порядков $\nu_1 = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} xe^{-\frac{(x-a)^2}{2\sigma^2}} dx = a$, $\mu_2 = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (x-a)^2 e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \sigma^2$ и соответствующие выборочные моменты $\nu_1^* = \frac{1}{n}\sum\limits_{i=1}^{k} x_1 n_i = \overline{x}$, $\mu_2^* = \frac{1}{n}\sum\limits_{i=1}^{k} (x_i - \overline{x})^2 n_i = \sigma_в^2$ для оценки двух неизвестных параметров a и σ плотности нормального распределения вероятностей $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$.

Таким образом, приравниваем $\tilde{a} = \overline{x}$, $\tilde{\sigma} = \sigma_в$.

Значение оценки параметра a получено в *Описательной стати-*

стике в строке *Среднее*. Оценка параметра σ получается по формуле =(B15-1)/B15*B8, в которой вводятся ссылки на соответствующие ячейки по формуле (СЧЕТ–1)/СЧЕТ*ДИСПЕРСИЯ ВЫБОРКИ.

Согласно *методу максимального правдоподобия* для плотности нормального распределения вероятностей $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$ составляется функция правдоподобия

$$L(x_1,x_2,\ldots,x_{50},\lambda)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{50}e^{-\frac{\sum\limits_{i=1}^{50}(x_i-a)^2}{2\sigma^2}},$$

логарифмическая функция правдоподобия которой имеет следующий вид:

$$\ln L=-50\cdot\ln\left(\sigma\sqrt{2\pi}\right)-\frac{\sum\limits_{i=1}^{50}\left(x_i-a\right)^2}{2\sigma^2}.$$

Данная функция исследуется на экстремум в следующем порядке:

• находятся частные производные первого порядка:

$$\frac{\partial\ln L}{\partial a}=\frac{\sum\limits_{i=1}^{50}x_i-50a}{\sigma^2},\ \frac{\partial\ln L}{\partial\sigma}=-\frac{50}{\sigma}+\frac{\sum\limits_{i=1}^{50}\left(x_i-a\right)^2}{\sigma^3};$$

• находится точка экстремума, т. е. частные производные приравниваются к нулю, и решается система уравнений:

$$\begin{cases}\dfrac{\sum\limits_{i=1}^{50}x_i-50a}{\sigma^2}=0,\\ -\dfrac{50}{\sigma}+\dfrac{\sum\limits_{i=1}^{50}\left(x_i-a\right)^2}{\sigma^3}=0,\end{cases}\Rightarrow\begin{cases}a=\dfrac{1}{50}\sum\limits_{i=1}^{50}x_i,\\ \sigma^2=\dfrac{1}{50}\sum\limits_{i=1}^{50}\left(x_i-a\right)^2,\end{cases}\Rightarrow\begin{cases}a=\bar{x},\\ \sigma^2=\sigma_в^2.\end{cases}$$

Следовательно, в качестве точечных оценок неизвестных параметров a и σ нормального распределения берутся $\tilde{a}=\bar{x}$ и $\tilde{\sigma}=\sigma_в$.

3. Интервальные оценки. Пусть X – выборка из генеральной совокупности значений нормально распределенной случайной величины

$\xi \sim N(a; \sigma^2)$, где $a = M(\xi)$ и $\sigma^2 = D(\xi)$.

Доверительный интервал для a при известной σ^2. Результаты нахождения доверительного интервала представлены в виде таблицы в среде *MS Excel* (рис. 15).

	A	B
1	Объем выборки	50
2	Среднее	120,88
3	Стандартное отклонение	6,5204
4	Предельная ошибка среднего $\Delta_x = t\frac{\sigma}{\sqrt{n}}$	1,807328355
5	Нижняя граница	119,0726716
6	Верхняя граница	122,6873284

Рис. 15. **Границы доверительного интервала для *a* при известной σ^2**

Содержимое ячеек на новом листе *Доверит. интервал 1* заполняется следующим образом:

- в ячейки B1–B3 копируются данные из *Описательной статистики* (объем выборки, среднее, стандартное отклонение);
- ячейка B4 содержит формулу =ДОВЕРИТ(0,05;B3;B1);
- ячейка B5 содержит формулу =B2–B4;
- ячейка B6 содержит формулу =B2+B4.

Вывод. Доверительный интервал (119,07;122,69) накрывает с вероятностью 0,95 математическое ожидание *a* случайной величины ξ при известной дисперсии σ^2. Значит, средний расход используемого сырья по данной технологии (математическое ожидание) находится в указанном интервале в 95% случаев наблюдений, и только в 5% параметр может выйти из этого интервала.

Доверительный интервал для a при неизвестной σ^2. Результаты нахождения доверительного интервала представлены в виде таблицы в среде *MS Excel* (рис. 16).

	A	B
1	Объем выборки	50
2	Среднее	120,88
3	Стандартное отклонение	6,5204
4	Критическая точка распределения Стьюдента $t(\alpha, n-1)$	2,009574018
5	Предельная ошибка среднего $\Delta_x = t(\alpha; n-1) \cdot \frac{S}{\sqrt{n}}$	1,853076052
6	Нижняя граница	119,0269239
7	Верхняя граница	122,7330761

Рис. 16. **Границы доверительного интервала для *a* при неизвестной σ^2**

Содержимое ячеек на новом листе *Доверит. интервал 2* заполняется следующим образом:

- в ячейки B1–B3 копируются данные из *Описательной статистики* (объем выборки, среднее, стандартное отклонение);
- ячейка B4 содержит формулу =СТЬЮДРАСПОБР(0,05;B1-1);
- ячейка B5 содержит формулу =B4*B3/КОРЕНЬ(B1);
- ячейка B6 содержит формулу =B2–B5;
- ячейка B7 содержит формулу =B2+B5.

Замечание. В ячейку B6 можно сразу вписать формулу

=B2-(СТЬЮДРАСПОБР(0,05;B1-1)*B3)/КОРЕНЬ(B1).

Аналогичная формула может быть внесена в ячейку B7.

Вывод. Доверительный интервал (119,027; 122,733) накрывает с вероятностью 0,95 математическое ожидание $M(\xi) = a$ при неизвестной дисперсии σ^2.

Доверительный интервал для σ^2 при известном a. Результаты нахождения доверительного интервала представлены в виде таблицы в среде *MS Excel* (рис. 17).

Содержимое ячеек на новом листе *Доверит. интервал 3* заполняется следующим образом:

- в ячейки B1–B3 копируются данные из *Описательной статистики* (объем выборки, среднее, дисперсия выборки);
- ячейка B4 содержит формулу =(B1-1)*B3/B1 или =ДИСПР(A2:A51), где A2:A51 – ячейки листа *Исходные данные*;
- ячейка B5 содержит формулу =ХИ2ОБР(0,05/2;B1);
- ячейка B6 содержит формулу =ХИ2ОБР(1-(0,05/2);B1);
- ячейка B7 содержит формулу =B1*B4/B5;

• ячейка B8 содержит формулу =B1*B4/B6.

	A	B
1	Объем выборки	50
2	Среднее	120,88
3	Дисперсия выборки (несмещенная)	42,51591837
4	Дисперсия выборочная (смещенная)	41,6656
5	Значение $\chi^2(\frac{\alpha}{2};n)$	71,42019353
6	Значение $\chi^2(1-\frac{\alpha}{2};n)$	32,35738454
7	Нижняя граница $\frac{n\sigma_e^2}{\chi^2(\frac{\alpha}{2};n)}$	29,16934129
8	Верхняя граница $\frac{n\sigma_e^2}{\chi^2(1-\frac{\alpha}{2};n)}$	64,38344847

Рис. 17. **Границы доверительного интервала для σ^2 при известном *a***

Замечание. В ячейки B7 и B8 можно сразу вписать соответственно следующие формулы:

=B1*((B1-1)*B3/B1)/ХИ2ОБР(0,05/2;B1),

=B1*((B1-1)*B3/B1)/ХИ2ОБР(1-(0,05/2);B1).

Вывод. Доверительный интервал (29,169; 64,383) накрывает с вероятностью 0,95 дисперсию $D(\xi) = \sigma^2$ при известном математическом ожидании *a*.

Доверительный интервал для σ^2 при неизвестном a. Результаты нахождения доверительного интервала представлены в виде таблицы в среде *MS Excel* (рис. 18).

	A	B
1	Объем выборки	50
2	Среднее	120,88
3	Дисперсия выборки (несмещенная)	42,51592
4	Значение $\chi^2(\frac{\alpha}{2};n-1)$	70,22236
5	Значение $\chi^2(1-\frac{\alpha}{2};n-1)$	31,55493
6	Нижняя граница $\frac{(n-1)S^2}{\chi^2(\frac{\alpha}{2};n-1)}$	29,6669
7	Верхняя граница $\frac{(n-1)S^2}{\chi^2(1-\frac{\alpha}{2};n-1)}$	66,02074

Рис. 18. **Границы доверительного интервала для σ^2 при неизвестном *a***

Содержимое ячеек на новом листе *Доверит. интервал 4* заполняется следующим образом:

- в ячейки B1-B3 копируются данные из *Описательной статистики* (объем выборки, среднее, дисперсия выборки);
- ячейка B4 содержит формулу =ХИ2ОБР(0,05/2;B1-1);
- ячейка B5 содержит формулу =ХИ2ОБР(1-(0,05/2);B1-1);
- ячейка B6 содержит формулу =(B1-1)*B4/B5;
- ячейка B7 содержит формулу =(B1-1)*B4/B6.

Замечание. В ячейки B6 и B7 можно сразу вписать соответственно следующие формулы:

=(B1-1)*B3/ХИ2ОБР(0,05/2;B1-1),

=(B1-1)*B3/ХИ2ОБР(1-(0,05/2);B1-1).

Вывод. Доверительный интервал (29,67; 66,02) накрывает с вероятностью 0,95 дисперсию $D(\xi) = \sigma^2$ при неизвестном математическом ожидании *a*.

Тема 4. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ О ПАРАМЕТРАХ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

План

1. Основные понятия.

2. Понятие *P*-значения.
3. Проверка статистических гипотез.
4. Лабораторная работа.

1. Основные понятия

Пусть ξ – наблюдаемая дискретная или непрерывная случайная величина. *Статистической гипотезой H* называется утверждение, в котором высказывается предположение относительно параметров или вида закона распределения случайной величины ξ. Гипотеза *H* называется *простой*, если она содержит только одно предположение, а гипотеза, которая состоит из конечного или бесконечного числа простых гипотез, называется *сложной*. Гипотезы о значениях параметров известного распределения случайной величины ξ называются *параметрическими*. *Непараметрическими* называются гипотезы, сформулированные относительно вида закона распределения случайной величины ξ. Основная выдвинутая гипотеза называется *нулевой* H_0.

Гипотеза, противоречащая нулевой гипотезе H_0, называется *альтернативной (конкурирующей) гипотезой* H_1. Выбор альтернативной гипотезы определяется конкретной формулировкой задачи.

Правило, по которому принимается решение принять или отклонить гипотезу H_0, называется *критерием K*. Случайная величина *K*, с помощью которой принимается решение о принятии или отклонении нулевой гипотезы, называется с*татистикой K* критерия *K*. Проверка статистической гипотезы основывается на *принципе отношения правдоподобия*: маловероятные события считаются невозможными, а события, имеющие большую вероятность, считаются достоверными. Зафиксируем некоторую малую вероятность α – *уровень значимости*. Пусть *W* – множество значений статистики *K*, $W_K \subseteq W$ – такое подмножество, для которого $P(K \in W_K \mid H_0 \text{ верна}) = \alpha$.

Наблюдаемым значением $K_{набл}$ называется значение статистики *K*, вычисленное по выборке $X = \{x_1, x_2, \ldots, x_n\}$ из генеральной совокупности случайной величины ξ. Правило проверки статистических гипотез состоит в следующем:

- если $K_{набл} \in W_K$, то гипотеза H_0 отклоняется с вероятностью $P(K_{набл} \in W_K \mid H_0 \text{ верна}) = \alpha$;
- если $K_{набл} \in W \setminus W_K$, то гипотеза H_0 принимается с вероятностью

$P(K_{набл} \in W \setminus W_K \mid H_0 \text{ верна}) = 1 - \alpha$.

Критерий, основанный на использовании заранее заданного уровня значимости α, называется *критерием значимости*. Вероятность $p = 1 - \alpha$ называется *доверительной вероятностью*.

Критической областью W_K называется совокупность значений статистики K, при которых нулевую гипотезу отвергают. *Областью принятия* гипотезы называется совокупность $W \setminus W_K$ значений статистики K, при которых нулевая гипотеза принимается. *Критическими точками* $K_{кр}$ называются точки, отделяющие критическую область W_K от области принятия гипотезы $W \setminus W_K$. Уровень значимости α определяет «размер» критической области W_K. Положение критической области на множестве статистики W зависит от вида нулевой и альтернативной гипотез. Возможны три вида расположения критической области:

- правосторонняя критическая область $(K_{кр\ пр}; +\infty)$;
- левосторонняя критическая область $(-\infty; K_{кр\ лев})$;
- двусторонняя критическая область $(-\infty; K_{кр\ лев}) \cup (K_{кр\ пр}; +\infty)$.

Точки $K_{кр\ пр}$, $K_{кр\ лев}$ определяются в зависимости от вида закона распределения статистики K при выбранном уровне значимости α.

Выбор между гипотезами H_0 и H_1 может сопровождаться ошибками двух родов. *Ошибка первого рода* состоит в том, что будет отвергнута правильная нулевая гипотеза H_0. Вероятность ошибки первого рода равна уровню значимости α:

$$\alpha = P(\text{отвергнуть } H_0 \mid H_0 \text{ верна}).$$

Ошибка второго рода заключается в том, что будет принята неправильная нулевая гипотеза H_0 с вероятностью β:

$$\beta = P(\text{принять } H_0 \mid H_0 \text{ неверна}).$$

Вероятности ошибок должны быть малыми, они выбираются заранее.

При проверке гипотезы возникает одна из четырех ситуаций, приведенных в табл. 8.

Таблица 8. **Ошибки первого и второго рода**

Результаты проверки гипотезы	*Возможные состояния гипотезы*	
	H_0 верна	*H_0 неверна*
Гипотеза H_0 отклоняется	Ошибка первого рода	Правильный вывод

Гипотеза H_0 принимается	Правильный вывод	Ошибка второго рода

Мощностью критерия называется вероятность попадания статистики K в критическую область W_K при условии, что справедлива конкурирующая (альтернативная) гипотеза H_1. Мощность критерия равна вероятности γ правильного отклонения нулевой гипотезы H_0:

$$\gamma = 1 - \beta = P(\text{отвергнуть } H_0 \,|\, H_0 \text{ неверна}).$$

Поскольку критическая область W_K определяется по-разному на заданном уровне значимости α, то она выбирается так, чтобы мощность критерия γ была возможно большей:

$$P(K_{набл} \in W_K \,|\, H_0 \text{ неверна}) \to \max.$$

Чем больше мощность критерия, тем меньше вероятность принятия неверной гипотезы. Мощность критерия позволяет выбрать оптимальную статистику K для проверки гипотезы среди возможных статистик критерия.

На практике в качестве статистики K чаще всего используются специально подобранные случайные величины, распределения которых известны:

- Z (стандартизированное нормальное распределение);
- t (распределение Стьюдента);
- χ^2 (закон Пирсона «хи-квадрат»);
- F (распределение Фишера).

Общая схема проверки статистических гипотез несмотря на разнообразие гипотез и применяемых статистик выглядит следующим образом:

1. На основании выборочных данных выдвигаются нулевая гипотеза H_0 и альтернативная ей гипотеза H_1.

2. Выбирается уровень значимости α (в практических задачах пользуются стандартными значениями уровня значимости: α = 0,1; 0,05; 0,025; 0,01; 0,005; 0,001).

3. Выбирается статистика K, имеющая известный закон распределения.

4. Вычисляется наблюдаемое значение статистики $K_{набл}$ по выборочным данным.

5. Определяются критические точки $K_{кр}$ из условия $P(K \in W_K \,|\, H_0$

верна) = α и область принятия гипотезы в зависимости от формулировки альтернативной гипотезы.

6. Принимается статистическое решение: если $K_{набл}$ попадает в критическую область, то нулевая гипотеза H_0 отвергается, в противном случае H_0 принимается.

2. Понятие *P*-значения

Наряду с критерием значимости для проверки нулевой гипотезы существует правило, основанное на *P*-значении. Пусть $X = \{x_1, x_2, \ldots, x_n\}$ – выборка из генеральной совокупности значений случайной величины ξ, H_0 – выдвинутая гипотеза, K – выбранная статистика, которая является случайной величиной и при сформулированной гипотезе H_0 имеет известный закон распределения, $K_{набл}$ – наблюдаемое значение статистики, вычисленное по выборке. Из уравнения $P(K_{набл} \in W_K \mid H_0 \text{ верна}) = \alpha$ находится неизвестное значение α, называемое *P-значением*.

P-значение представляет собой максимальный уровень значимости, при котором K еще попадает в критическую область гипотезы, и называется *критическим уровнем значимости*.

В зависимости от найденного *P*-значения возможен один из следующих выводов:

- если $P > 0{,}1$, то имеется хорошее согласие с H_0;
- если $0{,}01 \le P \le 0{,}05$, то существуют сомнения в истинности H_0;
- если $P < 0{,}01$, то имеется довольно сильный довод против гипотезы H_0;
- если $P \le 0{,}01$, то гипотеза H_0 почти не подтверждается.

В общем случае, если критический уровень меньше заданного уровня значимости α, то гипотеза отклоняется.

3. Проверка статистических гипотез

Высказываемые в ходе решения задач гипотезы можно условно подразделить на следующие типы:

- о виде закона распределения исследуемой случайной величины;

• об однородности двух или нескольких выборок;

• о числовых значениях параметров исследуемого признака;

• об общем виде зависимости, существующей между компонентами исследуемого многомерного признака.

Гипотезы о законе распределения исследуемой случайной величины ξ. Проверка гипотезы о законе распределения осуществляется с помощью критериев согласия, основанных на выборе определенной меры (т. е. критерия) расхождения между теоретическим и эмпирическим распределениями. Если такая мера расхождения для рассматриваемого случая попадает в критическую область, то гипотеза отвергается, в противном же случае принимается.

Критерий согласия Пирсона χ^2. Пусть генеральная совокупность значений случайной величины ξ имеет неизвестный закон распределения. На основании выборки $X = \{x_1, x_2, \ldots, x_n\}$ выдвигается гипотеза H_0 о конкретном законе распределения ξ (нормальном, биномиальном, показательном и т. д.), выраженном через функцию распределения $F(x)$. Это распределение называется *теоретическим.* По выборке X находится эмпирическая функция распределения $F^*(x)$. Необходимо проверить гипотезу $H_0 : F^*(x) = F(x)$ при альтернативной гипотезе $H_1 : F^*(x) \neq F(x)$.

В критериях согласия иногда альтернативная гипотеза не указывается. Для проверки данной гипотезы используется статистика $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$, имеющая распределение χ^2 с $(k - r - 1)$ степенями свободы, где r – число параметров распределения $F(x)$, которые оцениваются по выборке X; n – объем выборки; k – число непересекающихся интервалов выборочных значений $(x_0 ; x_1)$, $(x_1 ; x_2)$, …; $(x_{k-1} ; x_k)$ …, $x_0 = -\infty$, $x_1 = x_{\min}$, $x_{k-1} = x_{\max}$, $x_k = +\infty$; n_i – число значений выборки, принадлежащих интервалу $[x_i ; x_{i+1})$, $i = 0,1, \ldots, k - 1$; p_i – вероятности попадания значений случайной величины в каждый из этих интервалов: $p_i = P(x_i \le \xi \le x_{i+1}) = \int_{x_i}^{x_{i+1}} p(x)dx = F(x_{i+1}) - F(x_i)$, $i = 0,1,\ldots,k-1$, где $p(x) = F'(x)$ – плотность распределения вероятностей случайной величины ξ. Если для некоторых интервалов условие $n \cdot p_i \ge 5$ не выполняется, то эти интервалы объединяются с соседними.

По выборке X вычисляется наблюдаемое значение $\chi^2_{набл}$ статистики. Для выбранного уровня значимости α по распределению χ^2 находится число $\chi^2_{кр} = \chi^2(\alpha; k - r - 1)$.

Вывод. Гипотеза H_0 не противоречит выборке на заданном уровне значимости, если $\chi^2_{набл} < \chi^2_{кр}$. Если же $\chi^2_{набл} \geq \chi^2_{кр}$, то гипотеза H_0 отвергается.

При работе в *Excel* с целью вычисления наблюдаемого и критического значений статистики χ^2 используются следующие статистические функции *Excel*:

- НОРМРАСП(x; среднее; стандартное_откл; интегральная);
- ХИ2РАСП(вероятность; степени_свободы).

Критерий согласия Колмогорова. Данный критерий основывается на мере γ отклонения эмпирической функции распределения $F^*(x)$ выборки $X = \{x_1, x_2, \ldots, x_n\}$ от теоретической функции распределения $F(x)$ случайной величины ξ. Он применяется в случае, когда гипотетически (по предположению) известны закон распределения $F(x)$ и все его параметры, а на основании опытных данных необходимо подтверждение его справедливости.

По выборке X из генеральной совокупности случайной величины ξ с неизвестной функцией распределения $F(x)$ выдвигается гипотеза $H_0 : F^*(x) = F(x)$.

Для проверки данной гипотезы используется статистика $\lambda = D\sqrt{n}$, где n – объем выборки. Величина $D = \max_x \left| F^*(x) - F(x) \right|$ является случайной величиной, имеющей распределение Колмогорова и характеризующей максимальное отклонение эмпирической функции распределения $F^*(x)$ от теоретической $F(x)$.

По выборке X вычисляется наблюдаемое значение $\lambda_{набл}$ статистики. Задавая уровень значимости α, по распределению Колмогорова определяется критическое значение статистики $\lambda_{кр} = \lambda(\alpha)$.

Вывод. Если $\lambda_{набл} < \lambda_{кр}$, то нулевая гипотеза принимается, т. е. считается, что теоретическая функция распределения $F(x)$ согласуется с результатами наблюдений. Если $\lambda_{набл} \geq \lambda_{кр}$, то H_0 отклоняется.

При работе в *Excel* с целью вычисления наблюдаемого и критиче-

ского значений статистики λ используются следующие статистические функции *Excel*:

- НОРМРАСП(*x*; среднее; стандартное_откл; интегральная);
- МАКС(число1;число2; …);
- ABS(число).

Гипотезы об однородности двух выборок. Пусть ξ_1 и ξ_2 – нормально распределенные случайные величины, $\xi_1 \sim N(a_1;\ \sigma_1^2)$ и $\xi_2 \sim N(a_2;\ \sigma_2^2)$. И пусть $X = \{x_1, x_2, \ldots, x_n\}$ является выборкой значений случайной величины ξ_1, а $Y = \{y_1, y_2, \ldots, y_m\}$ – выборкой значений случайной величины ξ_2. Необходимо по данным выборкам проверить равенства (однородность) математических ожиданий $a_1 = a_2$ и дисперсий $\sigma_1^2 = \sigma_2^2$.

Проверка гипотезы о равенстве математических ожиданий при известных и равных дисперсиях (большие независимые выборки). Будем считать, что дисперсии σ_1^2 и σ_2^2 известны и $\sigma_1^2 = \sigma_2^2$. Выдвигается гипотеза о равенстве математических ожиданий
$H_0 : a_1 = M(\xi_1) = M(\xi_2) = a_2$, причем σ_1^2 и σ_2^2 известны и равны.

Альтернативной гипотезой H_1 может быть одна из следующих гипотез:

- $H_1 : a_1 - a_2 = \delta,\ \delta \neq 0$;
- $H_1 : a_1 - a_2 = \delta > 0$;
- $H_1 : a_1 - a_2 = \delta < 0$.

Проверка нулевой гипотезы осуществляется с помощью *Z*-статистики $Z = \dfrac{\bar{x} - \bar{y} - \delta}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}}$, имеющей стандартное нормальное распределение, $Z \sim N(0;1)$, где *n*, *m* – объемы выборок *X* и *Y*; $\bar{x}$, $\bar{y}$, σ_1^2, σ_2^2 – выборочные средние и известные дисперсии выборок *X* и *Y* соответственно. В качестве оценок дисперсий σ_1^2, σ_2^2 при больших объемах выборок *X* и *Y* используются несмещенные выборочные дисперсии S_X^2, S_Y^2.

Вычисляется наблюдаемое значение $Z_{набл}$ данной статистики.

Для трех случаев альтернативной гипотезы H_1 при заданном уровне значимости α имеют место следующие критические точки $Z_{кр}$

и критические области статистики Z.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$, находится $Z_{кр}$ как квантиль порядка $\frac{1-\alpha}{2}$ нормального распределения, т. е как решение уравнения $\Phi(Z_{кр}) = \frac{1-\alpha}{2}$, где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{y^2}{2}} dy$ – функция Лапласа. Если $|Z_{набл}| < Z_{кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $|Z_{набл}| > Z_{кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ точка $Z_{кр}$ находится как квантиль порядка $\frac{1-2\alpha}{2}$ нормального распределения. Если $Z_{набл} < Z_{кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $Z_{набл} > Z_{кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta < 0$ точка $Z_{кр}$ также находится как квантиль порядка $\frac{1-2\alpha}{2}$ нормального распределения. Если $Z_{набл} > -Z_{кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $Z_{набл} < -Z_{кр}$, то нулевая гипотеза отвергается.

При работе в *Excel* с целью проверки данной гипотезы используется *Двухвыборочный z-тест для средних* из пакета *Анализ данных* меню *Сервис*. Результат анализа появится в виде таблицы. Формулы и соответствующие функции *Excel*, по которым выполняются расчеты в данном режиме, приводятся в табл. 9.

Таблица 9. **Двухвыборочный *z*-тест с одинаковыми дисперсиями**

	Переменная 1	*Переменная 2*
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ СРЗНАЧ	$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ СРЗНАЧ
	Переменная 1	*Переменная 2*
Дисперсия	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ДИСП	$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$ ДИСП
Наблюдения	n	m

	СЧЕТ	СЧЕТ
Гипотетическая разность средних	Число, равное предполагаемой разности математических ожиданий $\delta = a_1 - a_2$	
z	Наблюдаемое значение *Z*-статистики $Z_{набл} = \dfrac{\bar{x} - \bar{y} - \delta}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}}$	
P(Z < = z) одностороннее	Если $H_1 : a_1 - a_2 = \delta > 0$, то $P(Z <= z) = P(Z_{набл} > Z_{кр})$ Если $H_1 : a_1 - a_2 = \delta < 0$, то $P(Z <= z) = P(Z_{набл} < Z_{кр})$ 1-НОРМРАСП(*z*; 0; 1; 1)	
z критическое одностороннее	Модуль значений критических точек *Z*-статистики, определяемых из равенства $\Phi(Z_{кр}) = \dfrac{1 - 2\alpha}{2}$ НОРМСТОБР $(1 - \alpha)$	
P(Z < = z) двухстороннее	Если $H_1 : a_1 - a_2 \neq \delta$, то $P(Z <= z) = 2P(Z_{набл} > \lvert Z_{кр} \rvert)$ 2*(1-НОРМРАСП(*z*; 0; 1; 1))	
z-критическое двухстороннее	Модуль значений критических точек *Z*-статистики, определяемых из равенства $\Phi(Z_{кр}) = \dfrac{1 - \alpha}{2}$ НОРМСТОБР$(1 - \dfrac{\alpha}{2})$	

Проверка гипотезы о равенстве математических ожиданий при неизвестных и равных дисперсиях (малые независимые выборки). Будем считать, что дисперсии σ_1^2 и σ_2^2 неизвестны и равны. Выдвигается гипотеза $H_0 : a_1 = M(\xi_1) = M(\xi_2) = a_2$.

Альтернативной H_1 может быть одна из следующих гипотез:

- $H_1 : a_1 - a_2 = \delta, \delta \neq 0$;
- $H_1 : a_1 - a_2 = \delta > 0$;
- $H_1 : a_1 - a_2 = \delta < 0$.

Проверка нулевой гипотезы осуществляется с помощью *t*-статистики $t = \dfrac{\bar{x} - \bar{y} - \delta}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\dfrac{mn(m+n-2)}{m+n}}$, имеющей распределение Стьюдента с $(n + m - 2)$ степенями свободы, где n, m – объемы выборок X и Y; $\bar{x}$, $\bar{y}$, S_X^2, S_Y^2 – выборочные средние и исправленные дисперсии соответственно.

Вычисляется наблюдаемое значение $t_{набл}$ статистики t.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$, используя распределение Стьюдента, находится точка $t_{кр} = t_{двуст\ кр}(\alpha, n + m - 2)$. Если $|t_{набл}| < t_{двуст\ кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $|t_{набл}| > t_{двуст\ кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ по распределению Стьюдента находится точка $t_{кр} = t_{прав\ кр}(2\alpha; n + m - 2)$. Если $t_{набл} < t_{прав\ кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $t_{набл} > t_{прав\ кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta < 0$ по распределению Стьюдента находится $t_{кр} = t_{лев\ кр}(2\alpha; n + m - 2) = -t_{прав\ кр}(2\alpha; n + m - 2)$. Если $t_{набл} > -t_{прав\ кр}(2\alpha; n + m - 2)$, то нет оснований отвергнуть нулевую гипотезу. Если $t_{набл} < -t_{прав\ кр}(2\alpha; n + m - 2)$, то нулевая гипотеза отвергается.

При работе в *Excel* с целью проверки данной гипотезы используется статистический анализ *Двухвыборочный t-тест с одинаковыми дисперсиями* из *Анализа данных* меню *Сервис.*

В выбранном выходном диапазоне появится результат анализа в виде таблицы. Формулы и соответствующие функции *Excel*, по которым рассчитываются значения, приводятся в табл. 10.

Таблица 10. **Двухвыборочный *t*-тест с одинаковыми дисперсиями**

	Переменная 1	*Переменная 2*
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ СРЗНАЧ	$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ СРЗНАЧ
Дисперсия	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ДИСП	$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$ ДИСП
Наблюдения	n СЧЕТ	m СЧЕТ
Объединенная дисперсия	$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$	

Гипотетическая разность средних	Число, равное предполагаемой разности средних $\delta = a_1 - a_2$
df	Число степеней свободы $v = n + m - 2$
t-статистика	Наблюдаемое значение статистики t $t_{набл} = \dfrac{\bar{x} - \bar{y} - \delta}{S\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$
P(T < = t) одностороннее	Если $H_1 : a_1 - a_2 > \delta$, то $P = P(t_{кр} > t_{набл})$ Если $H_1 : a_1 - a_2 < \delta$, то $P = P(t_{кр} < t_{набл})$ СТЬЮДРАСП(*t*; *df*;1)
t критическое одностороннее	Критическая точка распределения Стьюдента $t_{кр} = t(2\alpha; n + m - 2)$ СТЬЮДРАСПОБР(2α; *df*)
P(T < = t) двухстороннее	Если $H_1 : a_1 - a_2 \neq \delta$, то $P = 2P(t_{кр} > \lvert t_{набл} \rvert)$ СТЬЮДРАСП(*t*; *df*;2)
t критическое двухстороннее	Критическая точка распределения Стьюдента $t_{кр} = t(\alpha; n + m - 2)$ СТЬЮДРАСПОБР(α; *df*)

Проверка гипотезы о равенстве математических ожиданий при неизвестных и неравных дисперсиях. Будем считать, что дисперсии σ_1^2 и σ_2^2 неизвестны и неравны. Выдвигается гипотеза

$$H_0 : a_1 = M(\xi_1) = M(\xi_2) = a_2.$$

Альтернативной H_1 может быть одна из следующих гипотез:

- $H_1 : a_1 - a_2 = \delta, \delta \neq 0$;
- $H_1 : a_1 - a_2 = \delta > 0$;
- $H_1 : a_1 - a_2 = \delta < 0$.

Проверка данной гипотезы осуществляется с помощью *t*-статистики $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}}$, распределение которой близко к распределению Стьюдента с числом степеней свободы, равным

$$\nu = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{\left(\frac{S_X^2}{n}\right)}{n+1} + \frac{\left(\frac{S_Y^2}{m}\right)}{m+1}} - 2,$$

округленным до целого числа. Здесь n, m – объемы выборок; $\bar{x}$, $\bar{y}$, S_X^2, S_Y^2 – выборочные средние и выборочные исправленные дисперсии соответственно. В случае равных объемов выборок ($n = m$)

$$\nu = n - 1 + \frac{2n-2}{\frac{S_X^2}{S_Y^2} + \frac{S_Y^2}{S_X^2}}.$$

Вычисляется наблюдаемое значение $t_{набл}$ статистики t.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$, используя распределение Стьюдента, находится точка $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Если $|t_{набл}| < t_{двуст\ кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $|t_{набл}| > t_{двуст\ кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ по распределению Стьюдента находится точка $t_{кр} = t_{прав\ кр}(2\alpha; \nu)$. Если $t_{набл} < t_{прав\ кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $t_{набл} > t_{прав\ кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta < 0$ находится точка $t_{кр} = t_{лев\ кр}(2\alpha; \nu) = -t_{прав\ кр}(2\alpha; \nu)$. Если $t_{набл} > -t_{прав\ кр}(2\alpha; \nu)$, то нет оснований отвергнуть нулевую гипотезу. Если $t_{набл} < -t_{прав\ кр}(2\alpha; \nu)$, то нулевая гипотеза отвергается.

При работе в *Excel* с целью проверки данной гипотезы используется статистический анализ *Двухвыборочный t-тест с разными дисперсиями* из *Анализа данных* меню *Сервис*. Результат анализа появится в виде таблицы. Формулы и соответствующие функции *Excel*, по которым выполняются расчеты в данном режиме, приводятся в табл. 11.

Таблица 11. **Двухвыборочный *t*-тест с различными дисперсиями**

	Переменная 1	*Переменная 2*
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$

	СРЗНАЧ	СРЗНАЧ
Дисперсия	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ ДИСП	$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2$ ДИСП
Наблюдения	*n* СЧЕТ	*m* СЧЕТ
Гипотетическая разность средних	Число, равное предполагаемой разности средних $\delta = a_1 - a_2$	
df	Число степеней свободы $\nu = \frac{\left(S_X^2/n + S_Y^2/m\right)^2}{\frac{S_X^2/n}{n+1} + \frac{S_Y^2/m}{m+1}} - 2$, округленное до целого числа. В случае равных объемов выборок ($n = m$) $\nu = n - 1 + \frac{2n-2}{\frac{S_X^2}{S_Y^2} + \frac{S_Y^2}{S_X^2}}$	
t-статистика	Наблюдаемое значение *t*-статистики $t_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{S_X^2/n + S_Y^2/m}}$	
P(T <= t) одностороннее	Если $H_1 : a_1 - a_2 > \delta$, то $P = P(t_{кр} > t_{набл})$ Если $H_1 : a_1 - a_2 < \delta$, то $P = P(t_{кр} < t_{набл})$ СТЬЮДРАСП(*t*; *df*;1)	
t критическое одностороннее	Критическая точка распределения Стьюдента $t = t(2\alpha; \nu)$ СТЬЮДРАСПОБР(2α; *df*)	
P(T <= t) двухстороннее	Если $H_1 : a_1 - a_2 \neq \delta$, то $P = 2P(t_{кр} > \lvert t_{набл} \rvert)$ СТЬЮДРАСП(*t*; *df*;2)	
t критическое двухстороннее	Критическая точка распределения Стьюдента $t = t(\alpha; \nu)$ СТЬЮДРАСПОБР(α; *df*)	

Проверка гипотезы о равенстве дисперсий нормальных выборок. Пусть ξ_1 и ξ_2 – нормально распределенные случайные величины, $\xi_1 \sim N(a_1; \sigma_1^2)$ и $\xi_2 \sim N(a_2; \sigma_2^2)$. И пусть $X = \{x_1, x_2, \ldots, x_n\}$ является выборкой значений случайной величины ξ_1, а $Y = \{y_1, y_2, \ldots, y_m\}$ – выборкой значений случайной величины ξ_2. Необходимо по данным выборкам проверить равенство дисперсий σ_1^2 и σ_2^2.

Выдвигается гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$, причем a_1, a_2 неизвестны.

Альтернативной H_1 может быть одна из следующих гипотез:

- $H_1 : \sigma_1^2 > \sigma_2^2$ (если $S_X^2 > S_Y^2$) или $H_1 : \sigma_1^2 < \sigma_2^2$ (если $S_X^2 < S_Y^2$);
- $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Проверка гипотезы H_0 осуществляется с помощью F-статистики $F = \dfrac{\max\{S_X^2; S_Y^2\}}{\min\{S_X^2; S_Y^2\}} = \dfrac{S_1^2}{S_2^2}$, имеющей распределение Фишера с ν_1 и ν_2 степенями свободы, где S_X^2, S_Y^2 – выборочные исправленные дисперсии, n, m – объемы выборок, ν_1 – число степеней большей исправленной дисперсии ($\nu_1 = n - 1$ или $\nu_1 = m - 1$), ν_2 – число степеней меньшей исправленной дисперсии ($\nu_2 = m - 1$ или $\nu_2 = n - 1$).

Вычисляется наблюдаемое значение $F_{набл}$ статистики F.

Для двух случаев альтернативной гипотезы H_1 при заданном уровне значимости α имеют место следующие критические точки статистики $F_{кр}$ и критические области.

При альтернативной гипотезе $H_1 : \sigma_1^2 > \sigma_2^2$ по распределению Фишера находится точка $F_{кр} = F(\alpha; \nu_1; \nu_2)$. Если $F_{набл} < F_{кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $F_{набл} > F_{кр}$, то нулевая гипотеза отвергается.

При альтернативной гипотезе $H_1 : \sigma_1^2 \neq \sigma_2^2$, используя распределение Фишера, находим следующие точки:

$$F_{кр\ лев} = F\left(1 - \frac{\alpha}{2}; \nu_1; \nu_2\right),\ F_{кр\ пр} = F\left(\frac{\alpha}{2}; \nu_1; \nu_2\right).$$

Если $F_{кр\ лев} < F_{набл} < F_{кр\ пр}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевая гипотеза отвергается.

При работе в *Excel* с целью проверки данной гипотезы используется статистический анализ *Двухвыборочный F-тест для дисперсии* из *Анализа данных* меню *Сервис*.

Результат анализа появится в виде таблицы. Формулы и соответствующие функции *Excel*, по которым выполняются расчеты в данном режиме, приводятся в табл. 12.

Таблица 12. **Двухвыборочный *F*-тест для дисперсии**

	Переменная 1	*Переменная 2*
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ СРЗНАЧ	$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ СРЗНАЧ
Дисперсия	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ДИСП	$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$ ДИСП
Наблюдения	n СЧЕТ	m СЧЕТ
df	Число степеней свободы	
	$n - 1$	$m - 1$
F	Наблюдаемое значение *F*-статистики $F_{набл} = \frac{\min\left\{S_X^2; S_Y^2\right\}}{\max\left\{S_X^2; S_Y^2\right\}} = \frac{S_1^2}{S_2^2}$	
P(F <= f) одностороннее	Если $H_1 : \sigma_X^2 > \sigma_Y^2$ ($H_1 : \sigma_X^2 < \sigma_Y^2$), то $P = P(F_{кр} < F_{набл})$ FРАСП(F; ν_1; ν_2) ν_1 – число степеней большей исправленной дисперсии ($\nu_1 = n - 1$ или $\nu_1 = m - 1$), ν_2 – число степеней меньшей исправленной дисперсии ($\nu_2 = m - 1$ или $\nu_2 = n - 1$)	
F критическое одностороннее	Критическое значение F-критерия определяется по таблицам распределения Фишера $F_{кр} = F(\alpha; \nu_1; \nu_2)$ FРАСПОБР($1 - \alpha$; ν_1; ν_2)	

Данный режим рассчитывает только односторонние оценки *P*-значения и $F_{кр} = F(\alpha; \nu_1; \nu_2)$, поскольку режим *Двухвыборочный F-тест для дисперсии* при проверке гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ в качестве альтернативной рассматривает гипотезу $H_1 : \sigma_1^2 > \sigma_2^2$ (если $S_X^2 < S_Y^2$) или $H_1 : \sigma_1^2 < \sigma_2^2$ (если $S_X^2 > S_Y^2$). Чтобы получить двустороннюю оценку для $F_{кр}$ (в этом случае рассматривается альтернативная гипотеза $H_1 : \sigma_1^2 \neq \sigma_2^2$), используется функция

FРАСПОБР(вероятность;степени_свободы1;степени_свободы2).

Тогда значение левосторонней критической точки равно

$$F_{кр\ лев}=\text{FРАСПОБР}(1-\frac{\alpha}{2};\nu_1;\nu_2),$$

значение правосторонней критической точки –

$$F_{кр\ пр}=\text{FРАСПОБР}(\frac{\alpha}{2};\nu_1;\nu_2).$$

Двусторонняя критическая область – объединение двух интервалов $(-\infty;F_{кр\ лев})\cup(F_{кр\ пр};+\infty)$.

Двустороннее *P*-значение рассчитывается с помощью функции ФТЕСТ(массив1;массив2).

Вопросы для самоконтроля

1. Что называется гипотезой?
2. Какая гипотеза называется нулевой, альтернативной, простой и сложной?
3. Что называется критерием и мощностью критерия?
4. Что определяет уровень значимости гипотезы?
5. Что понимают под критической областью критерия?
6. Как найти доверительную вероятность статистического критерия?
7. Какие виды ошибок могут быть при проверке гипотез?
8. Что называется мощностью критерия?
9. В чем состоит смысл *P*-значения?
10. Какие статистики используются при проверке гипотез о законе распределения?
11. Какие статистики используются при проверке гипотез о равенстве математических ожиданий двух нормальных выборок?
12. Какая статистика используется при проверке гипотезы о равенстве дисперсий нормальных выборок?

4. Лабораторная работа

Задание

Пусть случайная величина ξ_1 характеризует расход сырья при производстве продукции по одной технологии, ξ_2 – по другой технологии, причем предполагается, что ξ_1 и ξ_2 нормально распределены. В результате наблюдений получены выборка *X* из генеральной совокупности случайной величины ξ_1 и выборка *Y* из генеральной совокупности случайной величины ξ_2.

Выборка *X*:

114	112	132	124	119	124	119	116	129	116
124	119	119	114	129	116	124	129	116	119
110	124	140	119	124	129	119	124	124	124
116	129	119	124	110	124	112	114	129	116
119	116	129	116	119	114	132	119	124	112

Выборка *Y*:

109	137	111	133	111	126	111	114	114	114
119	122	119	122	122	122	119	122	122	122
114	119	114	114	119	119	114	119	119	119
122	123	122	123	123	123	122	123	123	123
123	133	123	111	126	126	126	126	126	126
133	114	133	133	111	135	133	135	135	137

Требуется при заданном уровне значимости $\alpha = 0{,}05$ проверить следующие гипотезы:

1) о нормальном законе распределения генеральной совокупности случайной величины ξ_1, используя критерии Пирсона и Колмогорова;

2) о равенстве дисперсий двух нормальных генеральных совокупностей, математические ожидания которых неизвестны;

3) о равенстве математических ожиданий двух нормальных генеральных совокупностей, дисперсии которых известны и равны;

4) о равенстве математических ожиданий двух нормальных генеральных совокупностей, дисперсии которых неизвестны и равны;

5) о равенстве математических ожиданий двух нормальных генеральных совокупностей, дисперсии которых неизвестны и различны.

Порядок выполнения работы

В ячейки A1:A51 листа *Исходные данные* вводится название «Выборка *X*» и ее значения, в ячейки B1:B61 – «Выборка *Y*» и ее значения.

Работа выполняется в следующем порядке:

1. Проверка гипотезы о законе распределения. Согласно к*ритерию Пирсона* по выборке *X* значений случайной величины ξ_1 при заданном уровне значимости $\alpha = 0{,}05$ выдвигается гипотеза $H_0 : F^*(x) = F(x)$, где $F(x)$ – известная функция нормального распределения

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt$$

с параметрами *a* и σ^2. При достаточно большом объеме выборки *X* в качестве оценок *a* и σ^2 используются $\bar{x}$, σ_X^2 соответственно.

Расчеты по нахождению $\chi^2_{набл}$ выполнены с использованием предварительно полученных результатов анализов *Описательная статистика* (см. лабораторную работу темы 3) и *Гистограмма* (см. лабораторную работу темы 2), представлены в виде таблицы в среде *MS Excel* (рис. 19) на листе *Критерий Пирсона*.

	A	B	C	D	E	F
1	*Правый конец интервала* $[x_i; x_{i+1})$	*Частота* n_i	*Значение гипотетической функции распределения* $F(x_{i+1})$	*Вероятность попадания в интервал* $p_i = F(x_{i+1}) - F(x_i)$	$n \cdot p_i$	$\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$
2			0			
3	110,00	2	0,047597984	0,047597984	2,37989921	0,060642655
4	114,29	7	0,155929775	0,10833179	5,41658952	0,462872206
5	118,57	8	0,36164987	0,205720095	10,2860048	0,508051269
6	122,86	11	0,619140008	0,257490138	12,8745069	0,272925104
7	127,14	12	0,831597835	0,212457827	10,6228914	0,178522792
8	131,43	7	0,947144358	0,115546523	5,77732613	0,258758354
9	135,71	2	0,988548555	0,041404197	2,07020987	0,002381124
10	400,00	1	1	0,011451445	0,57257225	0,319076727
11	*Среднее*	120,88				
12	*Стандартное отклонение*	6,52042317				
13	*Объем выборки*	50				
14	*Наблюдаемое значение статистики* $\chi^2_{набл}$	2,06323023				
15	*Критическое значение статистики* $\chi^2_{кр}$	11,07048				

Рис. 19. **Проверка гипотезы о законе распределения с помощью критерия Пирсона**

Содержимое ячеек заполняется следующим образом:

- В ячейки A3:B10 копируются столбцы *Карман* и *Частоты* из *Ги-*

стограммы (см. лабораторную работу темы 2). Поскольку правый конец последнего интервала есть неопределенность +∞, то заменяется *Еще* на число в 2–3 раза большее *Максимума* из *Описательной статистики*. Для этого вводится формула = 3* Описательная статистика! B13 (используется ссылка на ячейку B13 со значением максимума остатков листа *Описательная статистика*).

• В ячейки B11–B13 копируются значения *Среднее*, *Стандартное отклонение* и *Счет* из *Описательной статистики* соответственно.

• Ячейки C3:C10 содержат формулу массива
{=НОРМРАСП(A3:A10;B11;B12;ИСТИНА)}.

• В ячейку C2 вводится значение 0 (так как $F(-\infty) = 0$).

• Ячейки D3:D10 содержат формулу массива {=C3:C10-C2:C9}.

• Ячейки E3:E10 содержат формулу массива {=B13*D3:D10}.

• Ячейки F3:F10 содержат формулу массива
{=(B3:B10-E3:E10)^2/E3:E10}.

• Ячейка B14 содержит формулу =СУММ(F3:F10).

• Ячейка B15 содержит формулу =ХИ2ОБР(0,05; 9-2-1), где 9 – число интервалов [–∞;110), [110;114,29), [114,29;118,57), [118,57; 122,86), [122,86;127,14), [127,14;131,43), [131,43;135,71), [135,71;400), [400; +∞); 2 – число параметров (a, σ) нормального распределения.

Вывод. Так как $\chi^2_{набл} = 2{,}06323 < \chi^2_{0,05} = 12{,}59158$, то нет оснований для отклонения нулевой гипотезы о нормальном распределении случайной величины ξ_1 с параметрами $a = 120{,}88$, $\sigma = 6{,}52$. Значит, предположение о нормальном законе расхода сырья при производстве продукции верно в 95% случаев наблюдений.

Согласно *критерию Колмогорова* по выборке X значений случайной величины ξ_1 при заданном уровне значимости $\alpha = 0{,}05$ выдвигается гипотеза $H_0 : F^*(x) = F(x)$, где $F(x)$ – известная функция нормального распределения, $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)}{2\sigma^2}} dt$ с параметрами a и σ^2. При достаточно большом объеме выборки X в качестве оценок a и σ^2 используются $\bar{x}$, σ_X^2 соответственно.

Результаты нахождения $\lambda_{набл}$ представлены в виде таблицы в среде *MS Excel* (рис. 20) с использованием предварительно полученных результатов анализов *Описательная статистика* (см. лабораторную

работу темы 3) и *Гистограмма* (см. лабораторную работу темы 2) на новом листе.

	A	B	C	D
1	*Правый конец интервала* $[x_i; x_{i+1})$	*Значения эмпирической функции распределения* $F^*(x_{i+1})$	*Значения гипотетической функции распределения* $F(x_{i+1})$	*Модуль разностей значений функций* $\lvert F^*(x) - F(x) \rvert$
2	110,00	0,04	0,047597984	0,007597984
3	114,29	0,18	0,155929775	0,024070225
4	118,57	0,34	0,36164987	0,02164987
5	122,86	0,56	0,619140008	0,059140008
6	127,14	0,80	0,831597835	0,031597835
7	131,43	0,94	0,947144358	0,007144358
8	135,71	0,98	0,988548555	0,008548555
9	400,00	1,00	1	0,00E+00
10	*Среднее*	120,88		
11	*Стандартное отклонение*	6,520423174		
12	*Объем выборки*	50		
13	*Максимум модуля разностей* $D = \max \lvert F^*(x) - F(x) \rvert$	0,059140008		
14	*Наблюдаемое значение статистики* $\lambda_{набл} = D\sqrt{n}$	0,418183006		

Рис. 20. **Проверка гипотезы о законе распределения с помощью критерия Колмогорова**

Содержимое ячеек заполняется следующим образом:

• в ячейки A2:A9 копируется столбец *Карман* из *Гистограммы*, заменяется *Еще* на число в 2–3 раза большее *Максимума* из *Описательной статистики,* поскольку правый конец последнего интервала есть неопределенность +∞;

• в ячейки B2:B9 копируется столбец *Интегральный процент*, после этого необходимо изменить числовой формат в данных ячейках с процента на число с помощью команды *Формат – Ячейки…– Число – Числовой – ОК*;

• в ячейки B10–B12 копируются значения *Среднее*, *Стандартное отклонение* и *Счет* из *Описательной статистики*;

• ячейки C2:C9 содержат формулу массива
{=НОРМРАСП(A2:A9;B10;B11;ИСТИНА)};

• ячейки D2:D9 содержат формулу массива {=ABS(B2:B9-C2:C9)};

• ячейка B13 содержит формулу =МАКС(D2:D9);

• ячейка B14 содержит формулу =B13*КОРЕНЬ(B12).

Вывод. Поскольку $\lambda_{набл} = 0{,}4181 < \lambda_{0,05} = 1{,}358$, то нет оснований для отклонения нулевой гипотезы о нормальном законе расхода сырья при производстве продукции.

2. Проверка гипотезы о равенстве дисперсий нормальных выборок при неизвестных математических ожиданиях. По выборкам X и Y значений нормально распределенных случайных величин $\xi_1 \sim N(a_1; \sigma_1^2)$ и $\xi_2 \sim N(a_2; \sigma_2^2)$ с неизвестными математическими ожиданиями a_1, a_2 выдвигается гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ при альтернативной гипотезе H_1, которая может быть одной из следующих:

- $H_1 : \sigma_1^2 > \sigma_2^2$ (если $S_X^2 < S_Y^2$) или $H_1 : \sigma_1^2 < \sigma_2^2$ (если $S_X^2 > S_Y^2$);
- $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Для проверки данной гипотезы используется режим работы *Двухвыборочный F-тест для дисперсии.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 21):

- *Интервал переменной 1* (вводятся ссылки на ячейки A1:A51, в которых находятся название и значения выборки *X*);
- *Интервал переменной 2* (вводятся ссылки на ячейки B1:B61, в которых находятся название и значения выборки *Y*);
- *Гипотетическая средняя разность* (вводится число 0);
- *Метки* (устанавливается флажок);
- *Альфа* (вводится уровень значимости 0,05);
- *Выходной диапазон* (вводится ссылка на *Новый рабочий лист F-тест*).

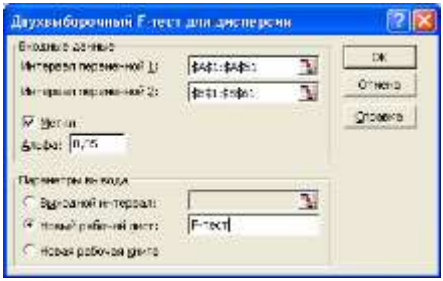


Рис. 21. **Заполнение диалогового окна *Двухвыборочный F-тест для дисперсии***

В выбранном выходном диапазоне ячеек A1:C10 появится результат анализа, представленный на рис. 22.

	A	B	C
1	Двухвыборочный F-тест для дисперсии		
2			
3		*Выборка X*	*Выборка Y*
4	Среднее	120,88	122,2166667
5	Дисперсия	42,51591837	53,6980226
6	Наблюдения	50	60
7	df	49	59
8	F	0,791759478	
9	P(F<=f) одностороннее	0,201061568	
10	F критическое одностороннее	0,631813712	
11	$F_{\text{кр лев}}$	0,577886183	
12	$F_{\text{кр пр}}$	1,70686576	
13	P(F<=f) двустороннее	0,402123136	

Рис. 22. **Результат проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$, где a_1, a_2 неизвестны**

Для определения двусторонней критической области выполняются следующие действия:

- в ячейку B11 вводится формула =FРАСПОБР(1-(0,05/2);B7;C7);
- в ячейку B12 вводится формула =FРАСПОБР(0,05/2; B7;C7);
- в ячейку B13 вводится формула =ФТЕСТ(A2:A51;B2:B61),

где A2:A51 и B2:B61 – ссылки на ячейки, содержащие значения выборок *X* и *Y* соответственно.

Вывод. Наблюдаемое значение $F_{набл}$ попадает в критическую область значений статистики, так как 0,7918 = $F_{набл} > F_{кр}$ = 0,6318. Поэтому есть основания отвергнуть нулевую гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$ на уровне значимости α = 0,05 при альтернативной гипотезе $H_1 : \sigma_1^2 < \sigma_2^2$ (так как $S_X^2 < S_Y^2$). Хотя для *P*-значения имеет место неравенство 0,2 > 0,05, что говорит о хорошем согласии с H_0.

При альтернативной гипотезе $H_1 : \sigma_1^2 \neq \sigma_2^2$ имеем 0,7918 = $F_{набл} \in$ (0,5779; 1,7069), т. е. $F_{набл}$ попадает в область допустимых значений статистики. Так как *P* = 0,4021 > 0,05, значит на уровне значимости α = 0,05 гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ о равенстве дисперсий расхода сырья по двум технологиям принимается.

3. Проверка гипотезы о равенстве математических ожиданий при известных и равных дисперсиях. По выборкам X и Y значений нормально распределенных случайных величин $\xi_1 \sim N(a_1; \sigma_1^2)$ и $\xi_2 \sim N(a_2; \sigma_2^2)$ выдвигается гипотеза о равенстве математических ожиданий $H_0 : a_1 = M(\xi_1) = M(\xi_2) = a_2$ в предположении, что σ_1^2 и σ_2^2 известны и равны, при альтернативной гипотезе H_1, которая может быть одной из следующих:

- $H_1 : a_1 - a_2 = \delta, \delta \neq 0$;
- $H_1 : a_1 - a_2 = \delta > 0$;
- $H_1 : a_1 - a_2 = \delta < 0$.

Для проверки данной гипотезы используется режим анализа *Двухвыборочный z-тест для средних.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 23):

- *Интервал переменной 1* (вводятся ссылки на ячейки A1:A51, в которых находятся название и значения выборки *X*);
- *Интервал переменной 2* (вводятся ссылки на ячейки B1:B61, в которых находятся название и значения выборки *Y*);
- *Гипотетическая средняя разность* (вводится число 0);
- *Дисперсия переменной 1* (*известная*) (вводится предварительно вычисленное с помощью функции ДИСП(A2:A51) значение несмещенной выборочной дисперсии S_X^2, которая является оценкой дисперсии σ_1^2 генеральной совокупности случайной величины ξ_1);
- *Дисперсия переменной 2 (известная)* (вводится предварительно вычисленное с помощью функции ДИСП(B2:B61) значение выборочной дисперсии S_Y^2, которая является оценкой дисперсии σ_2^2 генеральной совокупности случайной величины ξ_2);
- *Метки* (устанавливается флажок);
- *Альфа* (вводится уровень значимости 0,05);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист z-тест*).

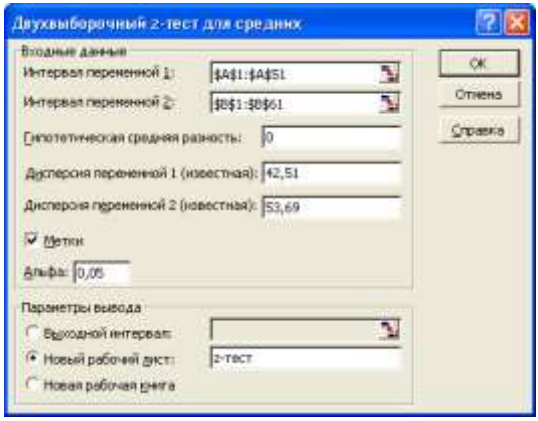


Рис. 23. **Заполнение диалогового окна *Двухвыборочный z-тест для средних***

Рассчитанные в данном режиме показатели представлены на рис. 24.

	A	B	C
1	Двухвыборочный z-тест для средних		
2			
3		*Выборка X*	*Выборка Y*
4	Среднее	120,88	122,2166667
5	Известная дисперсия	42,51	53,69
6	Наблюдения	50	60
7	Гипотетическая разность средних	0	
8	z	-1,011861925	
9	P(Z<=z) одностороннее	0,155802049	
10	z критическое одностороннее	1,644853476	
11	P(Z<=z) двухстороннее	0,311604097	
12	z критическое двухстороннее	1,959962787	

Рис. 24. **Результаты проверки гипотезы $H_0 : a_1 = a_2$, где σ_1^2 и σ_2^2 известны и $\sigma_1^2 = \sigma_2^2$**

Вывод. Так как наблюдаемое значение $Z_{набл}$ попадает в область допустимых значений, т. е. 1,0119 = $|Z_{набл}| < Z_{кр}$ = 1,9599, то нет оснований отвергнуть нулевую гипотезу $H_0 : a_1 = a_2$ на уровне значимости

$\alpha = 0{,}05$ при альтернативной гипотезе $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$. Для P-значения имеет место неравенство $0{,}3116 > 0{,}05$, что также говорит о хорошем согласии с H_0.

Следовательно, средний расход сырья для производства продукции по разным технологиям совпадает в 95% случаев наблюдений и только в 5% случаев различается.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ ($H_1 : a_1 - a_2 = \delta < 0$) имеет место неравенство

$$Z_{набл} = 1{,}0119 < 1{,}6448 = Z_{кр} \ (Z_{набл} = 1{,}0119 > -1{,}6448 = -Z_{кр}).$$

Значит, гипотеза H_0 принимается при заданном уровне значимости $\alpha = 0{,}05$. Следовательно, различие средних расходов сырья по двум разным технологиям незначительно. Это также подтверждается P-значением $P = 0{,}1558 > 0{,}05$.

4. Проверка гипотезы о равенстве математических ожиданий при неизвестных и равных дисперсиях (малые независимые выборки). По выборкам X и Y значений нормально распределенных случайных величин $\xi_1 \sim N(a_1; \sigma_1^2)$ и $\xi_2 \sim N(a_2; \sigma_2^2)$ выдвигается гипотеза о равенстве математических ожиданий $H_0 : a_1 = M(\xi_1) = M(\xi_2) = a_2$ в предположении, что σ_1^2 и σ_2^2 неизвестны и равны, при одной из следующих альтернативных гипотез H_1:

- $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$;
- $H_1 : a_1 - a_2 = \delta > 0$;
- $H_1 : a_1 - a_2 = \delta < 0$.

Для проверки данной гипотезы используется режим анализа *Двухвыборочный t-тест с одинаковыми дисперсиями.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 25):

- *Интервал переменной 1* (вводятся ссылки на ячейки A1:A51, в которых находятся значения выборки X);
- *Интервал переменной 2* (вводятся ссылки на ячейки B1:B61, в которых находятся значения выборки Y);
- *Гипотетическая средняя разность* (вводится число 0);
- *Метки* (устанавливается флажок);

• *Альфа* (вводится уровень значимости 0,05);

• *Параметры вывода* (вводится ссылка на *Новый рабочий лист t-тест-одинак-дисп*).

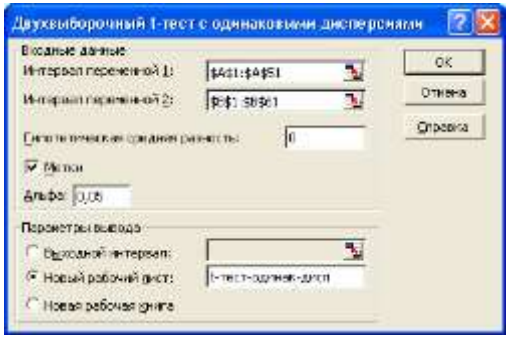


Рис. 25. **Заполнение диалогового окна *Двухвыборочный t-тест с одинаковыми дисперсиями***

На рис. 26 представлены рассчитанные в данном режиме показатели.

	A	B	C
1	Двухвыборочный t-тест с одинаковыми дисперсиями		
2			
3		*Выборка X*	*Выборка Y*
4	Среднее	120,88	122,2166667
5	Дисперсия	42,51591837	53,6960226
6	Наблюдения	50	60
7	Объединенная дисперсия	48,62466049	
8	Гипотетическая разность средних	0	
9	df	108	
10	t-статистика	-1,001057703	
11	P(T<=t) одностороннее	0,159518274	
12	t критическое одностороннее	1,659086593	
13	P(T<=t) двухстороннее	0,319036547	
14	t критическое двухстороннее	1,982170943	

Рис. 26. **Результат проверки гипотезы $H_0 : a_1 = a_2$, где σ_1^2 и σ_2^2 неизвестны и $\sigma_1^2 = \sigma_2^2$**

Вывод. Наблюдаемое значение $t_{набл}$ попадает в область допустимых значений, т. е. $1{,}001 = |t_{набл}| < t_{кр} = 1{,}982$, поэтому нет оснований отвергнуть нулевую гипотезу $H_0 : a_1 = a_2$ на уровне значимости $\alpha = 0{,}05$ при альтернативной гипотезе $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$. Для *P*-значения имеет место неравенство $0{,}319 > 0{,}05$, что также говорит о хорошем

согласии с H_0.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ имеем

$$t_{набл} = 1{,}001 < 1{,}6591 = t_{кр},\ P = 0{,}1595 > 0{,}05,$$

значит гипотеза $H_0 : a_1 = a_2$ принимается при уровне значимости $\alpha = 0{,}05$.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta < 0$ имеем

$$t_{набл} = -1{,}001 > -1{,}6591 = t_{кр},\ P = 0{,}1595 > 0{,}05,$$

значит гипотеза $H_0 : a_1 = a_2$ принимается при уровне значимости $\alpha = 0{,}05$.

5. Проверка гипотезы о равенстве математических ожиданий при неизвестных и неравных дисперсиях. По выборкам X и Y значений нормально распределенных случайных величин $\xi_1 \sim N(a_1; \sigma_1^2)$ и $\xi_2 \sim N(a_2; \sigma_2^2)$ выдвигается гипотеза о равенстве математических ожиданий $H_0 : a_1 = M(\xi_1) = M(\xi_2) = a_2$ в предположении, что σ_1^2 и σ_2^2 неизвестны и неравны, при одной из следующих альтернативных гипотез H_1:

- $H_1 : a_1 - a_2 = \delta,\ \delta \neq 0$;
- $H_1 : a_1 - a_2 = \delta > 0$;
- $H_1 : a_1 - a_2 = \delta < 0$.

Для проверки данной гипотезы используется режим анализа *Двухвыборочный t-тест с различными дисперсиями.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 27):

- *Интервал переменной 1* (вводятся ссылки на ячейки A1:A51, в которых находятся значения выборки *X*);
- *Интервал переменной 2* (вводятся ссылки на ячейки B1:B61, в которых находятся значения выборки *Y*);
- *Гипотетическая средняя разность* (вводится число 0);
- *Метки* (устанавливается флажок);
- *Альфа* (вводится уровень значимости 0,05);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист t-тест-различ-дисп*).

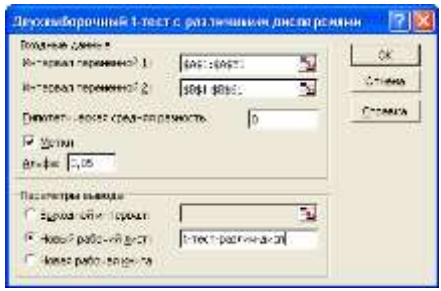


Рис. 27. **Заполнение диалогового окна *Двухвыборочный t-тест с различными дисперсиями***

На рис. 28 представлены рассчитанные в данном режиме показатели.

	A	B	C
1	Двухвыборочный t-тест с различными дисперсиями		
2			
3		*Выборка X*	*Выборка Y*
4	Среднее	120,66	122,2166667
5	Дисперсия	42,51591837	53,69000226
6	Наблюдения	60	60
7	Гипотетическая разность средних	0	
8	df	108	
9	t-статистика	-1,011799846	
10	P(T<=t) одностороннее	0,156950226	
11	t критическое одностороннее	1,659088593	
12	P(T<=t) двухстороннее	0,313900452	
13	t критическое двухстороннее	1,982170943	

Рис. 28. **Результаты проверки гипотезы $H_0 : a_1 = a_2$, где σ_1^2 и σ_2^2 неизвестны и $\sigma_1^2 \neq \sigma_2^2$**

Вывод. Наблюдаемое значение $t_{набл}$ попадает в область допустимых значений, т. е. $1{,}0118 = |t_{набл}| < t_{кр} = 1{,}9832$. Поэтому нет оснований отвергнуть нулевую гипотезу $H_0 : a_1 = a_2$ на уровне значимости $\alpha = 0{,}05$ при альтернативной гипотезе $H_1 : a_1 - a_2 = \delta$, $\delta \neq 0$. Для P-значения имеет место неравенство $0{,}3139 > 0{,}05$, что также говорит о хорошем согласии с H_0.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ имеем

$$t_{набл} = 1{,}0118 < 1{,}6591 = t_{кр},\ P = 0{,}1569 > 0{,}05,$$

значит гипотеза $H_0 : a_1 = a_2$ также принимается при уровне значимости $\alpha = 0{,}05$.

При альтернативной гипотезе $H_1 : a_1 - a_2 = \delta > 0$ имеем

$$t_{набл} = -1{,}0118 > -1{,}6591 = t_{кр},\ P = 0{,}1569 > 0{,}05,$$

значит гипотеза $H_0 : a_1 = a_2$ также принимается при уровне значимости $\alpha = 0{,}05$.

Тема 5. ДИСПЕРСИОННЫЙ АНАЛИЗ

План

1. Дисперсионный анализ.
2. Однофакторный дисперсионный анализ.
3. Двухфакторный дисперсионный анализ.
4. Лабораторная работа.

1. Дисперсионный анализ

Изучение реальных процессов предполагает получение не только прогнозной оценки исследуемого показателя, но и количественной характеристики степени влияния на него различных факторов, а также оценки возможных последствий их изменений в будущем. В результате опыта проводятся наблюдения над целым рядом случайных величин. При этом возникает задача изучения взаимосвязи между случайными величинами, которая решается в три этапа:

- проводится оценка существенности влияния одного фактора на другой с помощью дисперсионного анализа;
- проводится численная оценка связи с помощью корреляционного анализа;
- строятся функциональные зависимости посредством регрессионного анализа.

Дисперсионный анализ служит для статистического установления влияния отдельных факторов на изменчивость какого-либо признака, значения которого могут быть получены опытным путем в виде выборки из генеральной совокупности случайной величины ξ. Под *факторами* понимаются различные независимые показатели, количество которых может быть различным. Конкретная реализация фактора A называется *уровнем* (*группой*) этого фактора. В зависимости от количества факторов различают однофакторный и многофакторный дисперсионный анализ. Величина ξ называется *результативным признаком* (*фактором*) Y. Идея дисперсионного анализа состоит в том, что дисперсия признака Y разлагается на сумму дисперсий, вызван-

ных влиянием факторов, дисперсий, вызванных взаимодействием факторов, и случайной дисперсии, вызванной неучтенными случайными факторами. Затем указанные дисперсии сравниваются.

2. Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ позволяет статистически обосновать степень влияния на результативный признак Y одного фактора A для различных уровней или групп A_1, A_2, …, A_k, например, установление зависимости выполненных на стройке за смену работ (признак Y) от работающей бригады (группы A_1, A_2, …, A_k). В табл. 13 представлены выборочные значения результативного признака Y для различных групп.

Таблица 13. **Данные наблюдений для однофакторного дисперсионного анализа**

Номера наблюдений	Уровни (группы) фактора A			
	A_1	A_2	…	A_k
1	y_{11}	y_{12}	…	y_{1k}
2	y_{21}	y_{22}	…	y_{2k}
…	…	…	…	…
nj	$y_{n_1 1}$	$y_{n_2 2}$	…	$y_{n_k k}$
Групповая средняя	$\bar{y}_1$	$\bar{y}_2$	…	$\bar{y}_k$

Здесь n_1, n_2, …, n_k – число наблюдений в уровнях A_1, A_2, …, A_k соответственно, при этом $n = n_1 + n_2 + \ldots + n_k$ – число всех наблюдений.

Из таблицы находятся групповые средние $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$, j = 1, 2, …, k, и общая средняя $\bar{y} = \frac{1}{n}\sum_{j=1}^{k} \bar{y}_j n_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}$.

Изменчивость значений y_{ij} от одного уровня к другому объясняется влиянием фактора A на результативный признак Y, а изменчивость тех же значений в пределах одного уровня характеризуется влиянием

неучтенных случайных (остаточных) факторов.

Пусть a_1, a_2, …, a_k – математические ожидания результативного признака Y соответственно при уровнях A_1, A_2, …, A_k. Если для различных групп фактора A математические ожидания не изменяются, то считается, что результативный признак Y не зависит от фактора A, в противном случае такая зависимость существует.

Поскольку числовые значения математических ожиданий неизвестны, то возникает задача проверки гипотезы $H_0 : a_1 = a_2 = \ldots = a_k$ в предположении выполнения следующих условий для каждой группы фактора:

- наблюдения независимы и проводятся в одинаковых условиях;
- результативный признак Y имеет нормальный закон распределения с постоянной дисперсией для различных групп.

Поскольку числовые значения дисперсий различных групп неизвестны, то можно установить их равенство, проверив гипотезу о равенстве дисперсий $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ с помощью статистики Бартлетта $V = \frac{1}{C} \cdot \left(\sum_{j=1}^{k} (n_j - 1) \ln \frac{\bar{S}^2}{S_j^2} \right)$, имеющей распределение, близкое к χ^2-распределению с $(k - 1)$ степенями свободы, где

$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{j=1}^{k} \frac{1}{n_j - 1} - \frac{1}{n_1 + n_2 + \ldots + n_k - k} \right), \quad \bar{S}^2 = \frac{\sum_{j=1}^{k} (n_j - 1) S_j^2}{\sum_{j=1}^{k} (n_j - 1)},$$

S_j^2 – выборочная дисперсия j-й группы ($j = 1, 2, \ldots, k$).

При заданном уровне значимости α находится критическая точка $V_{кр} = \chi^2(\alpha; k - 1)$, определяющая правостороннюю критическую область $(V_{кр}; +\infty)$. По выборочным данным вычисляется наблюдаемое значение статистики $V_{набл}$. Если $V_{набл} > V_{кр}$, то гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ отвергается, в противном случае она принимается.

Если гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ подтверждается, то можно приступать непосредственно к процедуре однофакторного дисперсионного анализа, т. е. к проверке гипотезы $H_0 : a_1 = a_2 = \ldots = a_k$.

Общая сумма квадратов отклонений (общая вариация) отдельных

наблюдений y_{ij} от общей средней $\bar{y}$, вызванная влиянием на Y фактора A и случайных неучтенных (остаточных) факторов, вычисляется по формуле $Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$. Сумма Q разлагается на сумму $Q = Q_ф + Q_o$, где $Q_ф = \sum_{j=1}^{k}(\bar{y}_j - \bar{y})^2 n_j$ – сумма квадратов отклонений между группами (межгрупповая вариация), вызванная влиянием фактора A на Y и характеризующая рассеяние групповых средних около общей средней; $Q_o = \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \bar{y}_j)^2$ – сумма квадратов отклонений внутри групп (остаточной вариации), вызванная влиянием на Y остаточных факторов и характеризующая рассеяние отдельных наблюдений группы около ее средней.

По выборке значений результативного признака Y находятся три несмещенные оценки $S^2 = \frac{1}{n-1}Q$, $S_ф^2 = \frac{Q_ф}{k-1}$, $S_o^2 = \frac{Q_o}{n-k}$ дисперсии σ^2 случайной величины ξ, причем S_o^2 всегда является несмещенной оценкой, а S^2, $S_ф^2$ – только при выполнении гипотезы $H_0 : a_1 = a_2 = \ldots = a_k$, т. е. только в том случае, когда фактор A не влияет на результативный признак Y.

Проверка нулевой гипотезы H_0 основывается на сравнении оценок $S_ф^2$, S_o^2. Для этого используется F-статистика с $\nu_1 = k - 1$, $\nu_2 = n - k$ степенями свободы $F = \frac{S_ф^2}{S_o^2} = \frac{\frac{1}{k-1}\sum_{j=1}^{k}(\bar{y}_j - \bar{y})^2 n_j}{\frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}$.

По статистическим данным вычисляется наблюдаемое значение статистики $F_{набл}$. Для заданного уровня значимости α по F-распределению Фишера находится критическая точка статистики $F_{кр} = F(\alpha; k - 1; n - k)$ и строится правосторонняя критическая область $(F_{кр}; +\infty)$. Если $F_{набл} < F_{кр}$, то нулевая гипотеза не отвергается, и в этом случае

говорят, что влияние фактора *A* на признак *Y* не подтвердилось выборочными наблюдениями. Если в процессе анализа выявлено влияние фактора *A* на результативный признак *Y*, то степень данного влияния измеряется с помощью *выборочного коэффициента детерминации* $R^2 = \frac{(\sigma_в^2)_{фактора}}{(\sigma_в^2)_{признака}}$, показывающего, какая доля вариации *Q* объясняется зависимостью результативного признака *Y* от влияющего фактора *A*.

При работе в *Excel* с целью анализа данных с помощью критерия Бартлетта используются статистические функции ДИСП(число1; число2;…), ХИ2ОБР(вероятность;степени_свободы).

Для проведения однофакторного анализа используется *однофакторный дисперсионный анализ* из *Анализа данных* меню *Сервис*. После заполнения одноименного диалогового окна результат анализа появляется в виде двух таблиц. Формулы, по которым выполняются расчеты в *Excel*, представлены в таблицах 14 и 15 соответственно.

Таблица 14. **Однофакторный дисперсионный анализ: дисперсии групп**

ИТОГИ				
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	n_1	$\sum_{i=1}^{n_1} y_{i1}$	$\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_{i1}}{n_1}$	$S_1^2 = \frac{\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2}{n_1 - 1}$
Столбец 2	n_2	$\sum_{i=1}^{n_2} y_{i2}$	$\bar{y}_2 = \frac{\sum_{i=1}^{n_2} y_{i2}}{n_2}$	$S_2^2 = \frac{\sum_{i=1}^{n_2} (y_{i2} - \bar{y}_2)^2}{n_2 - 1}$
…	…	…	…	…
Столбец k	n_k	$\sum_{i=1}^{n_k} y_{ik}$	$\bar{y}_k = \frac{\sum_{i=1}^{n_k} y_{ik}}{n_k}$	$S_k^2 = \frac{\sum_{i=1}^{n_k} (y_{i2} - \bar{y}_2)^2}{n_k - 1}$

Таблица 15. **Однофакторный дисперсионный анализ: расчет параметров для проверки гипотезы**

Дисперсионный анализ			
Источник	Между группами	Внутри групп	Итого

вариации			
SS	$Q_\phi = \sum_{j=1}^{k} (\bar{y}_j - \bar{y})^2 n_j$	$Q_o = \sum_{j=1}^{k} \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_j)^2$	$Q = Q_\phi + Q_o$
df	$k - 1$	$n - k$	
MS	$\frac{1}{k-1} Q_\phi$	$\frac{1}{n-k} Q_o$	
F	$F_{набл} = \frac{\frac{1}{k-1} Q_o}{\frac{1}{n-k} Q_\phi}$		
P-Значение	$P(F_{набл} > F_{кр})$ FРАСП(x; $k - 1$; $n - k$)		
Значимость F	$F_{кр} = F(\alpha; k - 1; n - k)$ FРАСПОБР(α; $k - 1$; $n - k$)		

Замечание. В *Excel* на экран выводится таблица, строки которой записаны в столбцах табл. 16.

3. Двухфакторный дисперсионный анализ

Дисперсионный анализ рассматривает влияние двух независимых факторов A и B на изменчивость результативного признака Y. Пусть фактор A имеет I уровней A_1, A_2, …, A_I, фактор B – J уровней B_1, B_2, …, B_J, а число всевозможных сочетаний уровней этих факторов – $I \cdot J$. На каждом сочетании A_i и B_j, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, имеется n выборочных значений результативного признака Y. Поэтому общее число наблюдаемых значений признака Y равно $N = I \cdot J \cdot n$. Результаты наблюдений и групповые средние (описанные ниже) могут быть представлены в виде табл. 16.

Таблица 16. **Результаты наблюдений для двухфакторного анализа**

		Уровни (группы) фактора B				Групповые средние уровней фактора А
		B_1	B_2	…	B_J	
(группы) фактор	A_1	y_{111}	y_{121}	…	y_{1J1}	$\bar{y}_{A_1}$

		…	…	…	…	
		y_{11n}	y_{12n}	…	y_{1Jn}	
	A_2	y_{211}	y_{221}	…	y_{2J1}	$\bar{y}_{A_2}$
		…	…	…	…	
		y_{21n}	y_{22n}	…	y_{2Jn}	
	…	…	…	…	…	…
	A_I	y_{I11}	y_{I21}	…	y_{IJ1}	$\bar{y}_{A_I}$
		…	…	…	…	
		y_{I1n}	y_{I2n}	…	y_{IJn}	
Групповые средние уровней фактора B		$\bar{y}_{B_1}$	$\bar{y}_{B_2}$	…	$\bar{y}_{B_J}$	

Пусть a_i – математическое ожидание результативного признака Y на уровне A_i, $i = 1, 2, \ldots, I$; b_j – математическое ожидание результативного признака Y на уровне B_j, $j = 1, 2, \ldots, J$; c_{ij} – математическое ожидание результативного признака Y на сочетании уровней A_i и B_j, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$. Если при изменении уровня фактора A групповые математические ожидания не изменяются, т. е. $a_1 = a_2 = \ldots a_I$, то считается, что результативный признак Y не зависит от фактора A, в противном случае такая зависимость имеется. Аналогично, если при изменении уровня фактора B сохраняется равенство $b_1 = b_2 = \ldots = b_J$, то считается, что Y не зависит от фактора B.

Если $c_{11} = c_{12} = \ldots = c_{1J} = c_{21} = \ldots = c_{2J} = \ldots = c_{I1} = \ldots = c_{IJ}$, то считается, что результативный признак Y не зависит от взаимодействия факторов A и B. Поскольку числовые значения математических ожиданий неизвестны, то возникает задача проверки следующих гипотез:

$$H_0^A : a_1 = a_2 = \ldots = a_I;$$

$$H_0^B : b_1 = b_2 = \ldots = b_J;$$

$$H_0^{AB} : \begin{array}{l} c_{11} = c_{12} = \ldots = c_{1J} = \\ = c_{21} = c_{22} = \ldots = c_{2J} = \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ = c_{I1} = c_{I2} = \ldots = c_{IJ}. \end{array}$$

Проверка данных гипотез, как и в задаче однофакторного дисперсионного анализа, возможна только при выполнении следующих

условий:

• при различных сочетаниях уровней факторов *A* и *B* наблюдения независимы и проводятся в одинаковых условиях;

• при каждом сочетании уровней A_i и B_j, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, результативный признак *Y* имеет нормальный закон распределения с постоянной для различных сочетаний генеральной дисперсией σ^2.

Источниками изменчивости признака *Y* являются факторы *A* и *B*, их взаимодействие, а также влияние неучтенных случайных (остаточных) факторов.

Общая сумма квадратов отклонений (общая вариация) отдельных наблюдений y_{ijt} от общей средней $\bar{y}$, вызванная влиянием на признак *Y* факторов *A* и *B*, а также остаточных факторов, вычисляется по формуле $Q = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n}(y_{ijt} - \bar{y})^2$. Сумма *Q* равна $Q = Q_A + Q_B + Q_o$, где $Q_A = J\,n\sum_{i=1}^{I}(\bar{y}_{A_i} - \bar{y})^2$, $Q_B = I\,n\sum_{j=1}^{J}(\bar{y}_{B_j} - \bar{y})^2$ – суммы квадратов отклонений, вызванных влиянием соответственно факторов *A* и *B* на *Y*; $Q_o = Q_{AB} + Q_z$ – сумма квадратов отклонений, вызванная влиянием на *Y* одновременного взаимодействия факторов *A* и *B*, а также остаточных факторов,

$$Q_{AB} = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij} + \bar{y} - \bar{y}_{A_i} - \bar{y}_{B_j})^2,\ Q_z = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n}(y_{ijt} - \bar{y}_{ij})^2.$$

В приведенных формулах $\bar{y} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n} y_{ijt}$ – общая средняя; $\bar{y}_{A_i} = \frac{1}{J \cdot n}\sum_{j=1}^{J}\sum_{t=1}^{n} y_{jt}$ – средние значения признака *Y* на уровнях фактора *A* (по строкам); $\bar{y}_{B_j} = \frac{1}{I \cdot n}\sum_{i=1}^{I}\sum_{t=1}^{n} y_{it}$ – средние значения признака *Y* на уровнях фактора *B* (по столбцам); $\bar{y}_{ij} = \frac{1}{n}\sum_{t=1}^{n} y_{ijt}$ – средние значения признака *Y* при различных сочетаниях уровней A_i и B_j, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$.

По выборочным данным результативного признака Y находятся четыре несмещенные оценки $S^2 = \frac{Q}{N-1}$, $S_A^2 = \frac{Q_A}{I-1}$, $S_B^2 = \frac{Q_B}{J-1}$, $S_{\text{о}} = S_{AB}^2 + S_{\text{z}}^2$ ($S_{AB}^2 = \frac{Q_{AB}}{(I-1)(J-1)}$, $S_{\text{z}}^2 = \frac{Q_{\text{z}}}{IJ(n-1)}$) – дисперсии δ^2 случайной величины ξ, причем оценка $S_{\text{о}}^2$ всегда является несмещенной оценкой, оценка S_A^2 – при выполнении гипотезы H_0^A: $a_1 = a_2 = \ldots a_k$ (когда фактор A не влияет на результативный признак Y), оценка S_B^2 – при выполнении гипотезы $H_0^B: b_1 = b_2 = \ldots = b_m$ (когда фактор B не влияет на результативный признак Y), оценка S^2 – при выполнении гипотез H_0^A и H_0^B. Оценка S_{AB}^2 характеризует взаимодействие факторов A и B. Проверка гипотезы H_0^A (H_0^B) основывается на сравнении оценок S_A^2 (S_B^2) и $S_{\text{о}}^2$. Для этого используется статистика $F^A = \frac{S_A^2}{S_{\text{о}}^2}$ $\left(F^B = \frac{S_B^2}{S_{\text{о}}^2}\right)$, имеющая F-распределение с $\nu_1^A = I-1$ ($\nu_1^B = J-1$) и $\nu_2 = (I-1)(J-1)$ степенями свободы.

По статистическим данным вычисляется наблюдаемое значение статистики $F_{набл}^A$ ($F_{набл}^B$). Для заданного уровня значимости α по F-распределению Фишера находится критическая точка статистики

$$F_{кр}^A = F(\alpha; I-1; (I-1)(J-1)) \quad \left[F_{кр}^B = F(\alpha; J-1; (I-1)(J-1))\right]$$

и строится правосторонняя критическая область $(F_{кр}^A; +\infty)$ $\left[(F_{кр}^B; +\infty)\right]$. Если $F_{набл}^A < F_{кр}^A$ ($F_{набл}^B < F_{кр}^B$), то нулевая гипотеза не отвергается, и в этом случае говорят, что влияние фактора A (B) на признак Y не подтвердилось выборочными наблюдениями.

Проверка гипотезы H_0^{AB} основывается на сравнении оценок S_{AB}^2 и S_{z}^2. Для этого используется статистика $F^B = \frac{S_{AB}^2}{S_{\text{z}}^2}$, имеющая F-распределение с $\nu_1 = (I-1)(J-1)$ и $\nu_2 = IJ(n-1)$ степенями свободы. По статистическим данным вычисляется наблюдаемое значение

статистики $F_{набл}^{AB}$. Для заданного уровня значимости α по F-распределению находится критическая точка статистики

$$F_{кр}^{AB} = F(\alpha;(I-1)(J-1);IJ(n-1))$$

и строится правосторонняя критическая область $(F_{кр}^{AB};+\infty)$. Если $F_{набл}^{AB} < F_{кр}^{AB}$, то нулевая гипотеза не отвергается, и в этом случае говорят, что взаимное влияние факторов A и B на признак Y не подтвердилось выборочными наблюдениями.

Двухфакторный дисперсионный анализ имеет две разновидности: *без повторений* ($n = 1$) и *с повторениями* ($n \geq 2$). В первом случае каждому уровню фактора соответствует только одна выборка данных, во втором – определенным уровням факторов соответствует n выборок.

При работе в *Excel* с целью проведения двухфакторного анализа без повторений (или с повторениями) используется *Двухфакторный дисперсионный анализ без повторений* (или *Двухфакторный дисперсионный анализ с повторениями*) из *Анализа данных* меню *Сервис*.

После заполнения диалогового окна *Двухфакторный дисперсионный анализ без повторений* результат анализа появляется в виде двух таблиц. Формулы, по которым выполняются расчеты в *Excel*, представлены в таблицах 17 и 18 соответственно.

Таблица 17. **Двухфакторный дисперсионный анализ без повторений: дисперсии групп**

ИТОГИ				
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Строка 1	J СЧЕТ	$\sum_{j=1}^{J} y_{1j}$ СУММ	$\bar{y}_{A_1} = \frac{\sum_{j=1}^{J} y_{1j}}{J}$ СРЗНАЧ	$S_{A_1}^2 = \frac{\sum_{j=1}^{J} (y_{1j} - \bar{y}_{A_1})^2}{J-1}$ ДИСП
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
…	…	…	…	
Строка I	J СЧЕТ	$\sum_{j=1}^{J} y_{Ij}$ СУММ	$\bar{y}_{A_I} = \frac{\sum_{j=1}^{J} y_{Ij}}{J}$ СРЗНАЧ	$S_{A_I}^2 = \frac{\sum_{j=1}^{J} (y_{Ij} - \bar{y}_{A_I})^2}{J-1}$ ДИСП

Столбец 1	I СЧЕТ	$\sum_{i=1}^{I} y_{i1}$ СУММ	$\bar{y}_{B_1} = \frac{\sum_{i=1}^{I} y_{i1}}{I}$ СРЗНАЧ	$S^2_{B_1} = \frac{\sum_{i=1}^{I} (y_{i1} - \bar{y}_{B_1})^2}{I-1}$ ДИСП
…	…	…	…	…
Столбец J	I СЧЕТ	$\sum_{i=1}^{I} y_{iJ}$ СУММ	$\bar{y}_{B_J} = \frac{\sum_{i=1}^{I} y_{iJ}}{I}$ СРЗНАЧ	$S^2_{B_J} = \frac{\sum_{i=1}^{I} (y_{iJ} - \bar{y}_{B_J})^2}{I-1}$ ДИСП

Таблица 18. **Двухфакторный дисперсионный анализ без повторений: расчет параметров для проверки гипотезы**

Дисперсионный анализ				
Источник вариации	Строки	Столбцы	Погрешность	Итого
SS	$Q_A = J\sum_{i=1}^{I}(\bar{y}_{A_i} - \bar{y})^2$	$Q_B = I\sum_{j=1}^{J}(\bar{y}_{B_j} - \bar{y})^2$	$Q_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij} + \bar{y} - \bar{y}_{A_i} - \bar{y}_{B_j})^2$	$Q = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y})^2$
df	$I-1$	$J-1$	$(I-1)(J-1)$	$IJ-1$

Окончание табл. 18

Дисперсионный анализ				
Источник вариации	Строки	Столбцы	Погрешность	Итого
MS	$\frac{Q_A}{I-1}$	$\frac{Q_B}{J-1}$	$\frac{Q_{AB}}{(I-1)(J-1)}$	
F	$F^A_{набл} = \frac{\frac{1}{I-1}Q_A}{\frac{1}{n-k}Q_{AB}}$	$F^B_{набл} = \frac{\frac{1}{I-1}Q_B}{\frac{1}{n-k}Q_{AB}}$		
P-Значение	$P(F^A_{набл} > F^A_{кр})$ FРАСП	$P(F^B_{набл} > F^B_{кр})$ FРАСП		

Значимость F	$F_{кр}^{A} = F(\alpha; I-1; (I-1)(J-1))$ FРАСПОБР	$F_{кр}^{B} = F(\alpha; J-1; (I-1)(J-1))$ FРАСПОБР		

Замечание. В *Excel* на экран выводится таблица, строки которой записаны в столбцах табл. 18.

После заполнения диалогового окна *Двухфакторный дисперсионный анализ с повторениями* результат анализа появляется в виде двух таблиц. Формулы, по которым выполняются расчеты в *Excel*, представлены в таблицах 19 и 20 соответственно.

Таблица 19. **Двухфакторный дисперсионный анализ с повторениями: расчет параметров для проверки гипотезы**

ИТОГИ	Уровень B_1	…	Уровень B_J	Итого
Уровень A_1				
Счет	n	…	n	Jn
Сумма	$\sum_{t=1}^{n} y_{11t}$	…	$\sum_{t=1}^{n} y_{1Jt}$	$\sum_{j=1}^{J}\sum_{t=1}^{n} y_{1Jt}$
Среднее	$\bar{y}_{11} = \frac{1}{n}\sum_{t=1}^{n} y_{11t}$	…	$\bar{y}_{1J} = \frac{1}{n}\sum_{t=1}^{n} y_{1Jt}$	$\bar{y}_{A_1} = \frac{1}{Jn}\sum_{j=1}^{J}\sum_{t=1}^{n} y_{1Jt}$

Окончание табл. 19

ИТОГИ	Уровень B_1	…	Уровень B_J	Итого
Дисперсия	$\frac{\sum_{t=1}^{n}(y_{11t} - \bar{y}_{11})^2}{n}$	…	$\frac{\sum_{t=1}^{n}(y_{1Jt} - \bar{y}_{1J})^2}{n}$	$S_{A_1}^2 = \frac{\sum_{j=1}^{J}\sum_{t=1}^{n}(y_{1Jt} - \bar{y}_{A_1})^2}{Jn-1}$
…	…	…	…	…
Уровень A_I				
Счет	n	…	n	Jn
Сумма	$\sum_{t=1}^{n} y_{I1t}$	…	$\sum_{t=1}^{n} y_{IJt}$	$\sum_{j=1}^{J}\sum_{t=1}^{n} y_{IJt}$
Среднее	$\bar{y}_{I1} = \frac{1}{n}\sum_{t=1}^{n} y_{I1t}$	…	$\bar{y}_{IJ} = \frac{1}{n}\sum_{t=1}^{n} y_{IJt}$	$\bar{y}_{A_I} = \frac{1}{Jn}\sum_{j=1}^{J}\sum_{t=1}^{n} y_{IJt}$

Дисперсия	$\frac{\sum_{t=1}^{n}(y_{I1t}-\bar{y}_{11})^2}{n}$	…	$\frac{\sum_{t=1}^{n}(y_{IJt}-\bar{y}_{1J})^2}{n}$	$S_{A_I}^2=\frac{\sum_{j=1}^{J}\sum_{t=1}^{n}(y_{1Jt}-\bar{y}_{A_1})^2}{Jn-1}$
Итого				
Счет	In	…	In	
Сумма	$\sum_{i=1}^{I}\sum_{t=1}^{n}y_{i1t}$	…	$\sum_{i=1}^{I}\sum_{t=1}^{n}y_{iJt}$	
Среднее	$\bar{y}_{B_1}=\frac{\sum_{i=1}^{I}\sum_{t=1}^{n}y_{i1t}}{I\,n}$	…	$\bar{y}_{B_J}=\frac{\sum_{i=1}^{I}\sum_{t=1}^{n}y_{iJt}}{I\,n}$	
Дисперсия	$S_{B_1}^2=$ $=\frac{\sum_{i=1}^{I}\sum_{t=1}^{n}(y_{i1t}-\bar{y}_{B_1})^2}{I\,n-1}$	…	$S_{B_J}^2=$ $=\frac{\sum_{i=1}^{I}\sum_{t=1}^{n}(y_{iJt}-\bar{y}_{B_J})^2}{I\,n-1}$	

Вопросы для самоконтроля

1. В чем суть дисперсионного анализа?
2. В каком случае используется однофакторный дисперсионный анализ?
3. Каким условиям должны удовлетворять группы наблюдений фактора, для того чтобы применить однофакторный дисперсионный анализ?
4. С помощью какого критерия проверяется гипотеза о равенстве дисперсий нескольких выборок?
5. Каким образом можно представить сумму квадратов отклонений отдельных наблюдений от общей суммы?
6. Что характеризует межгрупповая вариация?

7. Что характеризует внутригрупповая вариация?

8. Какая статистика используется для проверки гипотезы о равенстве математических ожиданий групп фактора?

9. В чем разница между однофакторным и двухфакторным анализами?

10. Каким условиям должны удовлетворять группы наблюдений факторов, для того чтобы применить двухфакторный дисперсионный анализ?

11. Какая гипотеза проверяется в двухфакторном дисперсионном анализе?

12. Какие разновидности имеет двухфакторный дисперсионный анализ?

4. Лабораторная работа

Задание 1

Результаты наблюдений за расходом сырья при производстве одинаковой продукции по одной и той же технологии на пяти различных заводах равных мощностей представлены в таблице *MS Excel* на рис. 29 (считается, что расход сырья является нормально распределенной случайной величиной).

	A	B	C	D	E	F
1		Расход сырья				
2	Месяцы	Завод 1	Завод 2	Завод 3	Завод 4	Завод 5
3	1	114	112	132	124	124
4	2	124	119	124	114	116
5	3	110	124	129	119	119
6	4	116	116	129	124	119
7	5	119	116	129	116	132
8	6	119	124	124	116	129
9	7	129	112	114	129	116
10	8	124	119	119	124	119
11	9	110	119	124	114	
12	10	124	112		116	
13	11	119			129	
14	12	124				

Рис. 29. **Данные для однофакторного анализа**

При уровне значимости $\alpha = 0{,}05$ требуется выяснить, зависит ли расход сырья от того, на каком заводе произведена продукция. Необходимо оценить степень этой зависимости.

Задание 2

Данные о разрывной нагрузке пряжи, изготовленной на разных станках и из отличающегося некоторым образом друг от друга сырья, представлены в таблице *MS Excel* на рис. 30.

	A	B	C
1	*Тип станка*	*Вид сырья*	
2		*шелк натуральный*	*шелк искусственный*
3	*JANOME*	*10*	*50*
4	*HUSQVARNA*	*20*	*60*
5	*SINGER*	*30*	*100*

Рис. 30. **Данные для двухфакторного анализа**

Качество пряжи измеряется величиной разрывной нагрузки. Требуется при уровне значимости $\alpha = 0{,}05$ выяснить, влияют ли на качество пряжи тип станка и вид сырья, из которого пряжа производится.

Задание 3

Выборочные данные об урожайности пшеницы, выращенной на участках, на которые вносились различные виды удобрений и которые подвергались различной химической обработке, приведены в таблице *MS Excel* на рис. 31.

	A	B	C	D	E	F
1	Номер участка	Вид удобрения	Способ химической обработки			
2			Способ 1	Способ 2	Способ 3	Способ 4
3	Участок 1	Удобрение 1	21,4	20,9	19,6	17,6
4	Участок 2		21,2	20,3	18,8	16,6
5	Участок 3		20,1	19,8	16,4	17,5
6	Участок 1	Удобрение 2	12	13,6	13	13,3
7	Участок 2		14,2	13,3	13,7	14
8	Участок 3		12,1	11,6	12	13,9
9	Участок 1	Удобрение 3	13,5	14	12,9	12,4
10	Участок 2		11,9	15,6	12,9	13,7
11	Участок 3		13,4	13,8	12,1	13
12	Участок 1	Удобрение 4	12,8	14,1	14,2	12
13	Участок 2		13,8	13,2	13,6	14,6
14	Участок 3		13,7	15,3	13,3	14

Рис. 31. **Данные для двухфакторного анализа с повторениями**

Требуется при уровне значимости $\alpha = 0{,}05$ выяснить, влияют ли на урожайность пшеницы способы химической обработки почвы.

Порядок выполнения работы

Задание 1

По условию задачи результирующим признаком *F* является расход сырья, который зависит от влияющего фактора *A* (завода), представленного пятью группами. Решение задачи выполняется в следующие два этапа:

1. Проверка гипотезы о равенстве групповых дисперсий с помощью критерия Бартлетта. Пусть нулевая гипотеза имеет вид $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_5^2$, где σ_i^2 – дисперсия *i*-й группы фактора *A*, $i = 1, 2, \ldots, 5$. Показатели, рассчитанные в ходе проверки данной гипотезы, представлены в таблице на рис. 32.

Содержимое ячеек таблицы заполняется следующим образом:

- в ячейках B2:F2 находятся объемы выборок n_j, вычисленные с помощью функции СЧЕТ для каждого завода (СЧЕТ(B3:B14) и т. д.);
- в ячейках B3:F3 находятся несмещенные оценки S_j^2, вычисленные с помощью функции ДИСП для каждого завода (ДИСП(B3:B14) и т. д.);
- в ячейки B4:F4 вводится формула массива

{=СУММПРОИЗВ(B2:F2-1;B3:F3)/СУММ(B2:F2-1)};

- в ячейки B5:F5 вводится формула массива
  {=1+1/(3*(5-1))*(СУММ(1/(B2:F2-1))-1/(СУММ(B2:F2)-5))};
- в ячейки B6:F6 вводится формула массива
  {=1/B5*СУММПРОИЗВ(B2:F2-1;LN(B4/B3:F3))};
- ячейка B7 содержит формулу =ХИ2ОБР(0,05;5-1).

	A	B	C	D	E	F
1		Завод 1	Завод 2	Завод 3	Завод 4	Завод 5
2	Число наблюдений	12	10	9	11	8
3	Оценки групповых дисперсий S_j^2	35,87879	20,67778	32,11111	32,87273	35,92857
4	Значение $\overline{S^2}$	31,50851	31,50851	31,50851	31,50851	31,50851
5	Значение C	1,045638	1,045638	1,045638	1,045638	1,045638
6	Наблюдаемое значение статистики $V_{набл}$	0,829796	0,829796	0,829796	0,829796	0,829796
7	Критическая точка $V_{кр}$	9,487728				

Рис. 32. **Результаты проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_5^2$**

Так как $V_{набл}$ = 0,829796 не попадает в критическую область (9,487728; +∞), то гипотеза о равенстве дисперсий групп принимается, поэтому можно приступить к проверке гипотезы о равенстве математических ожиданий.

2. Однофакторный анализ – проверка нулевой гипотезы

$$H_0 : a_1 = a_2 = \ldots = a_5$$

(о равенстве средних значений объемов расходов пяти заводов).

Для этого используется режим анализа *Однофакторный дисперсионный анализ.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 33):

- *Входной интервал* (вводятся ссылки на ячейки B2:F14, в которых находятся наблюдаемые значения признака *F* и названия уровней фактора);
- *Группирование* (проставляется автоматически);
- *Метки* (устанавливается флажок);
- *Альфа* (вводится уровень значимости 0,05);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист Однофакт-анализ*).

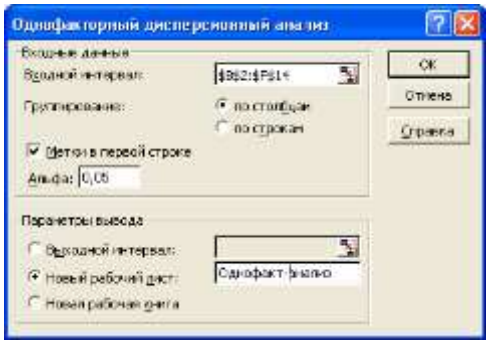


Рис. 33. **Диалоговое окно *Однофакторный дисперсионный анализ***

Показатели, рассчитанные в ходе анализа, представлены в виде двух таблиц «ИТОГИ» и «Дисперсионный анализ» на рис. 34.

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	Завод 1	12	1432	119,3333	35,8787879		
6	Завод 2	10	1173	117,3	20,6777778		
7	Завод 3	9	1124	124,8889	32,1111111		
8	Завод 4	11	1325	120,4545	32,8727273		
9	Завод 5	8	974	121,75	35,9285714		
10							
11							
12	Дисперсионный анализ						
13	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
14	Между группами	304,437	4	76,10929	2,41551966	0,0625637	2,57873724
15	Внутри групп	1417,88	45	31,50851			
16							
17	Итого	1722,32	49				

Рис. 34. **Результат *Однофакторного дисперсионного анализа***

Вывод. Вычисленное значение $F_{набл}$ = 2,4155 не попадает в критическую область (2,5787; ∞), поэтому нулевая гипотеза о равенстве математических ожиданий групп принимается. Следовательно, расход сырья статистически не зависит от завода, производящего продукцию. Для *P*-значения имеет место неравенство 0,05 < 0,06, что также говорит о хорошем согласии гипотезы $H_0 : a_1 = a_2 = ... = a_5$ с выборочными данными. С другой стороны, *P*-значение находится близко к 0,05, что вызывает сомнения в истинности гипотезы.

Выборочный коэффициент детерминации

$$R^2 = \frac{\sigma_A^2}{\sigma_F^2} = \frac{SS_{между\ группами} / n}{SS_{итого} / n} = \frac{304,437 / 50}{1722,32 / 50} = 0,1768$$

показывает, что только 17% общей выборочной вариации расхода сырья связано с выбором завода.

Задание 2

Результирующим признаком F является величина разрывной нагрузки пряжи, факторами – тип станка (A) и вид сырья (B) (см. рис. 30). Необходимо проверить следующие гипотезы:

- о равенстве математических ожиданий групп фактора A

$$H_0^A : a_1 = a_2 = a_3;$$

- о равенстве математических ожиданий групп фактора B

$$H_0^B : b_1 = b_2;$$

- о равенстве математических ожиданий взаимодействия факторов

$$H_0^{AB} : c_{11} = c_{12} = c_{21} = c_{22} = c_{31} = c_{32}.$$

Для решения данной задачи используется режим анализа *Двухфакторный дисперсионный анализ без повторений.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 35):

- *Входной интервал* (вводятся ссылки на ячейки A2:C5, в которых находятся значения наблюдаемого признака F и названия уровней факторов);
- *Метки* (устанавливается флажок);
- *Альфа* (вводится уровень значимости 0,05);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист Двухфакт-без-повторений*).

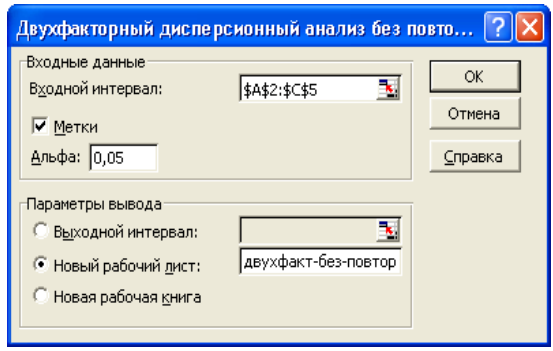


Рис. 35. **Диалоговое окно *Двухфакторный дисперсионный анализ без повторений***

Показатели, рассчитанные в ходе анализа, представлены в виде двух таблиц «ИТОГИ» и «Дисперсионный анализ» на рис. 36.

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ без повторений						
2							
3	ИТОГИ	Счет	Сумма	Среднее	Дисперсия		
4	JANOME	2	60	30	800		
5	HUSQVARNA	2	80	40	800		
6	SINGER	2	130	65	2450		
7							
8	шелк натуральный	3	60	20	100		
9	шелк искусственный	3	210	70	700		
10							
11							
12	Дисперсионный анализ						
13	Источник вариации	SS	df	MS	F	P-Значение	F критическое
14	Строки	1300	2	650	4,33333333	0,1875	19,00002644
15	Столбцы	3750	1	3750	25	0,037749651	18,51276466
16	Погрешность	300	2	150			
17							
18	Итого	5350	5				

Рис. 36. **Результат *Двухфакторного дисперсионного анализа без повторений***

Вывод. Вычисленное значение *F*-критерия фактора *A* (тип станка) $F^{A}_{набл}$ = 4,3333 не попадает в критическую область (19,00; +∞), образованную правосторонним интервалом, поэтому гипотеза

$$H_0^A : a_1 = a_2 = a_3$$

принимается, т. е. считается, что влияние типа изготавливающего станка на качество пряжи не подтвердилось. Для *P*-значения имеет место неравенство 0,1875 > 0,05, что говорит о хорошем согласии ги-

потезы H_0^A с выборочными данными.

Вычисленное значение *F*-критерия фактора *B* (вид сырья) $F_{набл}^B = 25$ попадает в критическую область $(18{,}5128; +\infty)$, образованную правосторонним интервалом, поэтому гипотеза $H_0^B : b_1 = b_2$ отвергается, т. е. считается, что вид сырья влияет на качество пряжи. Так как для *P*-значения имеет место неравенство 0,038 < 0,05, то гипотеза H_0^B также не принимается.

Выборочный коэффициент детерминации

$$R^2 = \frac{\sigma_B^2}{\sigma_F^2} = \frac{SS_{столбцы} / n}{SS_{итого} / n} = \frac{3750 / 6}{5350 / 6} = 0{,}7009$$

показывает, что на 70% всей выборочной вариации качества пряжи влияет вид сырья. Влиянию типа станка подвержено только 24% выборочной вариации

$$R^2 = \frac{\sigma_A^2}{\sigma_F^2} = \frac{SS_{строки} / n}{SS_{итого} / n} = \frac{1300 / 6}{5350 / 6} = 0{,}243.$$

Задание 3

Результирующим признаком *F* является урожайность пшеницы, факторами – вид удобрений (*A*) и способ химической обработки (*B*) (см. рис. 31).

Необходимо проверить следующие гипотезы:

- о равенстве математических ожиданий групп фактора *A*

$$H_0^A : a_1 = a_2 = a_3 = a_4;$$

- о равенстве математических ожиданий групп фактора *B*

$$H_0^B : b_1 = b_2 = b_3 = b_4;$$

- о равенстве математических ожиданий взаимодействия факторов

$$H_0^{AB} : c_{11} = c_{12} = c_{13} = c_{14} = ... = c_{44}.$$

Для решения данной задачи используется режим анализа *Двухфакторный дисперсионный анализ с повторениями.* Значения параметров в одноименном диалоговом окне устанавливаются следующим образом (рис. 37):

- *Входной интервал* (вводятся ссылки на ячейки B2:F14, в которых

находятся значения наблюдаемого признака F и названия уровней факторов);

• *Число строк для выборки* (вводится число наблюдений для каждой пары групп факторов A и B, т. е. число 3);

• *Альфа* (вводится уровень значимости 0,05);

• *Параметры вывода* (вводится ссылка на *Новый рабочий лист Двухфакт-без-повторений*).

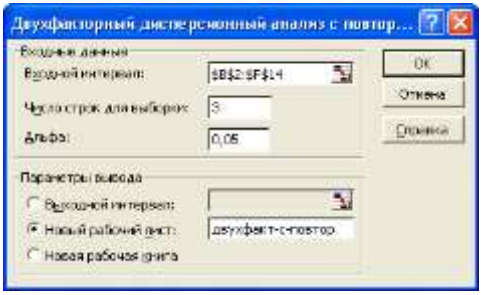


Рис. 37 **Диалоговое окно *Двухфакторный дисперсионный анализ с повторениями***

Показатели, рассчитанные в ходе анализа, представлены в виде двух таблиц «ИТОГИ» и «Дисперсионный анализ» на рис. 38.

Вывод. Вычисленное значение F-критерия фактора A (вид удобрений) $F^{A}_{набл}$ = 123,64 попадает в критическую область $(2{,}9011; +\infty)$, поэтому гипотеза $H^{A}_{0}: a_1 = a_2 = a_3 = a_4$ отвергается, т. е. считается, что вид используемого удобрения влияет на урожайность пшеницы. Для P-значения имеет место неравенство $1{,}1 \cdot 10^{-17} < 0{,}01$, поэтому гипотеза H^{A}_{0} не принимается.

Выборочный коэффициент детерминации

$$R^2 = \frac{\sigma^2_A}{\sigma^2_F} = \frac{SS_{строки} / n}{SS_{итого} / n} = \frac{309{,}259 / 48}{371{,}5881 / 48} = 0{,}8323$$

показывает, что на 83% всей выборочной вариации урожайности пшеницы влияет вид удобрения.

Вычисленное значение F-критерия фактора B (способ химической

обработки) $F_{набл}^{B}$ = 3,9862 попадает в критическую правостороннюю область $(2{,}1888; +\infty)$, поэтому гипотеза $H_0^B : b_1 = b_2 = b_3 = b_4$ отвергается, т. е. считается, что способ химической обработки также влияет на урожайность пшеницы. Так как для *P*-значения имеет место неравенство 0,016 < 0,05, то гипотеза H_0^B также не принимается.

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ с повторениями						
2							
3	ИТОГИ	Способ 1	Способ 2	Способ 3	Способ 4	Итого	
4	*Удобрение 1*						
5	Счет	3	3	3	3	12	
6	Сумма	62,7	61	54,8	51,7	230,2	
7	Среднее	20,9	20,33333	18,26667	17,23333	19,18333333	
8	Дисперсия	0,49	0,303333	2,773333	0,303333	3,134242424	
9							
10	*Удобрение 2*						
11	Счет	3	3	3	3	12	
12	Сумма	38,3	38,5	38,7	41,2	156,7	
13	Среднее	12,76667	12,83333	12,9	13,73333	13,05833333	
14	Дисперсия	1,543333	1,163333	0,73	0,143333	0,819015152	
15							
16	*Удобрение 3*						
17	Счет	3	3	3	3	12	
18	Сумма	38,8	43,4	37,9	39,1	159,2	
19	Среднее	12,93333	14,46667	12,63333	13,03333	13,26666667	
20	Дисперсия	0,803333	0,973333	0,213333	0,423333	0,986060606	
21							
22	*Удобрение 4*						
23	Счет	3	3	3	3	12	
24	Сумма	40,3	42,6	41,1	40,6	164,6	
25	Среднее	13,43333	14,2	13,7	13,53333	13,71666667	
26	Дисперсия	0,303333	1,11	0,21	1,053333	0,726969697	
27							
28	*Итого*						
29	Счет	12	12	12	12		
30	Сумма	180,1	185,5	172,5	172,6		
31	Среднее	15,00833	15,45833	14,375	14,38333		
32	Дисперсия	13,25902	9,706288	6,389318	3,519697		
33							
34							
35	Дисперсионный анализ						
36	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
37	Выборка	309,259	3	103,0863	123,6418	1,11281E-17	2,901117568
38	Столбцы	9,970625	3	3,323542	3,986257	0,016084218	2,901117568
39	Взаимодействие	25,67854	9	2,853171	3,422096	0,004729788	2,188762949
40	Внутри	26,68	32	0,83375			
41							
42	Итого	371,5881	47				

Рис. 38. **Результат *Двухфакторного дисперсионного анализа с повторениями***

Выборочный коэффициент детерминации для фактора *B*

$$R^2 = \frac{\sigma_B^2}{\sigma_F^2} = \frac{SS_{столбцы} / n}{SS_{итого} / n} = \frac{9{,}9706 / 48}{371{,}59 / 48} = 0{,}0268$$

показывает, что только 2,68% всей выборочной вариации урожайно-

сти пшеницы связано с влиянием способа химической обработки.

Значение *F*-критерия взаимодействия факторов $F_{набл}^{AB}$ = 3,4221 попадает в критическую правостороннюю область $(2{,}1888; +\infty)$. Значит, эффективность различных видов удобрений изменяется в зависимости от способа химической обработки почвы. Так как для *P*-значения имеет место неравенство 0,005 < 0,05, то гипотеза H_0^B отвергается.

Влиянию взаимодействия факторов подвержено только около 7% выборочной вариации

$$R^2 = \frac{\sigma_{AB}^2}{\sigma_F^2} = \frac{SS_{взаимодейс\,твие} / n}{SS_{итого} / n} = \frac{25{,}6785 / 48}{371{,}59 / 48} = 0{,}069105.$$

Тема 6. ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗИ МЕЖДУ ПРИЗНАКАМИ

План

1. Ковариация и корреляция.
2. Оценка тесноты связи.
3. Регрессия.
4. Лабораторная работа.

1. Ковариация и корреляция

В естественных науках важной задачей является анализ зависимостей между изучаемыми величинами. При обработке и использовании статистических данных с целью получения как научных, так и практических выводов, вызывает интерес, как изменяется один признак при изменении другого. Эти зависимости могут быть функциональными, стохастическими. В *функциональных* зависимостях каждому значению одной переменной соответствует единственное значение другой, т. е. результативный признак полностью определяется факторным. *Стохастической* называется зависимость, при которой изменение одной из величин влечет изменение распределения другой.

Факторный признак в таких связях не полностью определяет результативный признак, а влияет только на среднее значение, при этом отдельные результаты могут противоречить установленной связи.

В статистике изучаются выборочные данные случайных величин, поэтому стохастическая зависимость называется *статистической*, или *корреляционной.*

Задачами корреляционного анализа являются следующие:

- измерение степени связи;
- отбор факторов, оказывающих наибольшее влияние на результативный признак на основании степени связности между признаками;
- обнаружение неизвестных причинных связей.

Для оценки тесноты и вида связи между случайными величинами используются показатели ковариации и корреляции.

Пусть $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$ – выборка из генеральной совокупности двумерной случайной величиной (ξ; η), описывающей случайный эксперимент.

Выборочной ковариацией $\mathrm{cov}(X;Y)$ называется среднее произведений отклонений значений выборок $X = \{x_1, x_2, ..., x_n\}$ и $Y = \{y_1, y_2, \ldots, y_n\}$ от их средних $\bar{x}$, $\bar{y}$: $\mathrm{cov}(X;Y) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Ковариация характеризует рассеивание значений выборок *X* и *Y*, а также линейную связь между ними.

Выборочной ковариационной матрицей называется матрица вида

$$\begin{pmatrix} \mathrm{cov}(X;X) & \mathrm{cov}(X;Y) \\ \mathrm{cov}(Y;X) & \mathrm{cov}(Y;Y) \end{pmatrix}.$$

Очевидно, что $\mathrm{cov}(X;X) = \sigma_X^2$, $\mathrm{cov}(Y;Y) = \sigma_Y^2$.

Выборочная ковариационная матрица устанавливает взаимосвязь между выборками *X* и *Y* из генеральных совокупностей значений случайных величин ξ и η по величине ковариации следующим образом:

- при $\mathrm{cov}(X,Y) > 0$ большим значениям выборки *X* соответствуют большие значения выборки *Y*;
- при $\mathrm{cov}(X,Y) < 0$ большим значениям выборки *X* соответствуют меньшие значения выборки *Y* (или наоборот);

• при $\mathrm{cov}(X,Y)\to 0$ данные выборок X и Y не связаны.

Выборочным коэффициентом корреляции r_{XY} случайных величин ξ и η, между которыми предполагается линейная корреляционная связь, называется величина, определяемая по формуле

$$r_{XY}=\frac{\mathrm{cov}(X;Y)}{\sigma_X\sigma_Y}.$$

Качественная оценка тесноты связи между величинами выявляется по шкале Чеддока (табл. 21).

Таблица 21. **Шкала Чеддока**

Теснота связи	Значение коэффициента корреляции при наличии	
	прямой связи	обратной связи
Слабая	0,1–0,3	(–0,1)–(–0,3)
Умеренная	0,3–0,5	(–0,3)–(–0,5)
Заметная	0,5–0,7	(–0,5)–(–0,7)
Высокая	0,7–0,9	(–0,7)–(–0,9)
Весьма высокая	0,9–0,99	(–0,9)–(–0,99)

Выборочной корреляционной матрицей называется матрица вида

$$\begin{pmatrix} 1 & r_{XY} \\ r_{YX} & 1 \end{pmatrix}.$$

Выборочная корреляционная матрица также устанавливает взаимосвязь наборов выборочных данных по величине коэффициента корреляции следующим образом:

• при $0<r_{XY}<1$ большим значениям выборки X соответствуют большие значения выборки Y;

• при $-1<r_{XY}<0$ большим значениям выборки X соответствуют меньшие значения выборки Y (или наоборот);

• при $r_{XY}=0$ данные двух диапазонов некоррелированы;

• при $|r_{XY}|=1$ существует линейная функциональная зависимость между выборочными значениями X и Y.

Для нахождения выборочного коэффициента корреляции исполь-

зуются следующие формулы:

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

При исследовании связи между несколькими случайными величинами находятся выборочные коэффициенты ковариации и корреляции между парами всех исследуемых величин и строятся соответствующие ковариационные и корреляционные матрицы. Например, ковариационная матрица для трех выборок *X*, *Y*, Z имеет следующий вид:

$$K(X;Y;Z) = \begin{pmatrix} \operatorname{cov}(X;X) & \operatorname{cov}(X;Y) & \operatorname{cov}(X;Z) \\ \operatorname{cov}(Y;X) & \operatorname{cov}(Y;Y) & \operatorname{cov}(Y;Z) \\ \operatorname{cov}(Z;X) & \operatorname{cov}(Z;Y) & \operatorname{cov}(Z;Z) \end{pmatrix}.$$

При работе в *Excel* с целью нахождения ковариационной матрицы используется режим *Ковариация* из *Анализа данных* меню *Сервис*. Результат анализа представлен в виде таблицы. Формулы и функции *Excel*, по которым рассчитываются соответствующие числовые характеристики, приводятся в табл. 22.

Таблица 22. **Ковариационная матрица**

	Столбец 1	*Столбец 2*
Столбец 1	$\operatorname{cov}(X,X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ КОВАР	
Столбец 2	$\operatorname{cov}(Y,X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ КОВАР	$\operatorname{cov}(Y;Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ КОВАР

Для нахождения корреляционной матрицы используется режим *Корреляция* из *Анализа данных* меню *Сервис*. Результат анализа представлен в виде таблицы. Формулы и функции *Excel*, по которым рассчитываются соответствующие числовые характеристики, приводятся в табл. 23.

Таблица 23. **Корреляционная матрица**

	Столбец 1	*Столбец 2*
Столбец 1	$r_{XX} = \frac{\operatorname{cov}(X, X)}{\sigma_X \sigma_Y} = 1$ КОРРЕЛ	
Столбец 2	$r_{YX} = \frac{\operatorname{cov}(Y, X)}{\sigma_X \sigma_Y}$ КОРРЕЛ	$r_{YY} = \frac{\operatorname{cov}(Y, Y)}{\sigma_Y \sigma_Y} = 1$ КОРРЕЛ

2. Оценка тесноты связи

Пусть дана выборка $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$ из генеральной совокупности двумерной случайной величины (ξ; η). Оценка тесноты связи между случайными величинами ξ и η на основе выборочных данных проводится согласно следующему алгоритму:

1. Рассчитывается выборочный коэффициент корреляции r_{XY}.

2. Проверяется значимость (существенность) коэффициента корреляции, т. е. существенно ли r_{XY} отличается от нуля или это отличие можно приписать влиянию случайности, связанной с выборкой. Для этого выдвигается нулевая гипотеза о равенстве нулю коэффициента корреляции двумерной случайной величины (ξ; η) $H_0 : r_{\xi\eta} = 0$ при альтернативной гипотезе $H_1 : r_{\xi\eta} \neq 0$.

При проверке нулевой гипотезы используется *t*-статистика $t = \frac{|r_{XY}|}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}$, имеющая распределение Стьюдента с $(n - 2)$ степенями свободы. По выборке находится наблюдаемое значение статистики $t_{набл}$. Для заданного уровня значимости α по распределению Стьюдента определяется критическая точка $t_{кр} = t(\alpha; n - 2)$. Если $t_{набл} \geq t_{кр}$, то нулевая гипотеза об отсутствии корреляционной зависимости случайных величин ξ и η отвергается, т. е. линейный коэффициент корреляции значим, существует статистическая зависимость между случайными величинами ξ и η.

3. Для значимого коэффициента корреляции $r_{\xi\eta}$ доверительный

интервал при уровне значимости α имеет следующий вид:

$$r_{XY} - t(\alpha; n-2)\frac{1-r_{XY}}{\sqrt{n}} \le r_{\xi\eta} \le r_{XY} + t(\alpha; n-2)\frac{1-r_{XY}}{\sqrt{n}}.$$

4. Рассчитывается стандартная ошибка выборочного коэффициента корреляции по формуле $\sigma_{1-\alpha} = \sqrt{\frac{1-r_{XY}^2}{n-2}}$.

При работе в *Excel* с целью вычисления выборочного коэффициента корреляции используется статистическая функция

КОРРЕЛ(массив1;массив2).

Для вычисления критического значения *t*-статистики при построении доверительного интервала для коэффициента корреляции используется функция *Excel*

СТЬЮДРАСПОБР(вероятность;степени_свободы).

3. Регрессия

Наряду с корреляционным анализом проводится *регрессионный анализ*, который заключается в определении формы связи зависимой случайной величины η с независимыми случайными величинами ξ_1, ξ_2, …, ξ_m.

Форма связи результативного признака η с факторами ξ_1, ξ_2, …, ξ_m называется *уравнением регрессии*. В зависимости от типа выбранного уравнения различают *линейную* и *нелинейную регрессии* (квадратичная, экспоненциальная, логарифмическая и т. д.).

Исходя из числа взаимосвязанных признаков различают *парную* и *множественную регрессии*. Если исследуется связь между двумя признаками (результативным и факторным), то регрессия называется *парной*, если между тремя и более признаками – *множественной* (*многофакторной*) *регрессией*.

При изучении регрессии следует придерживаться определенной последовательности этапов, приведенных ниже.

Этап 1. Установление формы зависимости. Пусть в результате наблюдений двумерной случайной величины (ξ; η) получены данные, представляющие собой совокупность точек $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$. Графическое изображение этих точек в плоскости *Oxy* пред-

ставляет собой корреляционное поле (диаграмму рассеяния). Диаграмма рассеяния позволяет произвести визуальный анализ эмпирических данных и графически определить вид функции регрессии $\eta = \varphi(\xi)$. При $n > 2$ диаграмму рассеяния случайного вектора $(\xi_1, \xi_2, \ldots, \xi_m, \eta)$ достаточно сложно изобразить графически. В этом случае регрессионная зависимость имеет вид $\eta = \varphi(\xi_1, \xi_2, \ldots, \xi_m)$.

Этап 2. Определение вида уравнения регрессии и его параметров (*коэффициентов*). Пусть результативный признак η линейно зависит от факторов $\xi_1, \xi_2, \ldots, \xi_m$. В общем виде теоретическая линейная регрессия представляется в виде уравнения

$$\eta = \alpha_0 + \alpha_1\xi_1 + \alpha_2\xi_2 + ... + \alpha_m\xi_m + \varepsilon,$$

где $a_0, a_1, \ldots, a_m$ – неизвестные коэффициенты, ε – случайные отклонения.

Для определения значений неизвестных коэффициентов необходимо знать и использовать все значения переменных $\xi_1, \xi_2, \ldots, \xi_m$ и η генеральной совокупности, что практически невозможно. Поэтому по выборке ограниченного объема строится эмпирическое уравнение регрессии $\tilde{y} = a_0 + a_1x_1 + a_2x_2 + ... + a_mx_m$, где $\tilde{y}$ – теоретические значения результативного признака, полученные путем подстановки соответствующих значений факторных признаков в уравнение регрессии; $a_0, a_1, \ldots, a_m$ – оценки неизвестных коэффициентов уравнения регрессии.

Параметры уравнения регрессии определяются с помощью метода наименьших квадратов. Сущность данного метода заключается в нахождении параметров модели $a_0, a_1, \ldots, a_m$, при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений y_i, $i = 1, 2, \ldots, n$, результативного признака от теоретических $\tilde{y}_i = \tilde{y}(x_i)$, $i = 1, 2, \ldots, n$, полученных по следующему уравнению регрессии:

$$S = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a_0 + a_1x_{1_i} + a_2x_{2_i} + ... + a_mx_{m_i}))^2 \to \min .$$

Здесь $X_j = \{x_{j_1}, x_{j_2}, ..., x_{j_n}\}$ – выборка из генеральной совокупности случайной величины ξ_j, $j = 1, 2, \ldots, m$.

Функция является функцией переменных $a_0, a_1, \ldots, a_m$, т. е. $S = S(a_1, a_2, \ldots, a_m)$. Исследование на экстремум данной функции про-

водится методами дифференциального исчисления. После несложных математических преобразований получается система m нормальных уравнений (по числу параметров)

$$\begin{cases} na_0 + a_1\sum\limits_{i=1}^{n} x_{1_i} + a_2\sum\limits_{i=1}^{n} x_{2_i} + ... + a_m\sum\limits_{i=1}^{n} x_{m_i} = \sum\limits_{i=1}^{n} y_i, \\ a_0\sum\limits_{i=1}^{n} x_{1_i} + a_1\sum\limits_{i=1}^{n} x_{1_i}^2 + a_2\sum\limits_{i=1}^{n} x_{2_i}x_{1_i} + ... + a_m\sum\limits_{i=1}^{n} x_{m_i}x_{1_i} = \sum\limits_{i=1}^{n} y_i x_{1_i}, \\ ...., \\ a_0\sum\limits_{i=1}^{n} x_{m_i} + a_1\sum\limits_{i=1}^{n} x_{1_i}x_{m_i} + a_2\sum\limits_{i=1}^{n} x_{2_i}x_{m_i} + ... + a_m\sum\limits_{i=1}^{n} x_{m_i}^2 = \sum\limits_{i=1}^{n} y_i x_{m_i}, \end{cases}$$

из которой находятся неизвестные a_0, a_1, …, a_m, которые являются коэффициентами уравнения регрессии, где n – число наблюдений.

Этап 3. Определение степени связи результативного признака и факторов. Для определения величины степени стохастической взаимосвязи результативного признака η и факторов ξ_1, ξ_2, …, ξ_m необходимо знать следующие дисперсии:

• общую дисперсию результативного признака η, отображающую влияние как основных, так и остаточных факторов:

$$\sigma_\eta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 ,$$

где $\bar{y}$ – выборочное среднее значение результативного признака η по выборке $Y = \{y_1, y_2, \ldots, y_n\}$;

• факторную дисперсию результативного признака η, отображающую влияние только основных факторов: $\sigma_\Phi^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(\tilde{y}_i - \bar{y})^2$;

• остаточную дисперсию результативного признака η, отображающую влияние только остаточных факторов:

$$\sigma_O^2 = \frac{1}{n-(m+1)}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 .$$

При корреляционной связи результативного признака и факторов выполняется соотношение $\sigma_\Phi^2 < \sigma_\eta^2$, при этом $\sigma_\eta^2 = \sigma_\Phi^2 + \sigma_O^2$.

Этап 4. Проверка общего качества уравнения регрессии. Для анализа общего качества уравнения линейной регрессии обычно исполь-

зуется *множественный коэффициент детерминации* R^2, называемый также *квадратом коэффициента множественной корреляции*. Множественный коэффициент детерминации рассчитывается по формуле $R^2 = \frac{\sigma_{\Phi}^2}{\sigma_{\eta}^2}$ и определяет долю разброса результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную модель. Чем теснее линейная связь между признаками, тем ближе коэффициент детерминации к единице. Однако, при достаточно близком к единице коэффициенте детерминации не всегда наблюдается тесная взаимосвязь между случайными величинами. Поэтому необходимы дополнительные исследования.

В большинстве случаев уравнение регрессии строится на основе выборочных данных. Поэтому возникает вопрос о согласованности построенного уравнения генеральной совокупности случайного вектора (ξ_1, ξ_2, …, ξ_n, η) выборочным данным. Для ответа на этот вопрос выдвигается гипотеза о незначимости множественного коэффициента детерминации $H_0 : R^2 = 0$ при альтернативной гипотезе $H_1 : R^2 \neq 0$.

При проверке нулевой гипотезы используется F-статистика

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

имеющая распределение Фишера $F = F(\alpha, m, n-m-1)$, где n – число наблюдений, m – число факторов в уравнении регрессии.

По выборочным данным вычисляется наблюдаемое значение статистики $F_{набл}$. По F-распределению находится критическое значение статистики $F_{прав\ кр} = F(\alpha; m; n-m-1)$. Если $F_{набл}$ попадает в критическую область, т. е. $F_{набл} \in [F_{прав\ кр}; +\infty)$, то нулевая гипотеза отвергается, что говорит о соответствии теоретического уравнения регрессии выборочным данным.

Для оценки адекватности уравнения регрессии также используется *показатель средней ошибки аппроксимации* $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \tilde{y}_i|}{y_i} \cdot 100\,\%$.

Этап 5. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов. Возможна ситуация, когда часть вычисленных коэффициентов регрессии не обладает необходимой степенью значимости, т. е. значения данных коэффициентов будут меньше их стандартной ошиб-

ки. В этом случае такие коэффициенты должны быть исключены из уравнения регрессии. Поэтому проверка адекватности построенного уравнения регрессии наряду с проверкой значимости коэффициента детерминации R^2 включает в себя также и проверку значимости каждого коэффициента регрессии.

Для этого выдвигаются нулевые гипотезы о незначимости коэффициентов a_i, т. е. $H_0 : a_i = 0$ при альтернативных гипотезах $H_1 : a_i \neq 0$, $i = 1, 2, \ldots, m$.

Проверка данных гипотез осуществляется с помощью t-статистики $t = \dfrac{a_i}{\sigma_{a_i}}$, $i = 1, 2, \ldots, m$, имеющей распределение Стьюдента с $\nu = n - m - 1$ числом степеней свободы, где σ_{a_i}, $i = 1, 2, \ldots, m$, – стандартные значения ошибок для коэффициентов регрессии a_i соответственно.

По выборочным данным вычисляется наблюдаемое значение t-статистики $t_{набл}$, по распределению Стьюдента находится критическое значение $t_{кр} = t(\alpha;\ n - m - 1)$. Если $t_{набл}$ попадает в критическую область, т. е. $t_{набл} \in (-\infty; -t_{кр}) \cup (t_{кр}; +\infty)$, то нулевая гипотеза $H_0 : a_i = 0$, $i = 1, 2, \ldots, m$, отвергается.

Доверительные интервалы для каждого коэффициента a_i, $i = 1, 2, \ldots, m$, регрессии имеют вид $a_i \in (a_i - t_{кр}\sigma_{a_i};\ a_i + t_{кр}\sigma_{a_i})$.

Если незначимым окажется коэффициент a_0, то проводится пересчет уравнения регрессии в предположении, что $a_0 = 0$. Если же незначимым окажется коэффициент при факторе, то следует пересмотреть набор факторов в уравнении регрессии.

Коэффициенты $a_0, a_1, \ldots, a_m$ при факторах $\xi_1, \xi_2, \ldots, \xi_m$ показывают степень влияния каждого фактора на результативный признак η.

При практической интерпретации уравнения регрессии также широко используются *частные коэффициенты эластичности*, показывающие, на сколько процентов в среднем изменится значение результативного признака η при изменении значения соответствующего фактора ξ_i на 1%, и определяемые по формуле $E_{\xi_i} = a_i \dfrac{\overline{x}_i}{\overline{y}}$, где $\overline{x}_i$ – среднее значение соответствующего фактора ξ_i; $\overline{y}$ – среднее значение результативного признака η; a_i – коэффициент регрессии при соот-

ветствующем факторе.

При работе в *Excel* с целью нахождения коэффициентов множественной регрессии, множественного коэффициента детерминации, ковариационной матрицы используется режим *Регрессия* из *Анализа данных* меню *Сервис*. Результат анализа *Вывод итогов* представлен в виде пяти таблиц: *Регрессионная статистика*, *Дисперсионный анализ*, таблица, содержащая коэффициенты регрессии и их статистические оценки, *Вывод остатка* и *Вывод вероятности*. Формулы и функции *Excel*, по которым рассчитываются соответствующие числовые значения, приводятся в таблицах 24–28.

Таблица 24. **Результат анализа *Регрессия*: Регрессионная статистика**

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R (*коэффициент корреляции*)	$r_{XY} = \dfrac{\mathrm{cov}(X, Y)}{\sigma(X)\sigma(Y)}$ КОРРЕЛ

Окончание табл. 24

ВЫВОД ИТОГОВ	
Регрессионная статистика	
R-квадрат (*коэффициент детерминации*)	$R^2 = (r_{XY})^2$
Нормированный R-квадрат	$R^{2*} = 1 - (1 - R^2)\dfrac{n-1}{n-2}$
Стандартная ошибка (*остаточное стандартное отклонение*)	$\sigma_0 = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-(m+1)}}$ ЛИНЕЙН(с аргументом статистика=1)
Наблюдения (*число наблюдений*)	n СЧЕТ

Таблица 25. **Продолжение анализа *Регрессия*: Дисперсионный анализ**

Дисперсионный анализ			
	Регрессия	Остаток	Итого
df	$k_Ф = m$	$k_O = n - (m + 1)$	$k_Y = k_Ф + k_O$

SS	$SS_{\Phi}^2 = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2$	$SS_O^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$	$SS_Y^2 = SS_{\Phi}^2 + SS_O^2$
MS	$\frac{SS_{\Phi}^2}{m}$	$\frac{SS_O^2}{n-(m+1)}$	
F	$F_{набл} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$		
Значимость F	Значение уровня значимости при соответствующем $F_{набл}$ FРАСП($F_{набл}$; m; $n-(m+1)$)		

Замечание. В *Excel* на экран выводится таблица, строки которой записаны в столбцах табл. 24.

Таблица 26. **Результат анализа *Регрессия*: коэффициенты регрессии и их статистические оценки**

	Y-пересечение	Переменная *X*1	…	Переменная *Xm*
Коэффициенты	a_0	a_1	…	a_m

Окончание табл. 26

	Y-пересечение	Переменная *X*1	…	Переменная *Xm*
Стандартная ошибка	σ_{a_0}	σ_{a_1}	…	σ_{a_m}
t-статистика (*наблюдаемые значения статистики*)	$t_{0_{набл}} = \frac{a_0}{\sigma_{a_0}}$	$t_{1_{набл}} = \frac{a_1}{\sigma_{a_1}}$	…	$t_{m_{набл}} = \frac{a_m}{\sigma_{a_m}}$
P-значение (*значения уровня значимости, соответствующие вычисленным значениям статистики*)	СТЬЮДРАСП ($t_{0_{набл}}$; $n-m-1$)	СТЬЮДРАСП ($t_{1_{набл}}$; $n-m-1$)	…	СТЬЮДРАСП ($t_{m_{набл}}$; $n-m-1$)
Нижние 95% (*нижние границы доверительных интервалов*)	$a_0 - t_{кр}\sigma_{a_0}$	$a_1 - t_{кр}\sigma_{a_1}$	…	$a_m - t_{кр}\sigma_{a_m}$
	$t_{кр}$=СТЬЮДРАСПОБР(α; $n-m-1$)			
Верхние 95% (*верхние границы доверительных интервалов*)	$a_0 + t_{кр}\sigma_{a_0}$	$a_1 + t_{кр}\sigma_{a_1}$	…	$a_m + t_{кр}\sigma_{a_m}$
	$t_{кр}$=СТЬЮДРАСПОБР(α; $n-m-1$)			

Замечание. В *Excel* на экран выводится таблица, строки которой записаны в столбцах табл. 26.

Таблица 27. **Результат анализа *Регрессия*: Вывод остатка**

ВЫВОД ОСТАТКА			
Наблюдение	Предсказанное *Y*	Остатки	Стандартные остатки
1	$\tilde{y}_{1} = \quad \tilde{y}(x_1)$	$e_1 = y_1 - \tilde{y}_1$	$\dot{e}_1 = \dfrac{e_1 - \bar{e}}{\sigma_e}$
2	$\tilde{y}_{2} = \tilde{y}(x_2)$	$e_2 = y_2 - \tilde{y}_2$	$\dot{e}_2 = \dfrac{e_2 - \bar{e}}{\sigma_e}$
…	…	…	
n	$\hat{y}_{n} = \tilde{y}(x_n)$	$e_n = y_n - \tilde{y}_n$	$\dot{e}_n = \dfrac{e_n - \bar{e}}{\sigma_e}$

Замечание. В табл. 27 при вычислении стандартных остатков используется $\bar{e} = \dfrac{1}{n}\sum_{i=1}^{n} e_i$, $\sigma_e = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2}$.

Таблица 28. **Результат анализа *Регрессия*: Вывод вероятности**

ВЫВОД ВЕРОЯТНОСТИ	
Персентиль (*k-я персентиль для диапазона данных*) ПЕРСЕНТИЛЬ(массив; *k*)	*Y* (*эмпирические значения η, соответствующие k-й персентили*)
0/*n*-я персентиль	$y_{\min}$
…	…
i/*n*-я персентиль	y_i
…	…
n/*n*-я персентиль	$y_{\max}$

Анализ *Регрессия* выполняет также построение графиков *График*

остатков, График подбора, График нормальной вероятности.

График остатков представляет собой в плоскости Ox_me точечную диаграмму остатков e_i для каждой независимой переменной x_i, $i=\overline{1,m}$. *График подбора* представляет собой в плоскости Ox_iy диаграммы наблюдаемых y_i и предсказанных $\tilde{y}_i$ значений для каждой независимой переменной x_i, $i=\overline{1,m}$. *График нормальной вероятности* представляет собой в плоскости Owy диаграмму персентилей $w_i=\frac{i}{n}$ (в процентах) и соответствующих выборочных значений y_i, i = 1, 2, …, n.

Вопросы для самоконтроля

1. Какая зависимость называется стохастической?
2. Каковы основные задачи корреляционного анализа?
3. Какая зависимость называется корреляционной?
4. Какая величина называется выборочной ковариацией? Что она характеризует?
5. Как по коэффициенту корреляции оценить тесноту связи между случайными величинами?
6. Какова формулировка алгоритма оценивания тесноты связи между случайными величины?
7. В чем разница между корреляционным и регрессионным анализами?
8. Как построить корреляционное поле?
9. Какое уравнение называется уравнением регрессии?
10. Какой вид имеет уравнение линейной регрессии?
11. В чем суть метода наименьших квадратов?
12. Как связаны общая и факторная дисперсии при корреляционной зависимости результативного признака и факторов?
13. Что используется для анализа общего качества уравнения линейной регрессии?
14. С помощью какого критерия проверяется гипотеза о незначимости множественного коэффициента детерминации?
15. Что оценивает показатель средней ошибки аппроксимации?

16. С помощью какого критерия осуществляется проверка значимости каждого коэффициента регрессии?

17. Что характеризуют частные коэффициенты эластичности?

4. Лабораторная работа

Задание 1

Показатели уровня образования, уровня преступности, а также отношение числа безработных к числу вакансий в областях Беларуси в 2002 г. приведены в таблице *MS Excel* на рис. 39 (уровень образования рассчитывался как численность лиц с высшим и средним специальным образованием на 1000 жителей области, уровень преступности – как число совершенных преступлений на 100 тыс. жителей области).

По выборочным данным требуется установить наличие взаимосвязи между указанными показателями в республике.

	A	B	C	D
1	Область	Уровень образования	Отношение числа безработных к числу вакансий	Уровень преступности
2	Минская	788	10,8	791
3	Гомельская	756	72,3	909
4	Могилевская	779	52,9	804
5	Брестская	740	10,4	695
6	Витебская	763	5,4	935
7				

Рис. 39. **Данные для корреляционного анализа**

Задание 2

Данные о прибыли шести предприятий, оборотных средств, стоимости основных фондов приведены в таблице *MS Excel* на рис. 40.

По предоставленным данным требуется определить параметры уравнения регрессии и провести его анализ.

	A	B	C	D
1	Номер предприятия	Прибыль η	Величина оборотных средств ξ_1	Стоимость основных фондов ξ_2
2	1	188	129	510
3	2	78	64	190
4	3	93	69	240
5	4	152	87	470
6	5	55	47	110
7	6	161	102	420
8				

Рис. 40. **Данные для множественной линейной регрессии**

Порядок выполнения работы

Задание 1

Для определения тесноты и вида связи между уровнем образования (ξ_1), отношением числа безработных к числу вакансий (ξ_2), уровнем преступности (ξ_3) используются режимы *Ковариация* и *Корреляция* из *Анализа данных* меню *Сервис*.

Значения параметров в диалоговом окне *Ковариация* устанавливаются следующим образом (рис. 41):

- *Входной интервал* (вводятся ссылки на ячейки B1:D6, в которых находятся названия и наблюдаемые значения исследуемых признаков ξ_1, ξ_2, ξ_3);
- *Группирование* (устанавливается флажок по столбцам);
- *Метки в первой строке* (устанавливается флажок);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист Ковариация*).

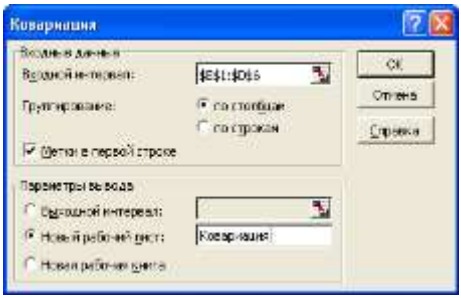


Рис. 41. **Диалоговое окно *Ковариация***

Показатели, рассчитанные в ходе анализа, представлены на рис. 42.

	A	B	C	D
1		*Уровень образования*	*Отношение числа безработных к числу вакансий*	*Уровень преступности*
2	Уровень образования	434,8		
3	Отношение числа безработных к числу вакансий	91,28	295,4024	
4	Уровень преступности	-58,4	-92,688	8091,36
5				

Рис. 42. **Результат анализа *Ковариация***

Так как все значения ковариаций между всем парами исследуемых признаков отличны от нуля, то между признаками существуют стохастические связи.

Значения параметров в диалоговом окне *Корреляция* устанавливаются следующим образом (рис. 43):

- *Входной интервал* (вводятся ссылки на ячейки B1:D6, в которых находятся названия и наблюдаемые значения исследуемых признаков ξ_1, ξ_2, ξ_3);
- *Группирование* (устанавливается флажок по столбцам);
- *Метки в первой строке* (устанавливается флажок);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист Корреляция*).

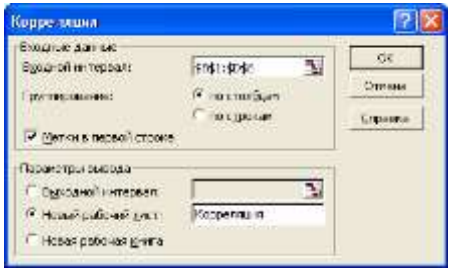


Рис. 43. **Диалоговое окно *Ковариация***

Очевидно, что характер всех выявленных связей заключается в следующем:

• связь «уровень образования – отношение числа безработных к числу вакансий» является слабой и прямой (r = 0,2447), т. е. с повышением образования отношение числа безработных к числу вакансий увеличивается;

• связь «уровень образования – уровень преступности» является слабой и обратной (r = –0,03115), т. е. с повышением уровня образования уровень преступности уменьшается;

• связь «отношение числа безработных к числу вакансий – уровень преступности» является слабой и обратной (r = –0,05998), т. е. с повышением отношения числа безработных к числу вакансий уровень преступности уменьшается.

Показатели, рассчитанные в ходе анализа, представлены на рис. 44.

	A	B	C	D
1		*Уровень образования*	*Отношение числа безработных к числу вакансий*	*Уровень преступности*
2	Уровень образования	1		
3	Отношение числа безработных к числу вакансий	0,254697014	1	
4	Уровень преступности	-0,031154863	-0,059989366	1

Рис. 44. **Результат анализа *Корреляция***

Проверим коэффициенты корреляции на значимость. В ячейки B5–B8 вводятся следующие формулы:

=ABS(B3)*КОРЕНЬ(5-2)/КОРЕНЬ(1-B3^2),
=ABS(B4)*КОРЕНЬ(5-2)/КОРЕНЬ(1-B4^2),
=ABS(C4)*КОРЕНЬ(5-2)/КОРЕНЬ(1-C4^2),
=СТЬЮДРАСПОБР(0,05;5-2).

Так как все наблюдаемые значения коэффициентов корреляции меньше критического (рис. 45), то все они незначимы, т. е. между указанными показателями отсутствует взаимосвязь.

	A	B
5	t1	0,456193
6	t2	0,053936
7	t3	0,103719
8	tкр	3,182446

Рис. 45. **Проверка коэффициентов корреляции на значимость**

Задание 2

При нахождении уравнения регрессии выполняются этапы, приведенные ниже.

На первом этапе устанавливается форма зависимости. Предполагается, что зависимость между результативным признаком (прибыль η) и такими факторными признаками, как величина оборотных средств ξ_1 и стоимость основных фондов ξ_2, является линейной.

На втором этапе определяются параметры регрессии. Для этого используется режим анализа *Регрессия*. Значения параметров в диалоговом окне *Регрессия* устанавливаются следующим образом (рис. 46):

- *Входной интервал Y* (вводятся ссылки на ячейки B1:B7, в которых находятся названия и наблюдаемые значения результативного признака η);
- *Входной интервал X* (вводятся ссылки на ячейки C1:D7, в которых находятся названия и наблюдаемые значения факторных признаков ξ_1 и ξ_2);
- *Метки* (устанавливается флажок);
- *Уровень надежности* (устанавливается флажок);
- *Константа-ноль* (не активизируется);
- *Параметры вывода* (вводится ссылка на *Новый рабочий лист Регрессия*);
- *Остатки* (устанавливается флажок);

• *Стандартизированные остатки* (устанавливается флажок);
• *График остатков* (устанавливается флажок);
• *Остатки* (устанавливается флажок);
• *График подбора* (устанавливается флажок);
• *Нормальная вероятность* (устанавливается флажок *График нормальной вероятности*).

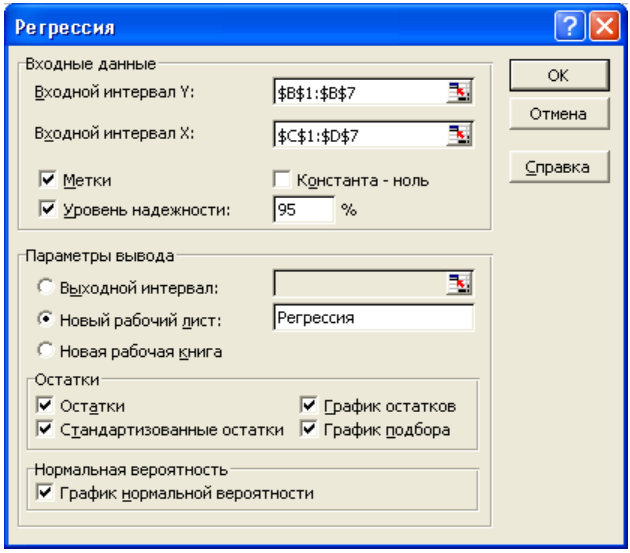


Рис. 46. **Заполнение диалогового окна *Регрессия***

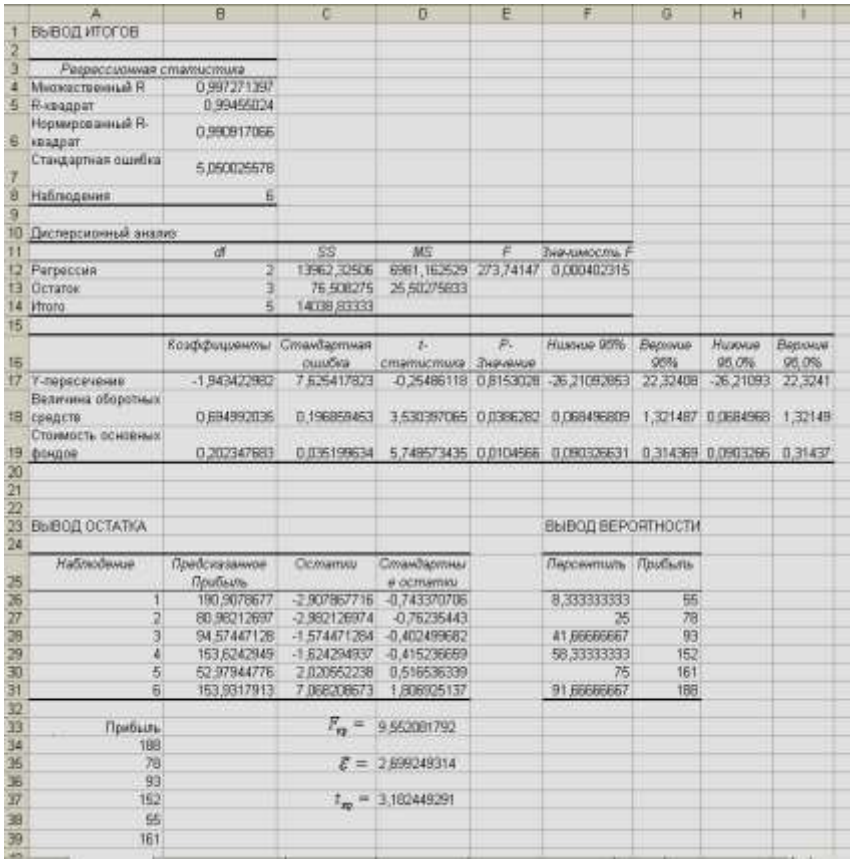
Рассчитанные в данном режиме показатели представлены на рис. 47.

Уравнение линейной множественной регрессии, учитывая рассчитанные коэффициенты (ячейки B17:B19), записывается в виде

$$\tilde{y} = -1{,}94 + 0{,}69x_1 + 0{,}20x_2 .$$

На следующем этапе проводится анализ общего качества уравнения регрессии. Значение множественного коэффициента детерминации R^2 = 0,995 (ячейка B5) показывает, что 99,5% общей вариации результативного признака η объясняется вариацией факторных признаков ξ_1 и ξ_2. Следовательно, выбранные факторы существенно влияют на прибыль предприятий, что подтверждает их правильность включения в построенную модель. Более того, значение R^2 близко к единице,

что говорит о том, что выборочные значения η, ξ_1 и ξ_2 достаточно близко расположены к линии регрессии.



	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,997271397							
5	R-квадрат	0,99455024							
6	Нормированный R-квадрат	0,990917066							
7	Стандартная ошибка	5,050025578							
8	Наблюдения	6							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	2	13962,32506	6981,162529	273,74147	0,000402315			
13	Остаток	3	76,508275	25,50275833					
14	Итого	5	14038,83333						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	-1,943422982	7,625417823	-0,25486118	0,8153028	-26,21092853	22,32408	-26,21093	22,3241
18	Величина оборотных средств	0,884992036	0,196859463	3,530397065	0,0386282	0,058496809	1,321487	0,0584968	1,32149
19	Стоимость основных фондов	0,202347883	0,035199634	5,749573435	0,0104566	0,090326631	0,314369	0,0903266	0,31437
20									
21									
22									
23	ВЫВОД ОСТАТКА					ВЫВОД ВЕРОЯТНОСТИ			
24									
25	*Наблюдение*	*Предсказанное Прибыль*	*Остатки*	*Стандартные остатки*		*Перцентиль*	*Прибыль*		
26	1	190,9078677	-2,907867716	-0,743370706		8,333333333	55		
27	2	80,98212697	-2,982126974	-0,76235443		25	78		
28	3	94,57447126	-1,574471264	-0,402499682		41,66666667	93		
29	4	153,6242945	-1,624294937	-0,415236659		58,33333333	152		
30	5	52,97944776	2,020552238	0,516536339		75	161		
31	6	153,9317913	7,068208673	1,806925137		91,66666667	188		
32									
33	Прибыль		$F_{кр}$ =	9,552091792					
34	188								
35	78		ξ =	2,699249314					
36	93								
37	152		$t_{кр}$ =	3,182449291					
38	55								
39	161								

Рис. 47. **Результат анализа *Регрессия***

Для проверки гипотезы $H_0 : R^2 = 0$ при альтернативной гипотезе $H_1 : R^2 \neq 0$ используется наблюдаемое значение F-статистики $F_{набл}$ = 273,74, находящееся в ячейке E12. Критическое значение $F_{кр}$ статистики вычисляется в ячейке D33 отдельно по формуле

=FРАСПОБР(0,05;B12;B13)

и равно $F_{кр}$ = 9,55. Поскольку 273,74 = $F_{набл} > F_{кр}$ = 9,55, то нулевая

гипотеза отклоняется, т. е. коэффициент детерминации является значимым. Для *P*-значения имеет место неравенство $P(F_{набл}) = 0{,}0004 < 0{,}05$, что также подтверждает значимость коэффициента детерминации (*P*-значение находится в ячейке F12 – *Значимость F*).

Вычисление показателя средней ошибки аппроксимации $\bar{\varepsilon}$ осуществляется также после выполнения анализа *Регрессия*. Для этого выполняются следующие операции:

- в ячейки A34:A39 вводятся значения результативного признака η (прибыль);
- в ячейку D35 вводится формула массива

{=(СУММ(ABS(C26:C31)/(A34:A39))/6)*100},

где C26:C31 – ячейки, содержащие остатки, т. е. разности между эмпирическими и теоретическими значениями результативного признака.

В результате получается значение $\bar{\varepsilon} = 2{,}699\%$, подтверждающее достаточно высокую адекватность построенного уравнения.

Затем проводится проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов. Для проверки гипотез $H_0 : a_0 = 0, a_1 = 0, a_2 = 0$ при альтернативных гипотезах $H_1 : a_i \neq 0, i = 0, 1, 2$ в анализе *Регрессия* рассчитываются в ячейках D17:D19 наблюдаемые значения *t*-статистики: для $a_0 : t_{набл}^{a_0} = -0{,}255$, для $a_1 : t_{набл}^{a_1} = 3{,}5303$, для $a_2 : t_{набл}^{a_2} = 5{,}7485$.

Поскольку в таблицах не приводится значение $t_{кр}$, то оно вычисляется в ячейке D37 по формуле =СТЬЮДРАСПОБР(0,05;B8-2-1) и равно $t_{кр} = 3{,}1824$. Здесь 0,05 – уровень значимости, B8 – число наблюдений, 2 – число факторов в уравнении регрессии, 1 – число свободных членов в уравнении регрессии.

Так как $t_{набл}^{a_0} \in (-\infty; -3{,}18) \cup (3{,}18; +\infty)$, то коэффициент a_0 не является значимым, он меньше своей стандартной ошибки

$$|a_0| = -1{,}94 < \sigma_{a_0} = 7{,}625.$$

Наблюдаемые значения t-статистики попадают в критическую область $t_{набл}^{a_1} \in (-\infty; -3{,}18) \cup (3{,}18; +\infty)$, $t_{набл}^{a_2} \in (-\infty; -3{,}18) \cup (3{,}18; +\infty)$, поэтому коэффициенты a_1 и a_2 являются значимыми. Для них

$$|a_1| = 0{,}69 > \sigma_{a_1} = 0{,}197, \; |a_2| = 0{,}2 > \sigma_{a_2} = 0{,}035.$$

P-значения коэффициентов a_0, a_1, a_2 удовлетворяют неравенствам

$$P(a_0) = 0{,}8153 > 0{,}05 = \alpha, \; P(a_1) = 0{,}0386 < 0{,}05 = \alpha,$$
$$P(a_2) = 0{,}0104 < 0{,}05 = \alpha,$$

что также говорит о незначимости коэффициента a_0 и значимости коэффициентов a_1, a_2.

Из предварительного анализа уравнения регрессии следует, что его необходимо пересчитать без свободного члена a_0, который не является статистически значимым.

Для пересчета уравнения регрессии в диалоговом окне *Регрессия* задаются те же параметры, но устанавливается флажок активизации параметра *Константа-ноль* (рис. 48).

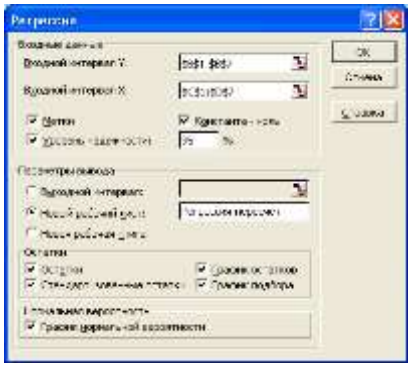


Рис. 48. **Заполнение диалогового окна *Регрессия* при условии $a_0 = 0$**

После пересчета на *Новом рабочем листе Регрессия-пересчет* генерируются таблицы (рис. 49), аналогичные таблицам, изображенным на рис. 47.

Таким образом, получается новое уравнение регрессии

$$\tilde{y} = 0{,}66x_1 + 0{,}21x_2.$$

Для нового выборочного коэффициента детерминации $R^2 = 0{,}9944$ аналогично предыдущим рассуждениям имеем

$$357{,}211 = F_{набл} > F_{кр} = 9{,}55;\ P(F_{набл}) = 0{,}00027 < 0{,}05.$$

Следовательно, R^2 является значимым.

Для коэффициентов a_1, a_2 при факторных признаках наблюдаемые значения t-статистики принадлежат критической области:

$$5{,}94654 = t_{набл}^{a_1} \in (-\infty; -3{,}18) \cup (3{,}18; +\infty);$$

$$7{,}6520 = t_{набл}^{a_2} \in (-\infty; -3{,}18) \cup (3{,}18; +\infty).$$

Следовательно, они являются значимыми.

Значения коэффициентов a_1, a_2 больше своих стандартных ошибок

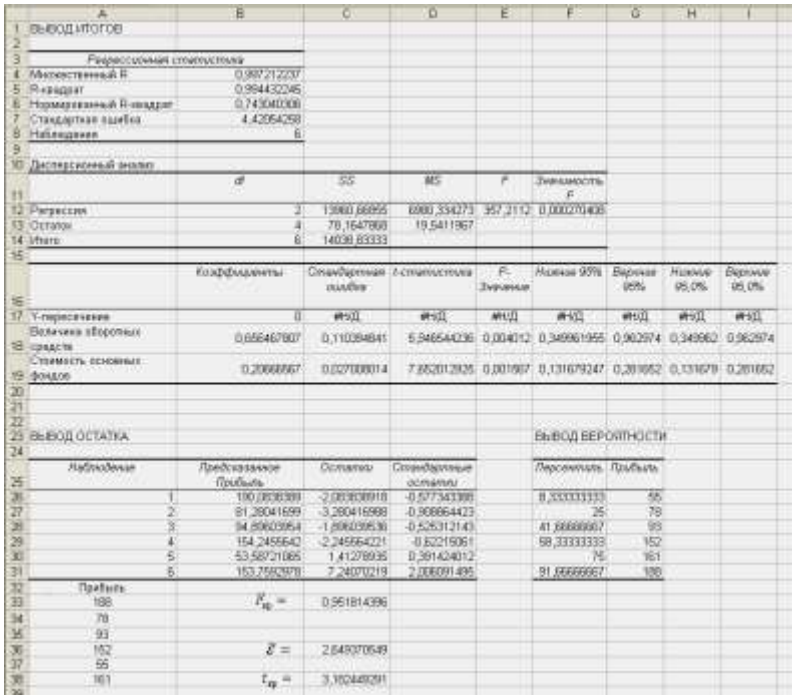
$$a_1 = 0{,}6564 > 0{,}11039 = \sigma_{a_1},\ a_2 = 0{,}2066 > 0{,}0270 = \sigma_{a_2},$$

и P-значения коэффициентов a_1, a_2 удовлетворяют неравенствам

$$a_1 : P = 0{,}0040 < 0{,}05 = \alpha,\ a_2 : P = 0{,}0015 < 0{,}05 = \alpha,$$

что говорит о значимости коэффициентов a_1, a_2.

Итак, полученное уравнение регрессии адекватно экспериментальным данным.



	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,997212237							
5	R-квадрат	0,994432245							
6	Нормированный R-квадрат	0,743040308							
7	Стандартная ошибка	4,42664259							
8	Наблюдения	6							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	2	13960,66895	6980,334473	357,2112	0,002704408			
13	Остаток	4	78,1647868	19,5411967					
14	Итого	6	14038,83333						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
18	Величина оборотных средств	0,656467807	0,110384841	5,946544236	0,004012	0,349661955	0,963274	0,349662	0,962674
19	Стоимость основных фондов	0,20668667	0,027088014	7,632012826	0,001567	0,131679247	0,281652	0,131679	0,281652
20									
21									
22									
23	ВЫВОД ОСТАТКА					ВЫВОД ВЕРОЯТНОСТИ			
24									
25	*Наблюдение*	*Предсказанное Прибыль*	*Остатки*	*Стандартные остатки*		*Персентиль*	*Прибыль*		
26	1	190,0836389	-2,083638916	-0,577343398		8,333333333	55		
27	2	81,28041699	-3,280416988	-0,908864423		25	78		
28	3	94,89603954	-1,896039536	-0,525312143		41,66666667	93		
29	4	154,2456642	-2,245664221	-0,62215061		58,33333333	152		
30	5	53,58721085	1,412789135	0,391424012		75	161		
31	6	153,7582878	7,240702119	2,006091495		91,66666667	188		
32	Прибыль								
33	188	R_{xy} =	0,961814396						
34	78								
35	93								
36	152	$\bar{x}$ =	2,649370549						
37	55								
38	161	t_{xy} =	3,192449291						

Рис. 49. **Результат анализа *Регрессия* (пересчет уравнения регрессии)**

Коэффициенты a_1, a_2 показывают степень влияния каждого фактора на прибыль предприятия. Увеличение оборотных средств на 1 млн р. ведет к росту прибыли на 0,66 млн р., а увеличение стоимости основных фондов на 1 млн р. ведет к росту прибыли на 0,21 млн р.

Коэффициенты эластичности E_{ξ_1} и E_{ξ_2} факторов ξ_1 и ξ_2 вычисляются на новом листе, для чего необходимо выполнить следующее:

- в ячейки A1:C7 копируются выборочные значения результативного признака η и факторов ξ_1 и ξ_2;
- в ячейки E2 и F2 копируются вычисленные коэффициенты регрессии a_1 и a_2 соответственно;
- в ячейках A9, B9, C9 вычисляются средние значения $\bar{y}$, $\bar{x}_1$, $\bar{x}_2$ соответственно с помощью функции СРЗНАЧ;
- в ячейке B12 вычисляется коэффициент эластичности E_{ξ_1} по формуле =E2*(B9/A9), в ячейке B14 – коэффициент эластичности E_{ξ_2} : =F2*(C9/A9).

Результаты вычислений коэффициентов E_{ξ_1} и E_{ξ_2} приведены на рис. 50.

	A	B	C	D	E	F
1	Прибыль η	Величина оборотных средств ξ_1	Стоимость основных фондов ξ_2		Коэффициент a_1	Коэффициент a_2
2	188	129	510		0,656467807	0,20686567
3	78	64	190			
4	93	69	240			
5	152	87	470			
6	55	47	110			
7	161	102	420			
8						
9	121,1667	83	323,3333333			
10						
11						
12	E_{ξ_1} =	0,44968496				
13						
14	E_{ξ_2} =	0,55148748				
15						

Рис. 50. **Вычисление коэффициентов эластичности E_{ξ_1} и E_{ξ_2}**

Поскольку E_{ξ_2} = 0,5514 > 0,4496 = E_{ξ_1}, то по абсолютному приросту наибольшее влияние на прибыль предприятий оказывает второй фактор: увеличение стоимости основных фондов ξ_2 на 1% вызывает рост прибыли на 0,55%, тогда как рост величины оборотных средств ξ_1 на 1% способствует росту прибыли на 0,45%.

ИНДИВИДУАЛЬНЫЕ ЗАДАНИЯ

1. Индивидуальное задание по теме «Выборка и ее анализ»

Цель: приобретение практических навыков первичной обработки статистических данных с помощью *Microsoft Excel*.

Задачи: представить выборочные данные из генеральной совокупности случайной величины в виде статистических рядов; построить полигоны частот и гистограммы; по виду полигона частот и гистограммы выдвинуть предположения о законе распределения случайной величины; сравнить результаты, полученные по разным выборкам.

Задание. В результате наблюдений получены выборки X и Y (табл. 29) из генеральных совокупностей случайных величин ξ_1 и ξ_2, характеризующих колебания курсов двух ценных бумаг в течение определенного времени.

Таблица 29. **Выборка X (x_i + 0,3k, где k – номер в журнале)**

57	46	33	49	29	50	38	41	27	34
37	49	51	26	55	42	59	43	46	30
31	43	58	41	35	47	33	45	49	37
47	34	54	39	60	49	25	50	31	53
Выборка Y (y_i + 0,3k, где k – номер в журнале)									
38	41	30	51	37	55	47	43	35	42
35	46	27	45	41	34	50	29	51	39
42	59	43	31	38	58	54	37	26	43
29	42	33	41	24	39	53	45	33	51
45	25	54	50	37	30	41	60	42	46
38	53	34	47	35	49	57	39	55	31

Требуется:
1) построить вариационный ряд и полигон относительных частот;
2) построить интервальный ряд, гистограмму, кумулятивную кривую.
Сделать выводы, представить отчет.

2. Индивидуальное задание по теме «Точечные и интервальные оценки»

Цель: научиться реализовывать методы точечного и интервального оценивания неизвестных параметров распределения случайной величины.

Задача: использовать возможности *MS Excel* при нахождении точечных и интервальных оценок неизвестных параметров распределения случайной величины.

Задание. В результате наблюдений получены выборки *X* и *Y* (табл. 29) из генеральных совокупностей случайных величин ξ_1 и ξ_2, характеризующих средний доход по акциям от определенного производства.

Требуется:

1) определить числовые характеристики выборки;

2) предполагая, что случайные величины ξ_1 и ξ_2 имеют нормальное распределение с плотностью $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, методами моментов и максимального правдоподобия оценить неизвестные параметры *a* и σ;

3) найти доверительные интервалы для математического ожидания и дисперсии при уровне надежности 95%, считая, что ξ_1 и ξ_2 имеют нормальное распределение с параметрами *a* и σ.

Сделать выводы и представить отчет.

3. Индивидуальное задание по теме «Проверка статистических гипотез о параметрах нормально распределенной генеральной совокупности»

Цель: приобретение навыков проверки статистических гипотез с помощью *MS Excel*.

Задачи: уметь формулировать нулевую и альтернативную гипотезы, выбирать соответствующие критерии для их проверки, использовать для проверки гипотез функции и дополнение *Анализ данных*.

Задание. Пусть случайная величина ξ_1 характеризует результат про-

веденного лабораторного анализа на наличие вредных веществ по одной технологии, а ξ_2 – по другой, причем предполагается, что ξ_1 и ξ_2 нормально распределены. В результате наблюдений получены выборка X из генеральной совокупности случайной величины ξ_1 и выборка Y из генеральной совокупности случайной величины ξ_2 (табл. 29).

Требуется при заданном уровне значимости $\alpha = 0{,}05$ проверить следующие гипотезы:

1) о нормальном законе распределения генеральной совокупности случайной величины ξ_1, используя критерии Пирсона и Колмогорова;

2) о равенстве дисперсий двух нормальных генеральных совокупностей, математические ожидания которых неизвестны;

3) о равенстве математических ожиданий двух нормальных генеральных совокупностей, дисперсии которых известны и равны;

4) о равенстве математических ожиданий двух нормальных генеральных совокупностей, дисперсии которых неизвестны и равны;

5) о равенстве математических ожиданий двух нормальных генеральных совокупностей, дисперсии которых неизвестны и различны.

Сделать выводы и представить отчет.

4. Индивидуальные задания по теме «Дисперсионный анализ»

Цель: приобретение навыков проведения однофакторного и двухфакторного (с повторениями и без повторений) анализов в пакете *MS Excel*.

Задачи: уметь выбрать область применения однофакторного и двухфакторного анализов, подобрать необходимые методы и формулы для расчетов, использовать функции и дополнение *Анализ данных* для проведения однофакторного и двухфакторного анализов.

Задание 1. На химическом заводе разработаны два новых варианта технологического процесса. Чтобы оценить, как изменится дневная производительность при переходе на работу по новым технологиям, завод в течение 10 дней работает по каждому варианту, включая существующий. Дневная производительность завода (в условных единицах) представлена в табл. 30.

Таблица 30. **Данные о дневной производительности по различным технологиям (каждое значение надо умножить на *k*, где *k* – номер в журнале)**

День работы	Дневная производительность		
	Существующая технология	Вариант 1	Вариант 2
1	119	53	76
2	26	128	149
3	96	176	97

Окончание табл. 30

День работы	Дневная производительность		
	Существующая технология	Вариант 1	Вариант 2
4	174	106	77
5	71	96	114
6	101	172	197
7	135	81	179
8	171	163	178
9	148	73	90
10	20	108	109

При уровне значимости $\alpha = 0{,}01$ требуется выяснить, как зависит дневная производительность от технологического процесса. Следует оценить степень этой зависимости.

Сделать выводы и представить отчет.

Задание 2. В процессе производства получается побочный нежелательный продукт, процентное содержание которого измеряется тремя различными катализаторами при двух значениях давления. Данные о процентном содержании побочного продукта для трех катализаторов в случаях верхнего или нижнего уровней давления представлены в табл. 31.

Таблица 31. **Данные о процентном содержании побочных продуктов**

(к каждому значению надо прибавить 0,02*k*, где *k* – номер в журнале)

Уровень давления	Катализатор		
	1	2	3
Верхний	1,31	1,22	1,09
Нижний	1,13	1,04	1,44

Требуется при уровне значимости $\alpha = 0,05$ определить, что оказывает более значимое влияние на выявление побочного продукта (катализатор или уровень давления).

Сделать выводы и представить отчет.

Задание 3. Партии мяса от пяти различных поставщиков загружаются в агрегат для упаковки в банки, который имеет шесть наполняющих цилиндров. Случайным образом отбираются по три банки, наполненные каждым из шести цилиндров. Масса соответствующих банок представлена в табл. 32.

Таблица 32. **Данные массы банок мяса, наполненных на различных цилиндрах (к каждому значению надо прибавить 0,6*k*, где *k* – номер в журнале)**

Номер банки	Номер цилиндра	Номер поставщика				
		1	2	3	4	5
1	Цилиндр 1	9,4	7,9	0,3	6,8	4,2
2		8,7	11,1	7,7	1,8	5,5
3		1,1	7,3	3,7	4,4	11,8
1	Цилиндр 2	10,0	11,6	5,2	2,1	5,7
2		8,3	7,6	0,7	7,3	3,4
3		2,3	4,7	9,7	11,3	5,8
1	Цилиндр 3	5,0	0,4	8,9	7,1	9,6
2		9,5	5,9	1,0	9,1	2,5
3		10,2	2,6	4,5	3,1	6,2
1	Цилиндр 4	1,2	6,4	11,9	6,9	8,1
2		11,5	5,9	9,0	5,3	2,4
3		11,9	7,8	1,7	5,1	8,8
1	Цилиндр 5	10,5	4,3	3,8	1,4	11,2

2		8,6	1,9	5,6	10,1	8,4
3		4,8	7,4	0,8	4,7	3,6
1	Цилиндр 6	11,4	5,7	9,9	2,2	7,2
2		6,0	8,2	3,2	10,9	6,1
3		11,7	6,3	0,2	7,0	9,2

Требуется при уровне значимости $\alpha = 0{,}05$ выяснить, влияет ли на массу банки наполняющий ее цилиндр.

Сделать выводы и представить отчет.

5. Индивидуальное задание по теме «Исследование взаимосвязи между признаками»

Цель: приобретение практических навыков проведения ковариационного и регрессионного анализов в пакете *MS Excel*, а также оценки параметров модели линейной регрессии.

Задачи: уметь проводить ковариационный и корреляционный анализы данных в пакете *MS Excel*, строить линейную регрессионную модель, проверять ее адекватность, определять степень влияния факторов на результативный признак.

Задание. Исследуется зависимость заработной платы (η) специалистов определенной профессии от возраста (ξ_1), образования (числа лет обучения) (ξ_2), стажа работы по данной специальности (ξ_3). Случайным образом выбраны 25 специалистов. Результаты обследования приведены в табл. 33.

Таблица 33. **Данные результатов обследования специалистов (*k* – номер в журнале)**

Заработная плата (в условных единицах)	Число лет обучения	Возраст, лет	Стаж работы по данной специальности, лет
225 + 10*k*	5	28	10
197 + 10*k*	3	26	8
198 + 10*k*	3	30	14

210 + 10*k*	7	35	15
309 + 10*k*	9	41	16
402 + 10*k*	15	45	18
220 + 10*k*	6	27	3
207 + 10*k*	10	35	13
207 + 10*k*	12	30	10
100 + 10k	0	23	2
227 + 10k	11	29	7
214 + 10k	8	33	11
203 + 10k	8	40	20
215 + 10k	10	41	21
184 + 10k	6	41	6
216 + 10k	4	23	1
190 + 10k	2	32	12

Окончание табл. 33

Заработная плата (в условных единицах)	Число лет обучения	Возраст, лет	Стаж работы по данной специальности, лет
217 + 10*k*	5	37	13
197 + 10*k*	5	31	7
200 + 10*k*	7	30	9

Требуется:

1) по выборочным данным установить наличие взаимосвязи между показателями η, ξ_1, ξ_2, ξ_3;

2) оценить параметры уравнения линейной регрессии и провести его анализ (проверить адекватность, определить степень влияния факторов на результативный признак).

Сделать выводы и представить отчет.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Афифи, А. Статистический анализ: Подход с использованием ЭВМ : [пер. с англ.] / А. Афифи, С. Эйзен. – М. : Мир, 1982. – 488 с.

Булдык, Г. М. Теория вероятностей и математическая статистика : учеб. пособие / Г. М. Булдык. – Минск : Выш. шк., 1989. – 285 с.

Вентцель, Е. С. Теория вероятностей : учеб. пособие / Е. С. Вентцель. – 7-е изд., стер. – М. : Высш. шк., 2001. – 575 с.

Жевняк, Р. В. Теория вероятностей и математическая статистика : учеб. пособие / Р. В. Жевняк, А. А. Карпук, В. Т. Унукович. – Минск : Харвест, 2000. – 384 с.

Закс, Л. Статистическое оценивание / Л. Закс. – М. : Статистика, 1976. – 598 с.

Ильина, О. П. Статистический анализ и пргнозирование экономической информации в электронной таблице Excel 5.0. Microsoft : учеб. пособие / О. П. Ильина, Н. В. Макарова. – СПб. : СПбГУЭФ, 1996. – 140 с.

Калинина, В. Н. Математическая статистика : учеб. пособие / В. Н. Калинина, В. Ф. Панкин. – М. : Высш. шк., 1998. – 336 с.

Колемаев, В. А. Теория вероятностей и математическая статистика : учеб. пособие / В. А. Колемаев, О. В. Староверов, В. Б. Турундаевский ; под ред. В. А. Колемаева. – М. : Высш. шк., 1991. – 400 с.

Мацкевич, И. П. Высшая математика: теория вероятностей и математическая статистика : учеб. пособие / И. П. Мацкевич, Г. П. Свирид. – Минск : Выш. шк., 1993. – 259 с.

Салманов, О. Н. Математическая экономика с применением Mathcad и Excel / О. Н. Салманов. – СПб. : БХВ-Петербург, 2003. – 464 с.

Сборник задач по математике для втузов. Специальные курсы : учеб. пособие / Э. А. Вуколов [и др.] ; под ред. А. В. Ефимова. – М. : Наука, 1984. – 608 с.

Справочник по вероятностным расчетам / Г. Г. Абезгауз [и др.]. – М. : Воен. изд-во, 1970. – 537 с.

Статистический ежегодник Республики Беларусь / под ред. В. И. Зиновского. – Минск : Информстат М-ва статистики Респ. Беларусь, 2002. – 600 с.

Тюрин, Ю. Н. Статистический анализ данных на компьютере : учеб. пособие / Ю. Н. Тюрин, А. А. Макаров ; под ред. В. Э. Фигурнова. – М. : ИНФРА-М, 1998. – 528 с.

Четыркин, Е. М. Вероятность и статистика / Е. М. Четыркин, И. Л. Калихман. – М. : Финансы и статистика, 1982. – 320 с.

СОДЕРЖАНИЕ

Учебное издание

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Пособие
для студентов экономических специальностей

Авторы-составители:
Авдашкова Людмила Павловна
Марченко Лариса Николаевна

Редактор И. А. Михайлова
Технический редактор И. А. Козлова
Компьютерная верстка Н. Н. Короедова

Подписано в печать 15.01.10. Бумага типографская № 1.
Формат 60 × 84 $^{1}/_{16}$. Гарнитура Таймс. Ризография.
Усл. печ. л. 7,21. Уч.-изд. л. 7,35. Тираж 400 экз.
Заказ №

Учреждение образования
«Белорусский торгово-экономический университет
потребительской кооперации».
ЛИ № 02330/0494302 от 04.03.2009 г.
246029, г. Гомель, просп. Октября, 50.

Отпечатано в учреждении образования
«Белорусский торгово-экономический университет
потребительской кооперации».
130

246029, г. Гомель, просп. Октября, 50.