МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
Белорусский национальный технический университет

Кафедра «Инженерная математика»

Н. К. Прихач
И. В. Прусова
В. М. Романчак

ПРИКЛАДНАЯ МАТЕМАТИКА. ВЫБОРКА И ЕЕ АНАЛИЗ

Учебно-методическое пособие
для студентов специальности
1-54 01 01 «Метрология, стандартизация и сертификация
(машиностроение и приборостроение)»

Рекомендовано учебно-методическим объединением по образованию в области обеспечения качества

Под редакцией М. А. Князева

Минск
БНТУ
2022

УДК 51-7(075.8)
ББК 22.1я7
П77

Р е ц е н з е н т ы:
профессор кафедры математики и физики УО «Белорусская государственная академия связи», доктор физ.-мат. наук, доцент *Л. Л. Гладков*;
заведующий кафедрой теории вероятности и математической статистики ФМПИ, главный научный сотрудник НИЛ СТАМ, доктор физ.-мат. наук, доцент *А. Ю. Харин*

Прихач, Н. К.
П77 Прикладная математика. Выборка и ее анализ : учебно-методическое пособие для студентов специальности 1-54 01 01 «Метрология, стандартизация и сертификация (машиностроение и приборостроение» / Н. К. Прихач, И. В. Прусова, В. М. Романчак; под ред. М. А. Князева. – Минск : БНТУ, 2022. – 76 с.
ISBN 978-985-583-757-3.

Учебно-методическое пособие содержит четыре лабораторные работы. Рассматриваются примеры автоматизации статистических вычислений с помощью пакетов Excel, Statistica. Приводятся типовые задачи с использованием компьютерных технологий, а также перечень вариантов задач для самостоятельного выполнения. Учебное пособие может быть использовано для самостоятельной работы студентами как заочного, так и дневного отделения специальности 1-54 01 01, изучающих дисциплину «Прикладная математика».

УДК 51-7(075.8)
ББК 22.1я7

ISBN 978-985-583-757-3

СОДЕРЖАНИЕ

Введение

Данное методическое пособие отражает опыт авторов при проведении лабораторных и чтении спецкурсов для студентов специальности «Метрология, стандартизация и сертификация». Пособие содержит лабораторных работы, которые должны содействовать знакомству со статистическими методами и алгоритмами анализа данных, выработке навыков использования современных информационных технологий

Для выполнения работ предлагается использовать пакеты Statistica и Excel. Пакет MS Excel позволяет решать простые прикладные задачи, но более универсальной программой для проведения инженерных расчетов, по мнению специалистов, является пакет прикладных программ Statistica. Причем данные из пакета Excel можно легко импортировать в пакет Statistica.

Изложение материала в пособие ведется на уровне, который предусматривает знание элементов математической статистики. Каждая работа содержит краткие теоретические сведения, основы работы с пакетами, практическую часть, задание для самостоятельной работы. В ходе выполнения каждой лабораторной работы студент оформляет и сдает отчет.

Пособие содержит четыре лабораторных работы. В первой работе описываются основные распределения непрерывных случайных величин. В остальных работах рассматриваются выборочный метод статистического анализа, расчет выборочных характеристик, проверка статистических гипотез о законе распределения нормальной совокупности и параметров нормального распределения.

Пособие предназначено для студентов специальности 1-54 01 01 дневного обучения приборостроительного факультета БНТУ, но может быть использовано студентами и магистрантами технических вузов для первоначального знакомства с инструментами статистического анализа.

Лабораторная работа № 1

ВЕРОЯТНОСТНЫЕ РАСПРЕДЕЛЕНИЯ. ЗНАКОМСТВО С ПАКЕТОМ STATISTICA

1.1. Краткие теоретические сведения

1. Равномерное распределение. Непрерывная случайная величина X называется *равномерно распределенной на интервале* $[a;b]$, если ее плотность вероятности равна константе $\frac{1}{b-a}$ на этом интервале и нулю вне его:

$$f(x) = \begin{cases} 0, & x < a; \\ \dfrac{1}{b-a}, & a \le x \le b; \\ 0, & x > b. \end{cases}$$

Соответствующая функция распределения вероятностей имеет следующий вид:

$$F(x) = \begin{cases} 0, & x < a; \\ \dfrac{x-a}{b-a}, & a \le x \le b; \\ 0, & x > b. \end{cases}$$

2. Экспоненциальное (показательное) распределение. Случайная величина X распределена по экспоненциальному закону, если ее плотность распределения вероятностей определяется формулой:

$$f(x) = \begin{cases} 0, & x < 0; \\ \lambda \cdot e^{-\lambda x}, & x \ge 0, \end{cases}$$

где $\lambda = \text{const}$, $\lambda > 0$ – параметр распределения.

Функция распределения показательного распределения имеет вид:

$$F(x) = \begin{cases} 0, & x < 0; \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Для вычисления значений плотности и функции экспоненциального распределения в Excel используется встроенная статистическая функция ЭКСП.РАСП.

Синтаксис:

ЭКСП.РАСП (X; Лямбда; Интегральная).

Здесь X – значение аргумента функции; лямбда – значение параметра; интегральная – логическое значение, которое определяет форму функции. При 0 рассчитывается плотность, а при 1 рассчитывается функция распределения.

3. Нормальное распределение.

Нормальным распределением (или *законом Гаусса*) называется распределение непрерывной случайной величины X, *плотность* которой определяется по формуле:

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

где a и σ – параметры распределения. Можно доказать, что параметр a равен *математическому ожиданию*, а параметр σ – *стандартному отклонению* случайной величины X.

Функция (интегральная) нормального распределения:

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

Для краткой записи нормального распределения с параметрами a и σ используют обозначение $N(a, \sigma)$. В частном случае параметры $a = 0$; $\sigma = 1$. Нормальное распределение $N(0, 1)$ называется *стандартным нормальным распределением*. В этом случае плотность распределения:

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

Функция стандартного нормального распределения иногда называется *функцией Лапласа*, она имеет специальное обозначение:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

Для вычисления значений плотности и функции нормального распределения в Excel используется встроенная статистическая функция НОРМ.РАСП. Чтобы открыть эту функцию необходимо нажать на значок $f(x)$ и в диалоговом окне выбрать «Статистические».

Синтаксис функции:

НОРМ.РАСП (X; Среднее; Стандартное_откл; Интегральная).

Здесь *X* – значение аргумента, на основе которого вычисляется нормальное распределение; *Среднее* – среднее значение распределения; *Стандартное откл* – стандартное отклонение распределения; при *Интегральной* = 0 рассчитывается функция плотности, а при 1 рассчитывается функция распределения.

Рассмотрим теперь другие виды законов распределений.

Распределение хи-квадрат определяется как сумма k независимых стандартных нормальных величин. Число k называется *числом степеней свободы*. Когда $k = 1$, случайная величина равна квадрату стандартной нормальной величины. Хи-квадрат распределение имеет только один параметр – число степеней свободы k, являющийся целым положительным числом. Функция, возвращающая значение плотности распределения хи-квадрат, находится в категории «*Статистические*» и называется ХИ2.РАСП (или ХИ2.РАСП.ПХ).

t-распределение Стьюдента важно в тех случаях, когда рассматриваются оценки среднего, оценки коэффициентов регрессионного уравнения, оценки параметров временных рядов. *t*-распределение Стьюдента с единственным параметром *k*, называемым *степенью свободы*, сосредоточено на всей действительной оси, симметрично относительно начала координат. Функция, возвращающая значение плотности распределения Стьюдента, находится в категории *Статистические* и называется СТЬЮД.РАСП. Функция имеет дополнительный чисто вычислительный параметр «Хвосты», который связан не с распределением Стьюдента, а с выводом полученных результатов программой *Excel.* Его всегда задаем равным 1.

F-распределение Фишера возникает в регрессионном, дисперсионном, дискриминантном анализе, а также в других видах многомерного анализа данных. Случайная величина, имеющая *F*-распределение с парой степеней свободы k_1 и k_2, определяется как отношение двух независимых случайных величин, имеющих распределение хи-квадрат со степенями свободы k_1 и k_2 с умножением на нормированный сомножитель k_2/k_1. *F*-распределение сосредоточено на положительной полуоси. Это распределение несимметрично. Функция, возвращающая значение плотности распределения Фишера находится в категории *Статистические* и называется F.РАСП.

4. Основы работы в пакете Statistica.

Введение в пакет Statistica. Универсальная интегрированная система, предназначенная для статистического анализа, визуализации данных и разработки пользовательских приложений *Statistica* – это современный пакет, в котором реализованы все новейшие компьютерные и математические методы статистического анализа данных.

Прикладное окно пакета *Statistica* имеет стандартную для окна приложения MS Windows структуру (рис. 1.1). Верхняя строка является заголовком окна и содержит имя приложения *Statistica* и имя активного модуля, например, справа расположены кнопки для изменения размеров окна. Вторая строка содержит команды прикладного меню, которое включает набор стандартных для приложения групп команд: *File, Edit, View, Options, Window, Help*, а также группы команд пакета: *Statisctics* (подключение статистических модулей), *Graphs* (команды для работы с графиками), *Data Mining* (сбор данных), *Tools* (инструменты), *Data* (управление табличными данными). По кнопке *Help* можно получить описание устройства и функций пакета как программной системы.

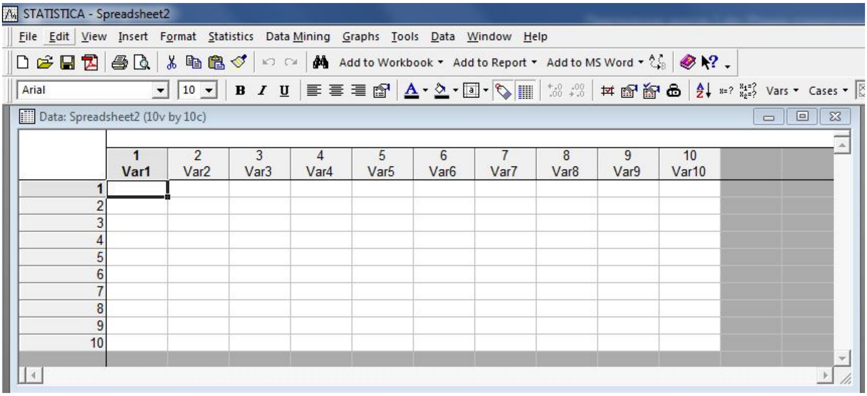


Рис. 1.1. Рабочее окно пакета *Statistica*

Третья и четвертая строки образуют панель инструментов (Toolbar), содержащую кнопки быстрого вызова команд меню.

Основным является рабочее окно, в котором вводятся исходные данные и выводятся результаты их статистической обработки в табличном или графическом виде. Ввод данных осуществляется в табличном виде.

Набор данных в пакете *Statistica* – это прямоугольная таблица, столбцам которой соответствуют обрабатываемые *переменные* (*Variables*), а строкам отвечают наблюдения (*Cases*) значений переменных. Для создания нового набора данных необходимо, прежде всего, завести файл с трафаретами таблицы нужных размеров. Для этого по меню *File – New* через раскрывшееся диалоговое окно необходимо выбрать нужное количество столбцов (*Variables*) и строк (*Cases*). При нажатии опции *Insert* в основном меню или кнопки *Vars* на панели инструментов становятся доступными команды редактирования переменных (столбцов): *Add* (добавить новые переменные); *Delete* (удалить переменные); *Move* (переместить) и др. При нажатии кнопки *Cases* становятся доступными аналогичные команды редактирования строк.

Файлы проекта. При работе с пакетом *Statistica* образуется пять различных типов документов: рабочая книга (*Workbook*), рабочий лист – мультимедийная таблица (*Spreadsheet*), отчет (*Report*), графическая область (*Graph*) и макрокоманда (*Macros*) для языка *Statistica Visual Basic*.

Рабочая книга представляет собой упорядоченный вывод данных, объединяя в себе рабочие листы и графики. Каждый документ представляет собой таблицу. Файл рабочей книги имеет расширение *.*stw*.

Рабочие листы пакета *Statistica* предназначены для ввода данных в числовом или текстовом формате, имеют расширение *.*sta*. Форматом рабочего листа является двумерная таблица с неограниченным количеством наблюдений (строк) и переменных (столбцов), каждый из которых содержит неограниченное количество символов. В рабочий лист могут быть внедрены звук, видеодокументы, графика, анимация.

Отчеты *Statistica* позволяют организовать вывод данных в текстовом формате, более удобном для вывода документов на печать. По умолчанию файл отчета имеет расширение *.*str*, но существует возможность преобразования отчета в стандартный файл формата RTF.

Вся графическая обработка данных сохраняется в отдельных файлах с расширением *.*stg*. При этом поддерживается внедрение графических объектов из других программ.

Макрокоманда представляет собой программный код, написанный на языке *Visual Basic.*

Процедура Probability Calculator. Расчет квантилей. Построение графиков плотности и функции распределения. Задача формулируется следующим образом: определить вероятность того, что случайная величина не превысит заданного значения; найти значение случайной величины, для которого функция распределения равна заданному значению p. Искомое значение случайной величины называется *квантилью*, соответствующей вероятности p.

Для работы с распределениями непрерывных случайных величин в пакете *Statistica* используется калькулятор вероятностных распределений. Для его вызова выполняется процедура *Probability Calculator* (рис. 1.2).

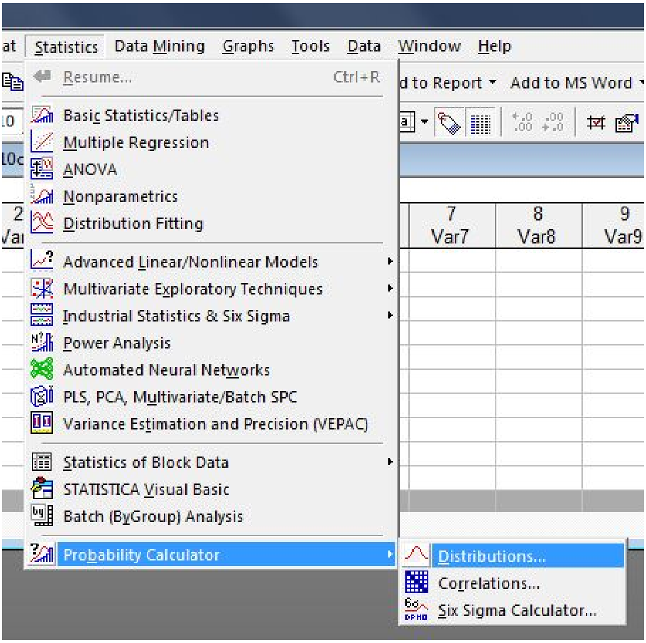


Рис. 1.2. Вызов процедуры *Probability Calculator*

С помощью вероятностного калькулятора могут решаться разнообразные вероятностные задачи, например, построение графиков плотностей и функций распределения, определение квантили для заданной вероятности и пр.

При запуске процедуры *Probability Calculator* открывается окно *Probability Distribution Calculator*, представленное на рис. 1.3. В левой части данного окна содержится список распределений непрерывных случайных величин, доступных пользователю. В средней части в строке p содержится вероятность $P(X \le x) = F(x)$, в строке Beta – значение случайной величины или квантиль, соответствующая этой вероятности. Далее содержатся поля для ввода параметров распределения. В верхней части окна содержатся опции, предназначенные для настройки режимов работы процедуры. Назначение этих опций следующее: *Inverse* – работа с обратной функцией распределения, которая используется при вычислении квантили; *Two-tailed* – построение двустороннего интервала для плотности распределения; *1-Cumulative p* – использование вместо вероятности p значения вероятности $1-p$; *Create Graph* – создание графика. Пользователь может выбрать закон распределения и для заданного значения случайной величины вычислить значение функции распределения или, наоборот, задать вероятность p и определить соответствующую ей квантиль $F^{-1}(x)$. В последнем случае следует переключить опцию *Inverse*.

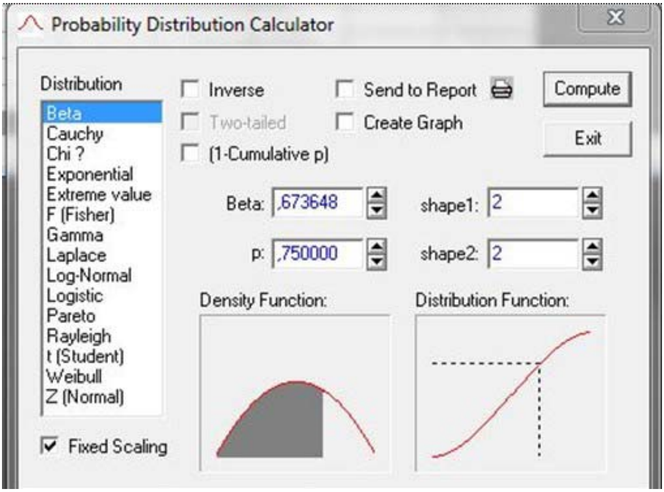


Рис. 1.3. Окно процедуры *Probability Calculator*

В случае работы с дискретными распределениями для расчета вероятностей и функций распределения применяют встроенные функции.

В пакете *Statistica* имеются встроенные функции для всех основных законов распределения, причем функции пакета, имена которых начинаются с буквы *I*, вычисляют значения функций распределения.

Для вычисления встроенной функции в пакете *Statistica* следует выделить незаполненный столбец таблицы *Spreadsheet* исходных данных и затем выполнить команду *Variable Specs*, нажав правую кнопку мыши. В нижней части открывшегося окна спецификации переменной находится рабочая область *Long name* (*label, link or formula with function*), которая предназначена для ввода выражений и комментариев. Набор формулы следует производить, начиная со знака равно (=), далее встроенные функции могут быть набраны непосредственно с клавиатуры либо с помощью конструктора (кнопка *Function*).

В последнем случае открывается окно *Spreadsheet Formulas*, в котором содержится список встроенных функций. Вставка необходимой функции производится с помощью двойного щелчка мыши по имени функции. Далее задаются фактические параметры распределения и значения аргумента, в качестве последних могут использоваться имена переменных (столбцов) или константы.

1.2. Практическая часть

Контрольный пример 1.1. Построить графики плотности распределения $f(x)$ и функции распределения $F(x)$ нормально распределенной случайной величины при x, меняющемся от 0 до 6 с шагом 0,2, математическим ожиданием $a = 3$ и стандартным отклонением $\sigma = 1$. Проанализировать влияние параметров a и σ на график. Задание выполнить в пакете *Excel*.

Решение. Запускаем программу *Excel* и задаем значения параметров a и σ. Для этого в ячейки A1 и A2 первого листа вводим подписи «a=» и «sig=» (кавычки здесь и далее вводить не надо), а в соседние B1 и B2 вводим значения 3 и 1. Для построения графиков протабулируем в столбцах C, D и E функции $f(x)$ и $F(x)$ на отрезке (0; 6) с шагом 0,2. Для этого вводим в C1 подпись «X», в D1 подпись «f» и в E1 – подпись «F». Вводим в C2 значение 0, в C3 – значение 0,2, обводим, выделяя, ячейки C2 и C3 и, захватив нижний правый угол рамки вокруг ячеек C2 и C3, перетягиваем его вниз до ячейки C32, что позволит автоматически занести в столбец значения от 0 до 6 с шагом 0,2. Ставим курсор в ячейку D2 и вызываем функцию плотности нормального распределения. Для этого нажимаем кнопку мастера функций *fx,* выбираем категорию «*Статистические*» и функцию НОРМ.РАСП.

Появляется окно, показанное на рис. 1.4.

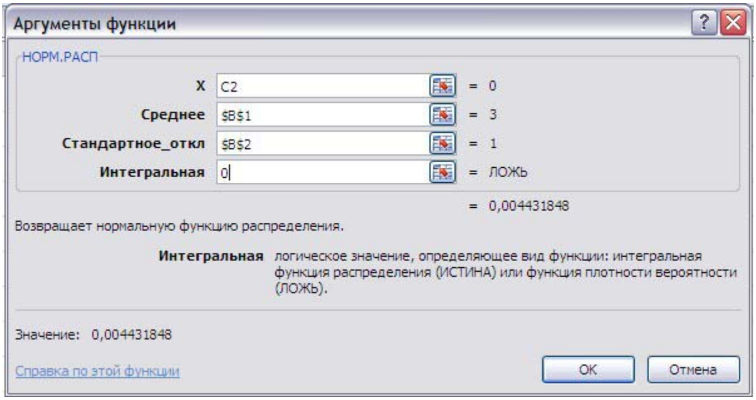


Рис. 1.4. Диалоговое окно функции НОРМ.РАСП

Вводим ссылкой на переменную *X*: «C2» (для ввода ссылки достаточно щелкнуть мышью по ячейке с данной адресацией), ссылкой на a и σ – «B1» и «B2». Эти ссылки абсолютные, т. к. ячейки со значениями a и σ – всегда B1 и B2, поэтому пишется знак $ (чтобы быстро относительную ссылку сделать абсолютной нужно после ввода ссылки нажать F4). В поле «*Интегральная*» ставим ноль, нажимаем «ОК». В ячейке D2 появился результат 0,004432, а в строке формул – запись «=НОРМ.РАСП(C2; B1; B2; 0)». Захватив нижний правый угол ячейки D2, автозаполнением растягиваем результат на ячейки D2:D32 (см. рис. 1.5).

	A	B	C	D	E
1	a=	3	X	f(x)	F(x)
2	sig=	1	0	0,00443	0,00135
3			0,2	0,00792	0,00256
4			0,4	0,01358	0,00466
5			0,6	0,02239	0,0082
6			0,8	0,03547	0,0139
7			1	0,05399	0,02275
8			1,2	0,07895	0,03593
9			1,4	0,11092	0,0548
10			1,6	0,14973	0,08076
11			1,8	0,19419	0,11507

Рис. 1.5. Исходные данные

Столбец E заполняем аналогично – в поле «Интегральная» нужно поставить 1.

Строим график по данным. Выделяем диапазон ячеек C1:E32. Вызываем мастер диаграмм, выбрав пункты меню *Вставка*/*Диаграммы*. Выбираем тип диаграммы «*Точечная*» и вид – «*Точечная с гладкими кривыми*». Получаем графики плотности и функции нормального распределения (см. рис. 1.6).

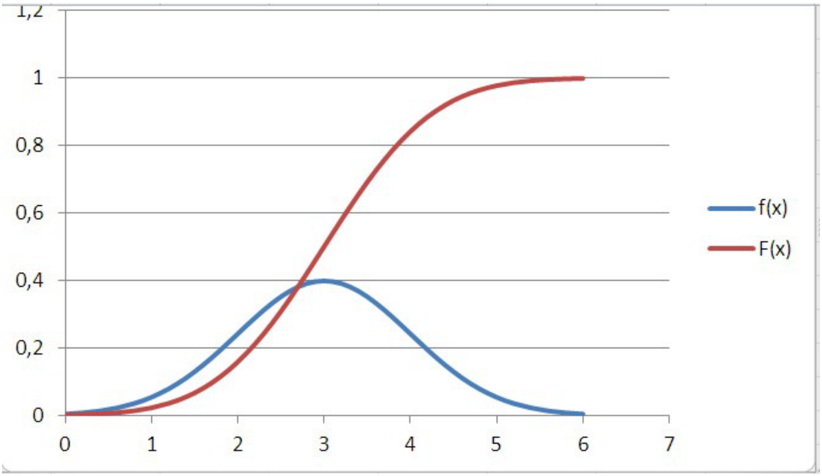


Рис. 1.6. Графики плотности и функции нормального распределения

Исследуем, как влияют параметры на вид графика $f(x)$. Для этого изменяем в ячейке B1 значение 3 на значение 4, нажимаем *Enter*. Видим, что график сместился вправо, изменяем на 2, график сместился влево. Возвращаем в B1 значение 3, и изменяем в B2 значение 1 на 2. График растянулся. Изменяем в B2 на 0,5 – график сжался.

Делаем вывод: параметр a изменяет положение графика, с увеличением параметра график смещается вправо; параметр σ влияет на ширину графика, с увеличением параметра график растягивается.

Контрольный пример 1.2. Исследование основных распределений в пакете *Statistica*.

1.2.1. *Нормальное распределение.*

Запустим процедуру *Probability Calculator*. Выполним расчеты для нормального распределения со средним значением $a = 0{,}5$

и $\sigma = 1$. Первая задача будет заключаться в поиске квантили для вероятности $p = 0{,}8$ и построении графиков плотности и функции распределения.

В окне калькулятора в соответствующие поля введем параметры распределения, вероятность p и отметим опции *Inverse* и *Create Graph* (рис. 1.7).

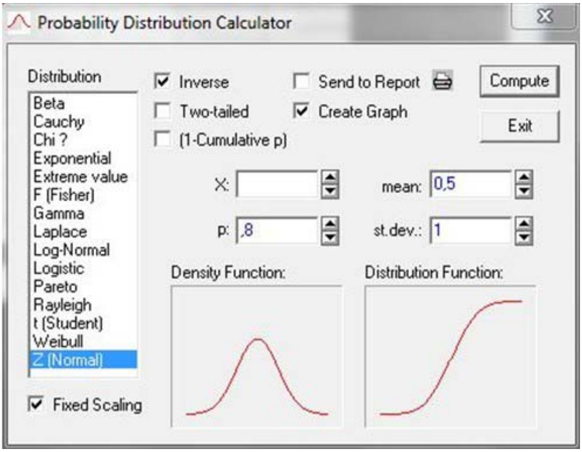


Рис. 1.7. Окно с исходными данными для расчета квантили

После нажатия кнопки *Compute* получим результат (рис. 1.8). Значение квантили для заданной вероятности равно 1,341612.

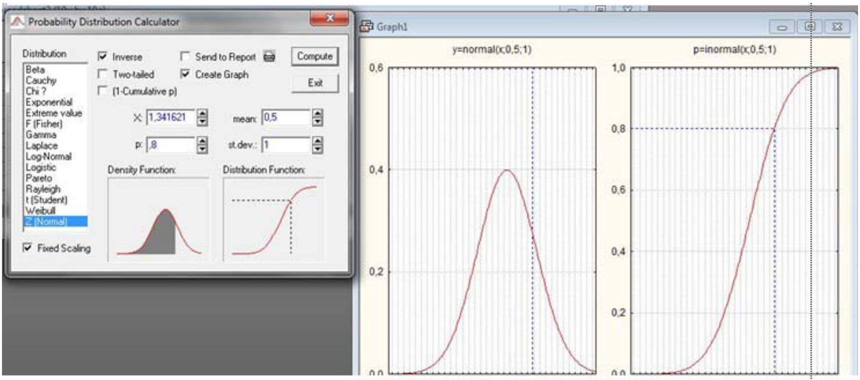


Рис. 1.8. Графики плотности и функции распределения для нормального распределения

Следующая задача заключается в определении значения функции распределения для заданного значения случайной величины $x = 1$. Введя в поле X значение, равное 1, и нажав кнопку *Compute*, в поле p получим значение функции распределения, равное 0,691462 (рис. 1.9).

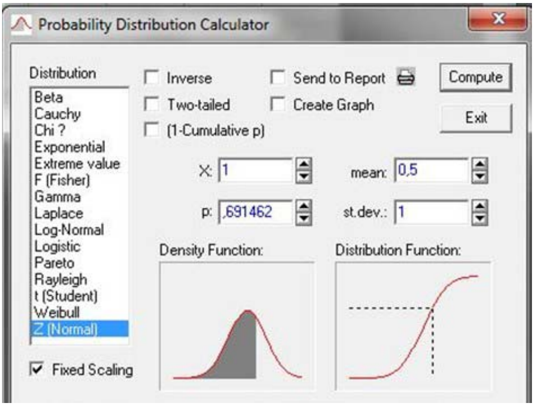


Рис. 1.9. Результаты расчета функции распределения для $x = 1$

1.2.2. Аналогичные действия выполним для распределений χ^2, *Фишера и Стьюдента*.

Для распределения χ^2 с числом степеней свободы $df = 6$ и вероятностью $p = 0{,}8$ определена квантиль *Chi_I* = 8,558060 и построены графики плотности и функции распределения (рис. 1.10).

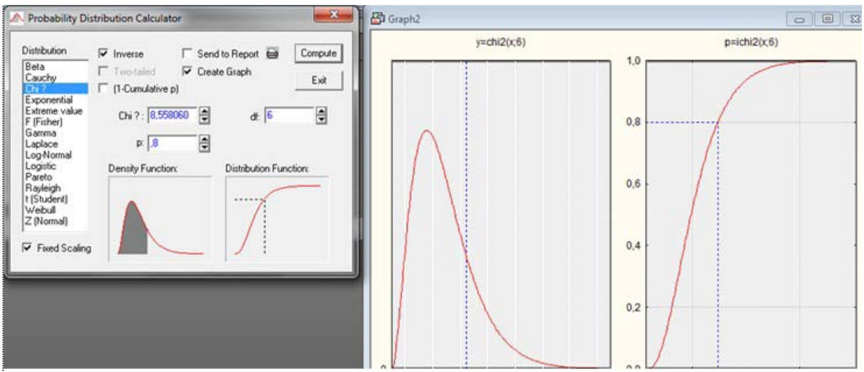


Рис. 1.10. Результаты расчета квантили и графики плотности и функции распределения для распределения χ^2

Для распределения Фишера со степенями свободы $k_1 = 4$ и $k_2 = 15$ результаты расчетов представлены на рис. 1.11. В поля *df1* и *df2* введены соответственно значения k_1 и k_2.

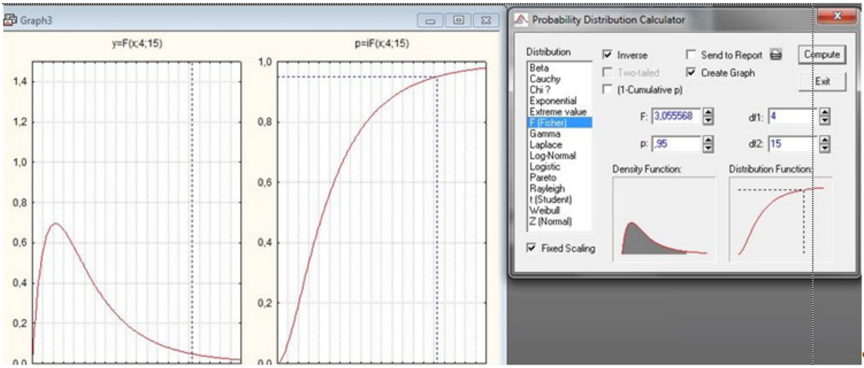


Рис. 1.11. Результаты расчета квантили и графики плотности и функции распределения для распределения Фишера

Для распределения Стьюдента с числом степеней свободы $df = 10$ результаты расчетов представлены на рис. 1.12. Квантиль этого распределения, соответствующая вероятности p = 0,95, равна $t = 1{,}812461$.

Графики плотности вероятности и функции распределения представлены на рис. 1.12.

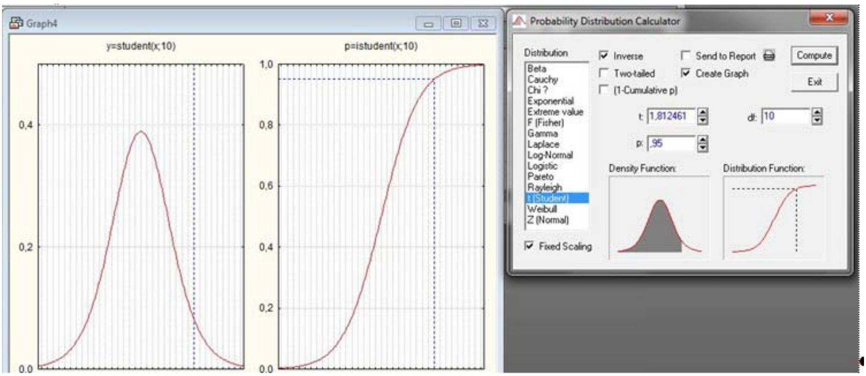


Рис. 1.12. Результаты расчета квантили и графики плотности и функции распределения для распределения Стьюдента

1.3. Задания для самостоятельной работы

Задание 1. В пакете *Excel:*

1. Ввести в таблицу значения аргумента x в диапазоне от x_1 до x_2 с шагом h.

2. Вычислить значение плотности стандартного нормального распределения, а также плотности нормального распределения с параметрами a и σ (табл. 1.1).

3. Используя мастер диаграмм, построить соответствующие кривые распределения.

4. Для заданных параметров нормального распределения построить семейство графиков функции распределения.

5. В пакете *Statistica*, используя вероятностный калькулятор, рассчитать квантили нормального распределения с параметрами a и σ из табл. 1.1 для $p = 0{,}8; 0{,}9; 0{,}95; 0{,}99$, а также построить графики $f(x)$ и $F(x)$.

Задание 2. Выполнить контрольный пример 1.2. Используя вероятностный калькулятор:

1. Составить таблицы нормального, хи-квадрат, Стьюдента и Фишера распределений (по 10 значений). Вычислить 0,95 и 0,99 квантили модельных распределений для различных значений параметров.

2. Проанализировать влияние параметров распределения на форму кривых плотностей для следующих непрерывных распределений: хи-квадрат, Стьюдента, экспоненциального, нормального, Фишера.

Таблица 1.1

Вариант	x_1	x_2	h	a	σ
1	–3	7	0,5	2	3
2	0	2	0,1	1	1
3	–4	4	0,4	0	2
4	–4	7,4	0,6	1	2
5	1	3	0,1	2	0,5
6	–1	3	0,2	0,2	1
7	2	4	0,1	3	0,6
8	–2	4	0,3	1	0,7
9	0,3	8,3	0,4	4	0,2
10	0	4	0,2	2	0,8

Лабораторная работа № 2

ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ ВАРИАЦИОННОГО РЯДА

2.1. Краткие теоретические сведения

Генеральной совокупностью называется совокупность объектов произвольной природы, обладающих признаками, доступными для наблюдения и количественного измерения.

Объекты, входящие в генеральную совокупность, называются ее *элементами*, а их общее число N – ее *объемом*.

Однако получение экспериментальных данных – достаточно трудоемкий, дорогой процесс, а в некоторых случаях и просто невозможный. Поэтому из всей генеральной совокупности приходится выбирать только определенную часть объектов, которую называют *выборочной совокупностью* или *выборкой объема n.*

Предположим, что над случайной величиной X производится ряд независимых опытов (наблюдений). В каждом из этих опытов случайная величина X принимает определенное значение $x_1, x_2, \ldots, x_n$. Совокупность этих значений рассматривается как простая выборка.

Наблюдаемое значение x_i называют *вариантой*, а их последовательность, записанную в возрастающем порядке, – *вариационным рядом.* Для каждой варианты можно указать *частоту* ее появления, которую обозначают n_i. Также может быть найдена *относительная частота* появления определенной варианты, как отношение частоты к объему выборки: $w_i = {n_i}/{n}$.

Статистическим распределением выборки (*вариационным рядом*) называют соответствие вариант (расположенных в возрастающем порядке) и их частот или относительных частот (табл. 2.1).

Таблица 2.1

Вариационный ряд

Варианта	x_1	x_2	…	x_k	$\sum$
Частота	n_1	n_2	…	n_k	n
Относительная частота	w_1	w_2	…	w_k	1

Если каждую пару (x_i, n_i) изобразить точкой на координатной плоскости и соединить эти точки ломаной линией, то будет получен *полигон частот*. Ломаная, соединяющая точки (x_i, w_i), называется *полигоном относительных частот*. Это аналог многоугольника распределения.

При большом числе наблюдений статистический ряд перестает быть удобной формой записи статистического материала – он становится громоздким и мало наглядным. Для придания ему большей компактности и наглядности строится так называемый *интервальный статистический (вариационный) ряд* (табл. 2.2). В этом случае весь диапазон наблюдаемых значений X разделяется на интервалы и подсчитывается количество значений n_i, W_i, приходящееся на каждый интервал.

Таблица 2.2

Интервальный статистический ряд

Интервал	$[x_1, x_2)$	$[x_2, x_3)$	…	$[x_{l-1}, x_l]$	$\sum$
Частота	n_1	n_2	…	n_l	n
Относительная частота	w_1	w_2	…	w_l	1

Длину интервала h проще выбирать одинаковой. Для нахождения длины интервала можно воспользоваться следующей формулой:

$$h = \frac{R}{l} = \frac{x_{\max} - x_{\min}}{l}. \qquad (2.1)$$

Здесь l – количество интервалов, вычисляемое по формуле $l = 1 + 3{,}322 \cdot \log n = \log_2 n + 1$; R – размах выборки.

Если в результате вычисления по формуле (2.1) длина интервала получится дробным числом, то выбирают либо близкое целое число, либо близкую простую дробь.

По этим данным можно построить гистограммы частот и относительных частот. *Гистограммой* называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которой служат частичные интервалы шириной h, а высоты равны ${n_i}/{h}$ (для частот) или ${w_i}/{h}$ (для относительных частот).

Гистограмма относительных частот является аналогом функции плотности распределения случайной величины.

Выборочной (эмпирической) функцией распределения называется функция $F^*(x) = {n_x}/{n}$, где n_x – число значений случайной величины, меньших x, а n – объем выборки.

При большом числе наблюдений эмпирическая функция распределения приближается к теоретической интегральной функции распределения генеральной совокупности.

Эмпирическая функция распределения обладает всеми свойствами теоретической функции распределения.

Кумулятивная кривая (кумулянта) – ломаная, соединяющая точки с координатами $\left(x_i, n_{x_i}\right)$ или $\left(x_i, \dfrac{n_{x_i}}{n}\right)$, где n_{x_i} – накопленные частоты; для интервального ряда n_{x_i} – число вариант меньших значений вариант i-го интервала.

Накопленная частота (частость) равна сумме всех частот (относительных частот) вариант, предшествующих данному значению. Накопленная частота характеризует число членов данной совокупности, в которых признак, интересующий нас, меньше данного значения.

2.2. Практическая часть

Контрольный пример 2.1. Дана выборка значений массы тела учащихся в килограммах:

Таблица 2.3

Эмпирические данные о массе тела учащихся

Наблюдения				
64	62	58	58	61
57	62	63	63	58
63	60	61	61	60
62	64	59	59	64
58	61	62	62	60
61	59	60	60	59
63	59	60	60	61
60	63	58	58	64
60	61	61	61	62
61	62	60	60	59
65	58	63	63	65

Построить дискретный вариационный ряд, провести интервальную обработку. Построить полигон частот, гистограмму относительных частот и кумулятивную кривую.

Задания выполнить в пакетах *Statistica* и *Excel*.

Решение.

1. Упорядочим выборку, т. е. составим вариационный ряд:

57	57	58	58	58	58	58	58	59	59	59
59	59	59	59	60	60	60	60	60	60	60
60	60	60	61	61	61	61	61	61	61	61
61	62	62	62	62	62	62	62	62	63	63
63	63	63	63	63	64	64	64	64	65	65

Дискретный вариационный ряд имеет вид табл. 2.4.

Таблица 2.4

x_i	57	58	59	60	61	62	63	64	65	$\sum$
n_i	2	6	7	10	9	8	7	4	2	**55**
w_i	$\frac{2}{55}$	$\frac{6}{55}$	$\frac{7}{55}$	$\frac{10}{55}$	$\frac{9}{55}$	$\frac{8}{55}$	$\frac{7}{55}$	$\frac{4}{55}$	$\frac{2}{55}$	**1**

Проведем интервальную обработку.

По условию, объем выборки $n = 55$. Определим оптимальную длину частичного интервала:

$$h = \frac{x_{\max} - x_{\min}}{l} = \frac{x_{\max} - x_{\min}}{1 + 3,322 \lg n} = \frac{65 - 57}{1 + 3,322 \lg 55} \approx \frac{8}{7} \approx 1,15.$$

Тогда:

$x_1 = 57;\ \ x_2 = 57 + 1,15 = 58,15;\ \ x_3 = 58,15 + 1,15 = 59,3;$

$x_4 = 59,3 + 1,15 = 60,45;\ \ x_5 = 60,45 + 1,15 = 61,6;$

$x_6 = 61,6 + 1,15 = 62,75;\ \ x_7 = 62,75 + 1,15 = 63,9;$

$x_8 = 63,9 + 1,15 = 65,05.$

Вариационный ряд представим в виде табл. 2.5.

Таблица 2.5

Интервал						
57 – 58,15	58,15 – 59,3	59,3 – 60,45	60,45 – 61,6	61,6 – 62,75	62,75 – 63,9	63,9 – 65,05
57	59	60	61	62	63	64
57	59	60	61	62	63	64
58	59	60	61	62	63	4
58	59	60	61	62	63	65
58	59	60	61	62	63	65
58	59	60	61	62	63	
58	59	60	61	62	63	
58		60	61	62		
		60	61			
		60				

Тогда интервальный статистический ряд примет вид табл. 2.6.

Таблица 2.6

Интервальный статистический ряд

x_i	x_{i+1}	x_i^*	n_i	$\frac{n_i}{n}$	Накопленная частота
57	58,15	58,575	8	8/55	8/55
58,15	59,3	58,725	7	7/55	15/55
59,3	60,45	59,875	10	10/55	25/55
60,45	61,6	61,025	9	9/55	34/55
61,6	62,75	62,175	8	8/55	42/55
62,75	63,9	63,325	7	7/55	49/55
63,9	65,05	64,475	6	6/55	1
$\sum$			**55**	**1**	

2. В пакете *Excel* в ячейку A1 введем слово «Наблюдения» (рис. 2.1), а в диапазон A2:A56 – эмпирические данные, приведенные в табл. 2.3.

	A	B	C	D	E	F	G	H
1	Наблюдения		Максимум	65	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	64		Минимум	57	57	2	0,0364	0,0364
3	57				58	6	0,1091	0,1455
4	63				59	7	0,1273	0,2727
5	62				60	10	0,1818	0,4545
6	58				61	9	0,1636	0,6182
7	61				62	8	0,1455	0,7636
8	63				63	7	0,1273	0,8909
9	60				64	4	0,0727	0,9636
10	60				65	2	0,0364	1,0000
11	61			Всего наблюдений		55	1	
12	65							
13	62							

Рис. 2.1. Результат вычислений абсолютных, относительных и накопленных частот

Рассчитаем максимальное и минимальное значения выборочных данных в ячейках D1 и D2, введя соответственно функции МАКС(A2:A56) и МИН(A2:A56) (рис. 2.1).

Построим вариационный ряд, считая массу тела дискретной случайной величиной. В ячейку E1 введем заголовок «Варианты», а ниже в столбце – все возможные неповторяющиеся значения массы тела учащихся (x_i), которые встречались в выборке (от минимального до максимального).

В ячейке F1 запишем заголовок «Абсолютные частоты». В этом столбце будут рассчитаны значения частот n_i, т. е. то количество раз, которое соответствующее значение x_i случайной величины встречалось в выборке. Для заполнения столбца абсолютных частот можно использовать стандартную функцию ЧАСТОТА().

Выделим мышью диапазон F2:F10, в котором разместятся найденные частоты, вызовем *Мастер функций* и в категории *Статистические* выберем функцию ЧАСТОТА. После этого заполним ее аргументы:

– *массив данных* – это диапазон эмпирических данных A2:A56;

– *массив интервалов* – это диапазон значений вариант E2:E10.

Закончить ввод функции нужно одновременным нажатием клавиш *Ctrl* + *Shift* + *Enter*, поскольку ее результатом является диапазон значений. В строке формул эта функция будет показана в фигурных скобках.

В ячейке F11 найдем общее число наблюдений, просуммировав значения в столбце абсолютных частот (см. рис. 2.1).

В ячейке G1 запишем заголовок «*Относительные частоты*». Для расчета относительных частот $w_i = \dfrac{n_i}{n}$ внесем в ячейку G2 формулу $= \mathrm{F2} / \$\mathrm{F}\$11$ и скопируем ее методом автозаполнения вниз по столбцу. Сумма относительных частот в этом столбце должна быть равна единице.

Последний столбец таблицы озаглавим «*Накопленные частоты*». В ячейку H2 скопируем значение относительной частоты из ячейки G2, а в ячейку H3 введем формулу $= \mathrm{H2} + \mathrm{G3}$. Методом автозаполнения скопируем введенную формулу вниз по столбцу в диапазон H4:H10.

Итоговый вид таблицы после форматирования показан на рис. 2.1.

Построим полигон частот по данным в столбце «Абсолютные частоты», как показано на рис. 2.2 (используем диаграмму типа «точечная с прямыми отрезками и маркерами»).

Рис. 2.2. Полигон частот

Построим также совместную диаграмму относительных и накопленных частот. Чтобы совместить в диаграмме несколько типов, (например, Гистограмму и Линейный график – как в нашем примере), необходимо сначала построить все диаграммы одного вида. Выделим диапазон G2:H10 и на вкладке «*Вставка*» выберем тип *Гистограмма.*

Теперь выбираем один ряд и для него меняем тип диаграммы. Щелкнув на ряде 2 правой кнопкой мыши, выбираем «*Изменить тип диаграммы для ряда*» и тип «*График*».

Так как гистограмма получилась незаметной на диаграмме, нужно добавить вспомогательную ось. Для этого нажмем правой кнопкой мыши на гистограмму или на название в легенде. Далее в появившимся диалоговом окне выберем «*Формат ряда данных»*. В открывшемся окне ищем *Параметры ряда* и меняем галочку на «*По вспомогательной оси»*. После минимального редактирования диаграмма будет иметь такой вид, как показано на рис. 2.3.

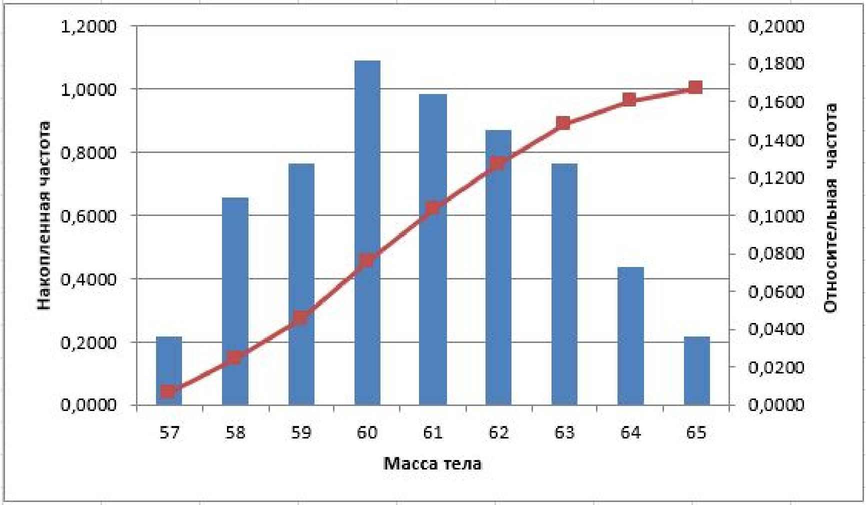


Рис. 2.3. Диаграмма относительных и накопленных частот

Гистограмма относительных частот есть аналог функции плотности распределения, а график накопленных частот проходит через левые концы «ступенек» эмпирической функции распределения и называется *кумулятивной кривой*.

Для построения интервального вариационного ряда разобьем диапазон наблюдавшихся значений [57; 65] на интервалы шириной 1,15. При этом минимальное значение должно попасть внутрь интервала. Данные в столбце «*Варианты*» (рис. 2.4) интерпретируются как правые границы интервалов. Значения, которые дает функция ЧАСТОТА – это частоты попадания в интервал. При этом если значение случайной величины попадает на границу интервала, то оно учитывается в левом интервале. Что касается самого первого значения в столбце «*Варианты*», то для него функция ЧАСТОТА дает количество наблюдений меньших или равных ему.

Остальные расчеты полностью аналогичны. На рис. 2.4 показан результат вычисления абсолютных, относительных и накопленных частот для интервального ряда.

F2 | {=ЧАСТОТА(A2:A56;E2:E8)}

	A	B	C	D	E	F	G	H
1	Наблюдения		Максимум	65	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	64		Минимум	57	58,15	8	0,1455	0,1455
3	57				59,3	7	0,1273	0,2727
4	63				60,45	10	0,1818	0,4545
5	62				61,6	9	0,1636	0,6182
6	58				62,75	8	0,1455	0,7636
7	61				63,9	7	0,1273	0,8909
8	63				65,05	6	0,1091	1,0000
9	60			Всего наблюдений		55	1	

Рис. 2.4. Результаты расчетов частот для интервального ряда

3. Теперь выполним данные задания в пакете *Statistica*.

После запуска пакета на экране появится сетка-таблица.

Преобразуем таблицу к размерам 1×55, выполнив следующие действия.

Нажимаем кнопку *Vars* (на экране), в раскрывающемся меню выбираем *Delete*; появится окно *Delete Variables*.

Укажем, какие переменные-столбцы убрать.

Нажимаем кнопку *Cases*, выбираем опцию *Add* (добавление), появится окно *Add Cases*: укажем, сколько строк добавить и куда.

Далее выделим столбец – переменную *Var*1 (щелчком правой кнопки мыши по ее заглавию) – в открывшемся меню выберем *Variable Specs* (спецификации переменной) – в появившемся окне *Variable 1* введем *Name* X .

Зададим исходные данные (или скопируем из пакета *Excel*).

В меню *Statistics – Basic Statistics/Tables* в окне *Descriptive Statistic* во вкладке *Quick* нажмем на кнопку *Frequency Table*. В результате получим таблицу частот (рис. 2.5). В первом столбце заданы интервалы для переменной *X*, причем последняя строка содержит пропущенные значения. Второй столбец содержит число попаданий переменной в интервалы (*Count*), третий столбец – накопленную частоту (*Cumulative Count*), четвертый и шестой – частоты в процентных соотношениях для имеющихся в наличии (не пропущенных) наблюдений (*Percent of Valid*) и для всех наблюдений (*% of Cases*), пятый и седьмой столбцы – накопленные частоты в процентах соответственно для (не пропущенных) наблюдений (*Cumul. % of Valid*) и для всех наблюдений (*Cumul. % of All*).

Category	Frequency table: X (Spreadsheet1) K-S d=,11619, p> .20; Lilliefors p<,10 Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
56,00000<x<=57,00000	2	2	3,63636	3,6364	3,63636	3,6364
57,00000<x<=58,00000	6	8	10,90909	14,5455	10,90909	14,5455
58,00000<x<=59,00000	7	15	12,72727	27,2727	12,72727	27,2727
59,00000<x<=60,00000	10	25	18,18182	45,4545	18,18182	45,4545
60,00000<x<=61,00000	9	34	16,36364	61,8182	16,36364	61,8182
61,00000<x<=62,00000	8	42	14,54545	76,3636	14,54545	76,3636
62,00000<x<=63,00000	7	49	12,72727	89,0909	12,72727	89,0909
63,00000<x<=64,00000	4	53	7,27273	96,3636	7,27273	96,3636
64,00000<x<=65,00000	2	55	3,63636	100,0000	3,63636	100,0000
Missing	0	55	0,00000		0,00000	100,0000

Рис. 2.5. Таблица частот

Для построения графика частот (полигона частот) выделим столбец *Percent of valid*, нажмем правую кнопку мыши и в контекстном меню выберем команду *Graph of Block Data – Line Plot: Entire Columns* (так как данные для построения графика расположены в столбцах). В результате получим график, представленный на рис. 2.6.

Рис. 2.6. График, построенный по исходным данным

Для построения гистограммы частот во вкладке *Quick* окна *Descriptive Statistic* нажмем на кнопку *Histograms*. Получим гистограмму, представленную на рис. 2.7.

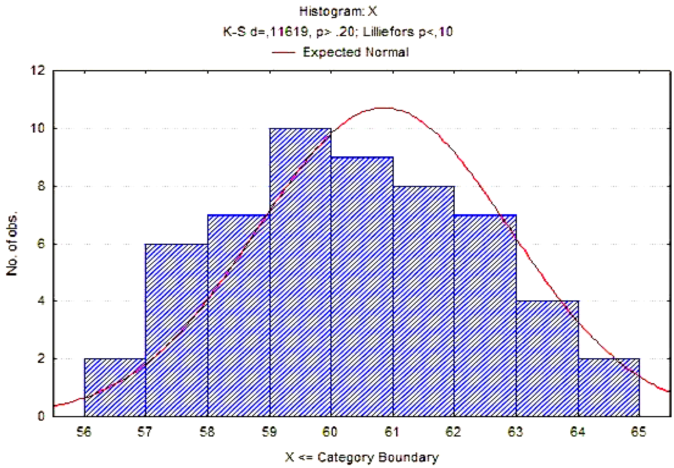


Рис. 2.7. Гистограмма, построенная по исходным данным

Для построения кумулятивной кривой (эмпирической функции распределения) перейдем в окно *Graphs* на вкладку *Histograms*. Во вкладке *Advanced* выберем *Fit type – Off*, установим: *Graph type: Regular, Showing type: Cumulative*, *Variables – X*; *Categories (число интервалов группировки) – 250 – OK.*

Наблюдаем график эмпирической функции распределения (рис. 2.8).

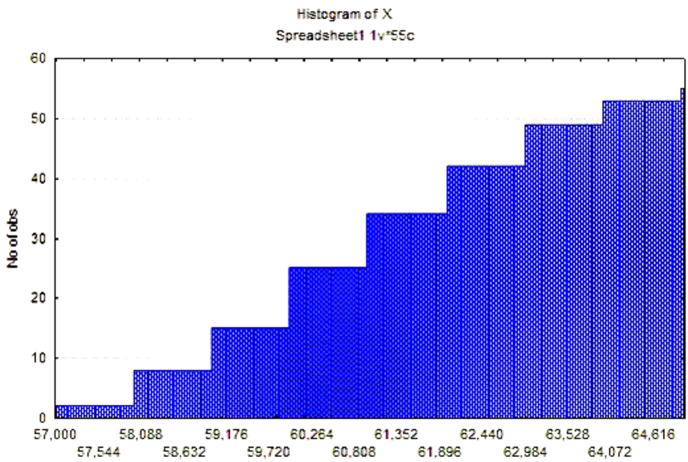


Рис. 2.8. Кумулятивная кривая

Контрольный пример 2.2. Необходимо в пакете *Statistica*:

1) сгенерировать следующие выборки объемом $n = 50$:

$R(2)$; $E(2)$; $N(2;\ 0{,}5)$;

2) построить вариационный ряд и провести группирование данных;

3) построить графики эмпирической функции распределения, а также полигона и гистограммы частот.

Решение.

После запуска пакета *Statistica* на экране появится сетка-таблица. Преобразуем таблицу к размерам 3×50, выполнив следующие действия:

– нажимаем кнопку *Vars* (на экране), в раскрывающемся меню выбираем *Delete*; появится окно окно *Delete Variables*.

– укажем, какие переменные-столбцы убрать (рис. 2.9).

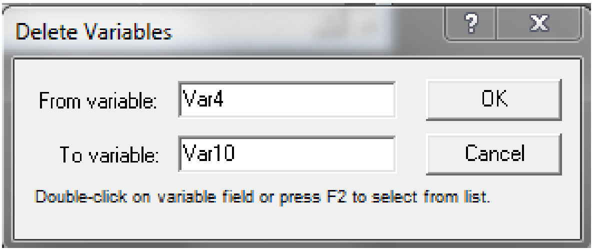


Рис. 2.9. Удаление ненужных переменных

– далее нажимаем кнопку *Cases*, выбираем опцию *Add* (добавление), появится окно *Add Cases*: укажем, сколько строк добавить и куда (рис. 2.10).

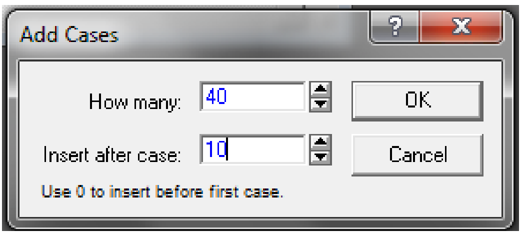


Рис. 2.10. Добавление строк

1. Сгенерируем выборку объема $n = 50$ с равномерным, показательным и нормальным распределениями согласно заданию. Выде-

лим столбец – переменную *Var1* (щелчком мыши по ее заглавию) – нажмем правую клавишу – в открывшемся меню выберем *Variable Specs* (спецификации переменной) – в появившемся окне *Variable 1* введем *Name R*, в нижнем поле *Long Name* вводится выражение, определяющее переменную (см. рис. 2.11). Для задания закона распределения следует ввести:

а) = *rnd*(2) – для случайной величины, распределенной равномерно на отрезке [0; 2];

б) = *VNormal*(*rnd(*1);2;0.5) – для случайной величины, распределенной по нормальному закону с параметрами $a = 2$; $\sigma^2 = 0{,}5^2$;

в) = *VExpon*(*rnd*(1);1/2) – для случайной величины, распределенной по показательному закону с параметром $\lambda = 2$.

Такая форма задания определяется способом генерации: с помощью функции, обратной (буква *V*) к функции распределения, и генератора случайных чисел (*rnd*(1)).

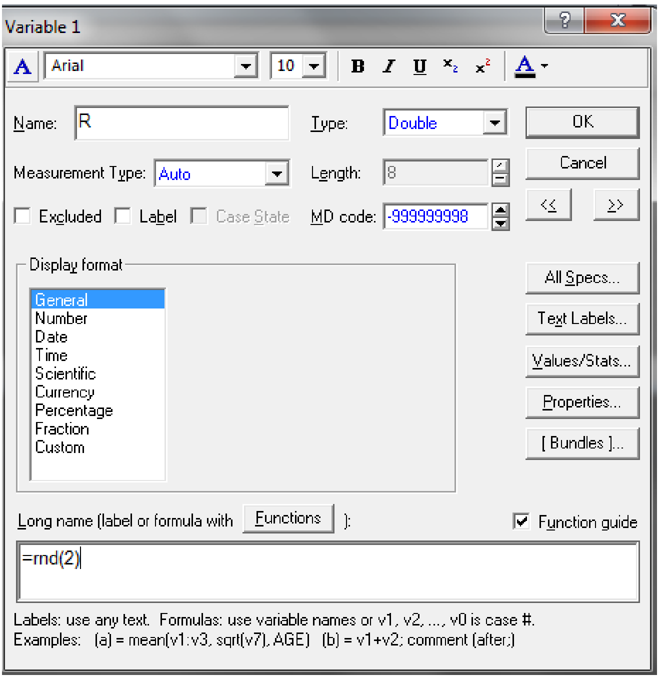


Рис. 2.11. Генерация равномерно распределенной случайной величины

Результат генерации (часть исходных данных) представлен на рис. 2.12.

	1 R	2 E	3 N
1	1,698925	0,382693	1,80235
2	1,203209	4,9181	2,545697
3	1,759833	1,116552	2,419282
4	0,770563	0,754625	2,026829
5	0,960855	1,049295	2,539117
6	0,155122	1,30223	2,091316
7	0,315295	0,675463	0,381524
8	1,592266	0,56612	2,91361
9	0,117848	2,62376	0,762583
10	0,31273	2,138003	2,342175
11	0,224994	3,291386	1,52493
12	0,61113	5,524334	2,330222
13	0,520845	0,931764	1,456521
14	1,635484	1,241677	2,264746
15	1,827226	0,471354	2,789487
16	1,983715	5,350762	1,911672

Рис. 2.12. Результат генерации

2. Для построения вариационного ряда выделим требуемую переменную (столбец) – нажмем правую кнопку мыши – выберем *Sort Cases*. Нажав кнопку *Add Vars*, выбираем переменные для сортировки и тип сортировки*: Ascending* (по возрастанию) – рис. 2.13. Далее нажимаем *OK*.

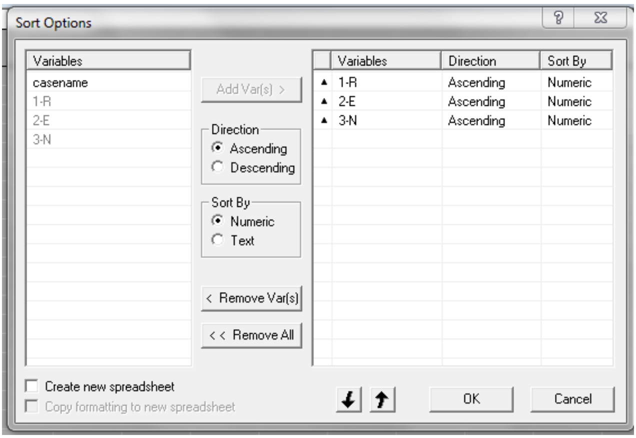


Рис. 2.13. Диалоговое окно для построения вариационного ряда

Для группирования данных (так же, как и в контрольном примере 2.1) выберем в меню *Statistics – Basic Statistic/Tables.* В окне *Descriptive Statistic* во вкладке *Quick* нажмем на кнопку *Frequency Tables*, в появившемся окне выделим необходимый нам столбец переменных (*R, N* или *E*), нажмем *OK*. В результате получим таблицу частот (на рис 2.14 приведена таблица частот для выборки *N*).

Workbook3*
Basic Statistics/Tabl
Descriptive stati
Frequency t
Frequency t
Frequency t

Frequency table: N: =VNormal(rnd(1);2;0;5) (Spreadsheet2.sta)
K-S d=,07270, p> .20; Lilliefors p> .20

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
0,000000<x<=,5000000	1	1	1,11111	1,1111	1,11111	1,1111
,5000000<x<=1,000000	2	3	2,22222	3,3333	2,22222	3,3333
1,000000<x<=1,500000	12	15	13,33333	16,6667	13,33333	16,6667
1,500000<x<=2,000000	21	36	23,33333	40,0000	23,33333	40,0000
2,000000<x<=2,500000	34	70	37,77778	77,7778	37,77778	77,7778
2,500000<x<=3,000000	18	88	20,00000	97,7778	20,00000	97,7778
3,000000<x<=3,500000	2	90	2,22222	100,0000	2,22222	100,0000
Missing	0	90	0,00000		0,00000	100,0000

Рис. 2.14. Таблица частот

3. Для построения полигона частот (как и в контрольном примере 2.1) выделим столбец *Percent of valid*, нажмем правую кнопку мыши и в контекстном меню выберем команду *Graph of Block Data – Line Plot: Entire Columns*.

Для построения гистограммы частот перейдем в окно *Graphs* на вкладку *Histograms.* Выбираем вкладку *Advanced.* Заполняем диалоговое окно, как это показано на рис. 2.15, и нажимаем *OK*.

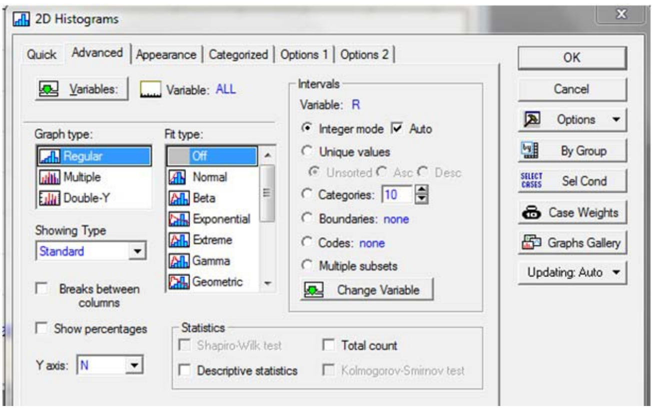


Рис. 2.15. Диалоговое окно вкладки *Advanced*

Наблюдаем гистограммы наших выборок (рис. 2.16), которые можно посмотреть, переходя по вкладкам внизу графика.

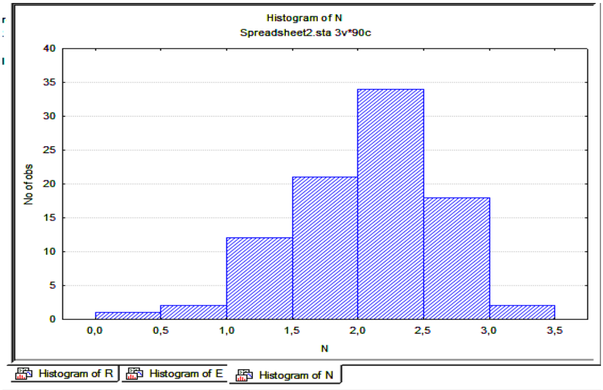


Рис. 2.16. Гистограммы выборок *N*, *R*, *E*

Построение кумулятивной кривой.

Вернемся в окно *Graphs* на вкладку *Histograms*. Во вкладке *Advanced* установим: *Graph type: Regular: Showing type: Cumulative; Fit type*: *Off; Variables – All; Categories (число интервалов группировки) – 250 – OK.*

Наблюдаем график кумулятивной кривой (рис. 2.17).

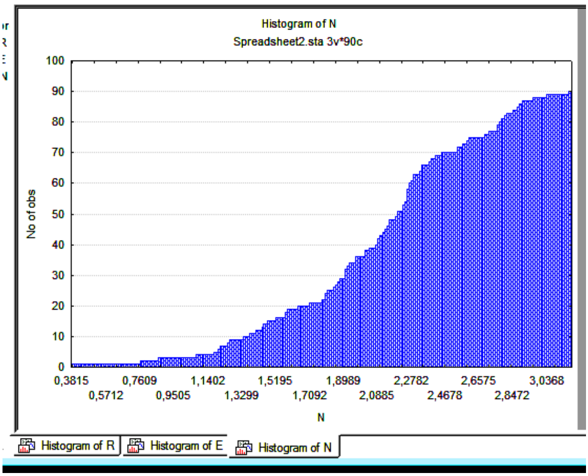


Рис. 2.17. Кумулянты заданных выборок

2.3. Задания для самостоятельной работы

Задание 1. В итоге многократных измерений некоторой физической величины одним прибором получены результаты, представленные в таблице. В пакетах *Excel* и *Statistica* необходимо построить по этим данным:

1) дискретный вариационный ряд, провести интервальную обработку;

2) полигон частот и гистограмму относительных частот, а также кумулянту.

Вариант 1

20	19	22	24	21	18	23	17	20	16
15	23	21	24	21	18	23	21	19	20
24	21	20	18	17	22	20	16	22	18
20	17	21	17	19	20	20	21	18	22
23	21	25	22	20	19	21	24	23	21
19	22	21	19	20	23	25	25	21	21

Вариант 2

10	0	1	2	3	4	5	6	7	8
9	1	2	3	4	5	6	7	8	9
1	3	4	5	9	1	5	7	3	7
3	5	3	5	3	5	5	3	5	9
0	1	2	3	4	5	6	7	8	9

Вариант 3

2	2	5	4	5	5	7	5	5	4
6	3	7	5	4	7	4	6	5	5
7	7	6	6	4	6	3	4	7	5
4	3	2	3	3	5	4	5	4	6
2	2	7	3	6	6	7	3	5	7

Вариант 4

1	1,1	1	0,8	0,6	0,6	0,6	0,7	0,7	0,7
1	0,7	0,7	0,6	0,5	0,7	0,7	0,8	0,8	0,9
0,7	1	0,9	0,6	0,7	0,8	0,8	0,8	0,7	0,7
1	0,9	1	1,1	0,8	0,8	0,7	0,6	0,6	0,6
0,9	0,8	0,7	0,7	0,8	0,7	0,8	0,9	0,8	0,8

Вариант 5

5	4	2	0	10	7	6	3	2	1
6	5	4	8	9	5	7	5	2	6
8	7	7	6	5	4	2	1	3	4
6	7	8	7	6	9	5	4	3	5
9	5	4	6	3	3	8	7	9	8

Вариант 6

0	5	2	4	7	5	2	0	1	8
6	10	10	5	8	7	4	6	5	5
8	7	8	6	5	3	3	4	2	4
4	3	2	2	1	1	5	1	4	1
2	8	7	9	9	6	7	3	5	5

Вариант 7

0	3	3	2	–1	4	1	1	3	3
8	2	5	1	5	3	8	2	6	5
1	1	4	5	1	3	6	1	2	5
1	4	5	3	3	7	0	1	2	0
5	5	4	2	2	3	4	3	4	2
3	2	2	3	3	4	2	2	3	8

Вариант 8

13	10	12	10	12	12	7	9	10	9
11	10	13	9	8	11	6	9	8	11
10	8	10	8	12	10	11	7	12	11
9	8	9	11	9	7	9	10	11	11
10	8	12	9	9	11	15	11	9	10
10	9	10	12	13	11	15	10	7	13

Вариант 9

4	2	4	3	1	2	0	1	5	0
0	1	4	3	0	2	1	4	2	2
3	0	0	2	4	1	1	2	0	1
4	5	4	3	4	0	2	6	1	6
3	0	3	1	2	2	3	5	1	3
4	0	5	3	3	0	1	3	4	6

Вариант 10

5	4	2	0	10	7	6	1	2	1
6	5	4	8	9	5	7	5	2	6
8	7	7	6	5	4	2	1	3	4
7	8	7	6	9	5	4	3	5	9
5	4	6	3	3	8	7	9	8	6

Задание 2. В пакете *Statistica* требуется выполнить:
– генерацию выборки объема n;
– построение вариационного ряда;
– группирование данных;
– построение полигона и гистограммы частот;
– построение эмпирической функции распределения.
Исходные данные находятся в табл. 2.7.

Таблица 2.7

№ варианта	Закон	n	№ варианта	Закон	n
1	$N(2,3)$ $E(3)$	55	2	$R[0,2]$ $N(2; 0,5)$	60
3	$R[0,7]$ $N(2; 0,8)$	70	4	$N(1,1)$ $E(2)$	80
5	$N(0; 0,3)$ $E(4)$	85	6	$R[0,1]$ $N(1; 0,3)$	70
7	$R[0,3]$ $N(2; 0,2)$	60	8	$N(0; 0,1)$ $E(7)$	65
9	$N(2; 3)$ $E(6)$	80	10	$R[0,5]$ $N(0; 4)$	85

Лабораторная работа № 3

ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ

3.1. Краткие теоретические сведения

По выборочным данным часто требуется определить (приближенно найти) параметры распределения исследуемой случайной величины. Выборочная оценка $\tilde{\theta}$ некоторого параметра θ (например, математического ожидания или дисперсии) сама является случайной величиной и должна удовлетворять определенным требованиям: быть *несмещенной*, *состоятельной* и *эффективной*.

Оценка $\tilde{\theta}$ называется *несмещенной* (оценкой без систематической ошибки), если ее математическое ожидание при любом n равно оцениваемому параметру: $M\left(\tilde{\theta}\right)=\theta.$

Оценка $\tilde{\theta}$ называется *состоятельной*, если при неограниченном увеличении выборки она сходится по вероятности к оцениваемому параметру: $\lim\limits_{n\to\infty}\left(P\left(\left|\tilde{\theta}-\theta\right|\right)<\varepsilon\right)=1$ для любого $\varepsilon>0.$

Оценка называется *эффективной* (в некотором классе оценок), если она имеет минимальную дисперсию в этом классе.

Выборочной средней $\overline{x}$ называют среднее арифметическое значение случайной величины X по выборочной совокупности объема n:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{k}x_i n_i. \qquad (3.1)$$

Выборочная средняя служит *несмещенной оценкой математического ожидания* признака X или генеральной совокупности.

Кроме выборочной средней в статистическом анализе применяются структурные средние: *медиана* и *мода*.

Мода выборки – варианта Mo с наибольшей частотой; медиана Me – варианта, которая делит вариационный ряд на равные части. Если $n=2m+1$, то $Me=x_{m+1}$, а если $n=2m$, то $Me=\dfrac{x_m+x_{m+1}}{2}$.

Средние величины не отражают изменчивости (вариации) значений признака. Чтобы охарактеризовать рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения $\overline{x}$ вводят свободную характеристику – *выборочную дисперсию*.

Выборочной дисперсией D_B называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\overline{x}$:

$$D_B = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2 \cdot n_i = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \overline{x}^2. \qquad (3.2)$$

Выборочным средним квадратическим отклонением (стандартом) называют квадратный корень из выборочной дисперсии:

$$\sigma_B = \sqrt{D_B}. \qquad (3.3)$$

Выборочная дисперсия является смещенной оценкой генеральной дисперсии. В качестве н*есмещенной оценки генеральной дисперсии* служит «исправленная» выборочная дисперсия:

$$S^2 = \frac{n}{n-1} D_B = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2 \cdot n_i. \qquad (3.4)$$

Стандартная ошибка среднего оценивает изменчивость выборочного среднего, приближенно показывая, насколько выборочное среднее отличается от среднего генеральной совокупности: $s_{\overline{x}} = \frac{s}{\sqrt{n}}$.

Скошенность кривой называется *асимметрией.* Для выборочной асимметрии $\tilde{A}_s$ справедлива формула: $\tilde{A}_s = \frac{1}{n \cdot s^3}\sum_i n_i\left(x_i - \overline{x}\right)^3$.

Отклонение крутизны называют *эксцессом*. Выборочный эксцесс $\tilde{E}_x$ определяется формулой: $\tilde{E}_x = \frac{1}{n \cdot s^4}\sum_i n_i\left(x_i - \overline{x}\right)^4 - 3$.

Так как асимметрия и эксцесс являются характеристиками формы кривой распределения, то по величине выборочных асимметрии и эксцесса можно делать предположения о его виде. Если выборочные асимметрия и эксцесс достаточно малы, т. е. близки к нулю, то можно выдвигать гипотезу о нормальном законе распределения генеральной совокупности.

Коэффициент вариации: $V = \frac{s}{\overline{x}} \cdot 100\,\%$.

Коэффициент вариации является относительной мерой рассеяния признака. Также он используется как показатель однородности выборочных наблюдений.

Считается, что если коэффициент вариации не превышает 10 %, то выборку можно считать однородной, т. е. полученной из одной генеральной совокупности.

Точечной называют оценку, которая определяется одним числом. Все оценки, приведенные выше, – точечные.

При малом числе наблюдений точечная оценка в значительной степени случайна, и замена истинного значения параметра на оценку может привести к серьезным ошибкам.

Чтобы дать представление о точности и надежности оценки, в математической статистике используют так называемые *доверительные интервалы* и *доверительную вероятность*.

Доверительный интервал (confidence interval) – вычисленный на основе выборки интервал значений признака, который с известной вероятностью содержит оцениваемый параметр генеральной совокупности.

Доверительная вероятность (или уровень доверия, confidence level) – это вероятность того, что доверительный интервал содержит значение параметра.

Доверительную вероятность принято устанавливать на уровнях 90, 95 и 99 %. Чем выше доверительная вероятность, тем более широкий и менее полезный получается интервал. Если доверительная вероятность не задана, считают, что она равна 0,95, или 95 %.

Уровень значимости α – это вероятность противоположного события (непопадания истинного значения параметра в доверительный интервал).

Точечной оценкой математического ожидания является выборочное среднее $\overline{x}$. Границы доверительного интервала определяются как $a_1 = \overline{x} - \delta$; $a_2 = \overline{x} + \delta$, где $\delta > 0$ – *точность доверительного интервала*, которая либо задается заранее, либо вычисляется.

Предположим, что наблюдается случайная величина X, имеющая нормальное распределение с параметрами a и σ. Для параметров строятся следующие доверительные интервалы.

1. Для неизвестного среднего a при известной дисперсии σ^2:

$$\overline{x} - t \cdot \frac{\sigma}{\sqrt{n}} < a < \overline{x} + t \cdot \frac{\sigma}{\sqrt{n}}. \tag{3.5}$$

Значение t находится из соотношения $\Phi(t) = \frac{\gamma}{2}$, где $\Phi(t)$ – *функция Лапласа*.

2. Для неизвестного среднего a при неизвестной дисперсии:

$$\overline{x} - t_\gamma \cdot \frac{s}{\sqrt{n}} < a < \overline{x} + t_\gamma \cdot \frac{s}{\sqrt{n}}, \tag{3.6}$$

где t_γ – критическая точка распределения Стьюдента (для двусторонней критической области) с числом степеней свободы $k = n - 1$ и уровнем значимости α;

s – «исправленное» стандартное отклонение.

Для вычисления критической точки распределения Стьюдента в MS Excel можно воспользоваться следующими функциями:

а) =СТЬЮДЕНТ.ОБР.2Х(α; n–1) – для двусторонней критической области;

б) =СТЬЮДЕНТ.ОБР(1–α; n–1) – для односторонней критической области.

В пакете *Statistica* все необходимые расчеты можно выполнить, используя вероятностный калькулятор.

3. Для неизвестной дисперсии нормально распределенной генеральной совокупности:

$$\frac{(n-1)\cdot s^2}{\chi^2\left(\frac{\alpha}{2}, n-1\right)} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi^2\left(1-\frac{\alpha}{2}, n-1\right)}. \qquad (3.7)$$

Значения $\chi^2\left(\frac{\alpha}{2}, n-1\right)$ и $\chi^2\left(1-\frac{\alpha}{2}, n-1\right)$ находятся:

– в пакете *Statistica* – с помощью вероятностного калькулятора;

– в пакете *Excel* – с помощью стандартной функции ХИ2.ОБР.ПХ(α; *k*), где *k – число степеней свободы.*

Иногда полученная точность не удовлетворяет пользователя, так как дает слишком широкий диапазон, в который попадает математическое ожидание с вероятностью *p*. Чем меньше точность доверительного интервала, тем ближе выборочная оценка к соответствующему генеральному показателю.

Точность зависит от числа наблюдений. Можно определить число наблюдений, которые необходимы для достижения заданной точности δ, по следующей формуле:

$$n(\delta) \geq t(\alpha, n-1)\cdot\frac{s^2}{\delta^2}+1. \qquad (3.8)$$

3.2. Практическая часть

Контрольный пример 3.1. По данным о массе тела учащихся из контрольного примера 2.1 лабораторной работы 2 рассчитать выборочные характеристики распределения. Найти 95 % доверительный интервал для математического ожидания и дисперсии. Определить, сколько нужно иметь наблюдений, чтобы точность определения маематичекого ожидания не превышала 0,2.

Задание выполнить в пакетах *Statistica* и *Excel*.

Решение.

В пакете *Microsoft Excel* для определения выборочных оценок параметров распределения используются следующие функции:

СРЗНАЧ() – вычисляет среднюю арифметическую аргументов (т. е. выборочную среднюю);

МЕДИАНА() – находит медиану заданной выборки;

МОДА.ОДН() – вычисляет наиболее часто встречающееся в выборке значение;

ДИСП.Г() – вычисляет выборочную дисперсию;

ДИСП.В() – вычисляет «исправленную» дисперсию;

СТАНДАРТОТКЛОН.В() – вычисляет «исправленное» СКО;

ЭКСЦЕСС() – вычисляет оценку эксцесса по выборке;

СКОС() – позволяет оценить асимметрию выборочного распределения.

Кроме того, в надстройке *Пакет анализа* имеется инструмент *Описательная статистика*, который дает возможность получить все выборочные характеристики случайной величины.

Введем эмпирические данные о весе учащихся на чистый лист *Excel* и оформим его, как показано на рис. 3.1.

	A	B	C	D	E	F	G	H
1	Наблюдения		Выборочные оценки (используя стандартные функции)				Выборочные оценки (пакет анализа)	
2	64		Среднее	60,855			*Наблюдения*	
3	57		Выборочная дисперсия	4,124				
4	63		Исправленная дисперсия	4,201			Среднее	60,855
5	62		Стандартное отклонение	2,050			Стандартная ошибка	0,276
6	58		Мода	60			Медиана	61
7	61		Медиана	61			Мода	60
8	63		Эксцесс	-0,744			Стандартное отклонение	2,050
9	60		Асимметрия	0,096			Дисперсия выборки	4,201
10	60		Количество наблюдений	55			Эксцесс	-0,744
11	61						Асимметричность	0,096
12	65						Интервал	8
13	62						Минимум	57
14	62						Максимум	65
15	60						Сумма	3347
16	64						Счет	55
17	61							

Рис. 3.1. Лист *Excel* с расчетом выборочных характеристик распределения

В ячейки D2:D9 для определения выборочных числовых характеристик введем стандартные функции Excel категории *Статистические*. Аргументами всех этих функций является диапазон выбранных значений веса A2:A56.

Распределение веса студентов является достаточно симметричным (асимметрия равна 0,096 и близка к нулю), а эксцесс имеет небольшое отрицательное значение (–0,744). Это означает, что распределение веса имеет более низкую и пологую вершину по сравнению с нормальным распределением.

Аналогичные данные можно получить с помощью инструмента *Описательная статистика* надстройки *Пакет Анализа*. Зададим команду *Данные/Анализ данных* и выберем инструмент *Описательная статистика*. Заполним диалоговое окно, как показано на рис. 3.2.

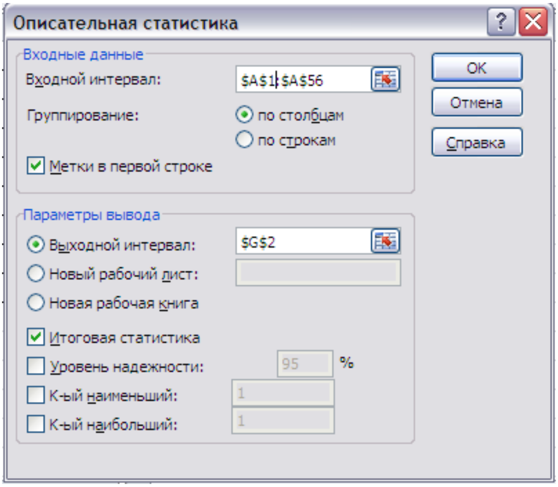


Рис. 3.2. Диалоговое окно для вывода описательной статистики

Данные каждой выборки должны быть расположены в одном столбце или одной строке. Переключатель *Группирование* в нашем случае установлен в положение по столбцам, так как эмпирические данные занесены в первый столбец.

Флажок *Метки в первой строке* установлен, поскольку входной интервал включает и заголовок данных в первой строке (слово «Наблюдения»).

Переключатель *Параметры вывода* установлен в положение *Выходной интервал*, так как нужно получить результаты расчетов на текущем листе *Excel*. В соответствующем поле указан адрес левой верхней ячейки выходного диапазона (G2).

Установим значок *Итоговая статистика* для вывода на листе Excel выборочных характеристик.

После заполнения этого диалогового окна нажимаем кнопку *OK*. Результаты расчетов с помощью *Пакета анализа* показаны на рис. 3.1 в столбцах G и H.

В пакете *Excel* существуют два основных варианта нахождения доверительного интервала для математического ожидания: когда дисперсия известна и когда неизвестна. В первом случае для вычислений применяется функция ДОВЕРИТ.НОРМ, а во втором – ДОВЕРИТ.СТЮДЕНТ.

Скопируем исходные данные на чистый лист *Excel* (см. рис. 3.3) и рассчитаем с помощью стандартных функций основные числовые характеристики: выборочное среднее, выборочное и «исправленное» стандартное отклонение.

Рассчитаем доверительный интервал с помощью функции ДОВЕРИТ.НОРМ. Ее синтаксис:

= ДОВЕРИТ.НОРМ (альфа; стандартное_откл; размер).

Здесь «альфа» – уровень значимости; «стандартное_откл» – стандартное отклонение предлагаемой выборки; «размер» – объем выборки.

Граница доверительного интервала определяется по формуле:

$$\bar{x} \pm \text{ДОВЕРИТ.НОРМ.}$$

Выделим ячейку, куда будет выводиться результат обработки данных (рис. 3.3 – ячейка Н3). Щелкаем по кнопке «*Вставить функцию*».

Появляется *Мастер функций*. Переходим в категорию «*Статистические*» и выделяем «ДОВЕРИТ.НОРМ». Нажимаем *OK*.

Открывается окно аргументов. Заполняем поля следующим образом: =ДОВЕРИТ.НОРМ(0,05;D2;55).

В ячейке D2 находится значение σ, рассчитанное с помощью стандартной функции Excel: =СТАНДОТКЛОН.Г(A1:A55).

Производим расчет левой границы доверительного интервала. Для этого выделяем отдельную ячейку (Н4), ставим знак «=» и вычитаем содержимое элементов листа, в которых расположены результаты вычислений выборочного среднего (ячейка D1) и ДОВЕРИТ.НОРМ. В нашем случае получилась следующая формула:

=D1 – H3.

Таким же образом производим вычисление правой границы доверительного интервала: =D1 + H3 (см. рис. 3.3).

Доверительный интервал для математического ожидания, найденный с помощью фунции ДОВЕРИТ.НОРМ, имеет вид:

$$60{,}319 < a < 61{,}391.$$

Функция ДОВЕРИТ.СТЮДЕНТ выполняет вычисление доверительного интервала генеральной совокупности с использованием распределения Стьюдента. Его удобно использовать в том случае, когда дисперсия и, соответственно, стандартное отклонение неизвестны. Синтаксис оператора таков:

=ДОВЕРИТ.СТЮДЕНТ(альфа; стандартное_откл; размер).

Рассчитаем границы доверительного интервала с неизвестным стандартным отклонением на примере нашей выборочной совокупности.

Выделим ячейку, в которую будет производиться расчет (H7 – рис. 3.3). Щелкаем по кнопке *«Вставить функцию»*.

В открывшемся *Мастере функций* переходим в категорию *«Статистические»*. Выбираем наименование «ДОВЕРИТ.СТЬЮДЕНТ» и нажимаем *OK*.

Заполняем поля: =ДОВЕРИТ.СТЬЮДЕНТ(0,05;D3;55).

В ячейке D3 находится «исправленное» стандартное отклонение, вычисленное с помощью функции *Excel* СТАНДОТКЛОН.В(A1:A55).

Далее, в ячейках H8 и H9 рассчитываем левую и правую границы доверительного интервала по формуле $D1 \mp H7$ (см. рис. 3.3).

	A	B	C	D	E	F	G	H	I	J	K
1	64		Среднее	60,855		Расчет доверительного интервала для M(X)					
2	57		Выборочное СКО	2,031		ДОВЕРИТ.НОРМ					
3	63		Исправленное СКО	2,050		Точность		0,537			
4	62		Доверительная вероятность	0,95		Левая граница		60,318			
5	58					Правая граница		61,391			
6	61		Расчет доверительного интервала для D(X)			ДОВЕРИТ.СТЮДЕНТ					
7	63		Верхний Хи2-квантиль	76,192		Точность		0,554			
8	60		Нижний Хи2-квантиль	35,586		Левая граница		60,300			
9	60		Левая граница интервала	2,977		Правая граница		61,409			
10	61		Правая граница интервала	6,374							
11	65					Количество наблюдений n для точности 0,2					
12	62							212			

Рис. 3.3. Расчет доверительных интервалов для параметров случайной величины

Доверительный интервал, найденный с помощью функции ДОВЕРИТ.СТЬЮДЕНТ: $60{,}3 < a < 61{,}409$.

Полученная точность доверительного интервала $\delta \approx 0{,}55$ превышает заданное значение 0,2. Определим по формуле (3.8), сколько необходимо иметь данных наблюдений для достижения этой точности. Для этого внесем, например, в ячейку H12 следующую формулу:

=СТЬЮДЕНТ.ОБР.2Х(0,05;54)*(D3^2)/(0,2^2)+1.

Полученное значение показывает, что нужно иметь не менее 212 наблюдений.

Для нахождения доверительного интервала для дисперсии рассчитаем квантили хи-квадрат распределения с помощью стандартной функции пакета ХИ2.ОБР.ПХ:

Верхний χ^2 квантиль: =ХИ2.ОБР.ПХ(0,05/2;54).

Нижний χ^2 квантиль: =ХИ2.ОБР.ПХ(1-0,05/2;54).

Границы доверительного интервала для дисперсии, рассчитанные по формуле 3.6, на рис. 3.3 находятся в ячейках D9 и D10.

Запишем доверительный интервал для дисперсии:

$$2{,}977 < \sigma^2 < 6{,}374.$$

Доверительный интервал для стандартного отклонения может быть получен путем извлечения квадратного корня из вышеуказанного выражения:

$$1{,}7254 < \sigma < 2{,}5247.$$

В пакете *Statistica.*

Для вычислений будем использовать исходную выборку (столбец) из 55 наблюдений над нормальной случайной величиной со средним $a = 60{,}855$ и дисперсией $s^2 = 2{,}0496^2 = 4{,}2009$. Скопируем данные с листа *Excel* в *Statistica*, предварительно создав в пакете новый документ.

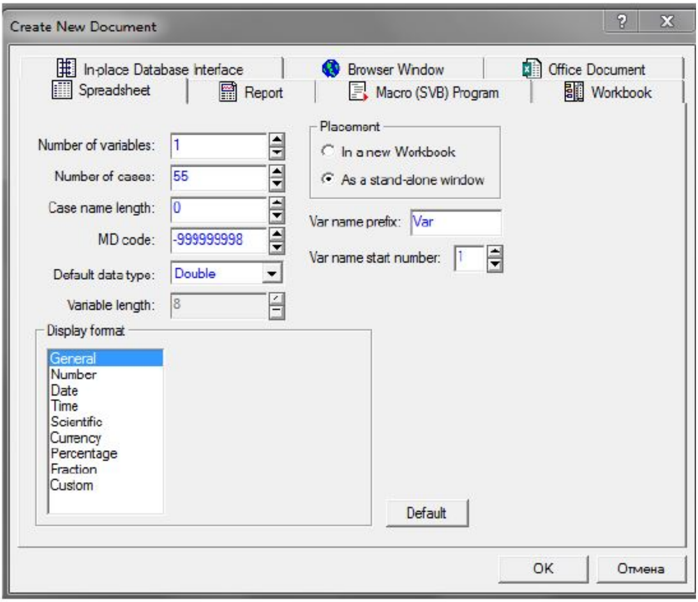


Рис. 3.4. Создание нового документа в пакете *Statistica*

В меню *Statistics – Basic Statistic/Tables* в окне *Descriptive Statistic* выберем вкладку *Advanced*, в результате появится окно, содержащее список числовых характеристик, которые могут быть вычислены. Отметим: *Mean* (среднее), *Median* (медиана), *Mode* (мода), *Standard Deviation* (среднее квадратическое отклонение), *Variance* (дисперсия), *Std.err.of mean* (стандартная ошибка среднего), *Skewness* (асимметрия), *Kurtosis* (эксцесс), *Range* (размах варьирования) – и активизируем кнопку *Summary*. В появившемся окне выделим переменную, для которой нужно произвести расчеты. В данном случае это переменная *X*.

Результаты вычислений размещаются в активном окне внизу (при анализе данных столбца) или слева (при анализе данных строки) от исходных данных (рис. 3.5).

	Descriptive Statistics (Spreadsheet1)									
Variable	Mean	Median	Mode	Frequency of Mode	Range	Variance	Std.Dev.	Standard Error	Skewness	Kurtosis
X	60,85455	61,00000	60,00000	10	8,000000	4,200673	2,049554	0,276362	0,096321	-0,743698

Рис. 3.5. Вычисленные параметры описательной статистики

Предположение о характере генерального распределения

Так как для нормального распределения эксцесс и асимметрия равны нулю, то достаточно малые значения эксцесса и асимметрии, а также то, что мода и медиана близки к среднему выборочному, позволяют выдвинуть гипотезу о нормальном распределении генеральной совокупности.

Определим доверительные интервалы для a и σ с уровнем доверия $\gamma = 0,95$.

Возвращаемся на вкладку *Advanced* процедуры *Descriptive Statistics*. Установим *Conf. Limits for means, Cl for Sample SD* и укажем значение *Interval: 95 %* (см. рис. 3.6), нажав на кнопку *Variables*, зададим имя переменной *X*.

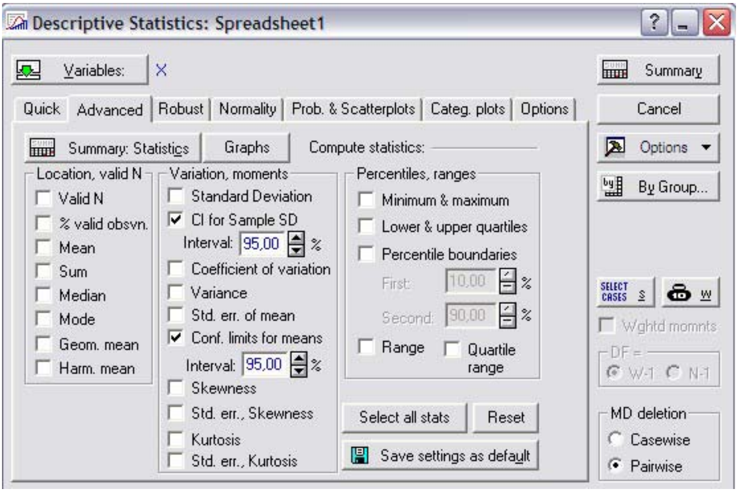


Рис. 3.6. Задание параметров расчета доверительных интервалов

Результаты вычислений будут представлены в отдельном окне (см. рис. 3.7). Первый столбец содержит имя исходной переменной, четыре других – левую и правую границы доверительного интервала.

	Descriptive Statistics (Spreadsheet1)			
Variable	Confidence -95,000%	Confidence 95,000	Confidence SD -95,000%	Confidence SD +95,000%
X	60,30047	61,40862	1,725447	2,524728

Рис. 3.7. Результаты вычислений

3.3. Задания для самостоятельной работы

Задание 1. Используя данные своего варианта из лабораторной работы № 2, вычислить выборочное среднее, моду, медиану распределения, выборочную и исправленные дисперсии, стандартное отклонение, коэффициент вариации, асимметрию и эксцесс, стандартную ошибку среднего. Сделать вывод.

Задание выполнить в пакетах *Excel* и *Statistica*.

Задание 2.

Вариант 1. Построить вариационный ряд по данным о росте (в см) группы из 40 мужчин. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь данных наблюдений, чтобы точность определения математического ожидания не превышала 1 см? Задание выполнить в пакетах *Excel* и *Statistica*.

181	169	178	171	179	172	181	179	168	174
167	169	171	179	181	181	183	172	176	165
178	176	1766	190	183	169	192	185	173	179
187	181	168	179	184	177	184	185	186	178

Вариант 2. Получены данные об успеваемости в группе из 20 студентов: 4, 4, 5, 3, 4, 5, 4, 5, 3, 5, 3, 3, 5, 4, 5, 4, 3, 5, 3, 5. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,2? Задание выполнить в пакетах *Excel* и *Statistica*.

Вариант 3. В рабочей зоне проводились замеры концентрации вредного вещества. Получен ряд значений (в мг/м^3): 12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16, 11, 12, 21, 17, 16, 15, 10, 13, 14, 19, 18, 17, 20, 14, 18. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,5? Задание выполнить в пакетах *Excel* и *Statistica.*

Вариант 4. В детской поликлинике измерялась масса тела детей одного возраста и получена следующая выборка: 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27. Построить 99 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,3? Задание выполнить в пакетах *Excel* и *Statistic*а.

Вариант 5. На складе исследовались объемы спроса на сахар-песок. Эмпирические данные по объемам спроса на сахар-песок за два месяца получены следующие: 900, 220, 180, 150, 50, 30, 50, 273, 250, 500, 50, 10, 730, 260, 20, 150, 50, 10, 150, 20, 410, 100, 100, 60, 60, 295, 250, 250, 50, 200, 400, 100, 50, 340, 200, 700, 520, 200. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 50? Задание выполнить в пакетах *Excel* и *Statistica.*

Вариант 6. Возраст студентов одного потока представляется следующими данными: 17, 20, 18, 19, 18, 17, 20, 21, 24, 22, 20, 21, 20, 19, 18, 20, 21, 22, 25, 20. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,5? Задание выполнить в пакетах *Excel* и *Statistica.*

Вариант 7. Служба контроля качества предприятия произвела замеры длин случайно отобранных заготовок (в см): 39, 41, 40, 43, 41, 44, 42, 41, 41, 43, 42, 39, 40, 42, 43, 42, 41, 39, 42, 42, 41, 42, 40, 41, 43, 41, 39, 40, 41, 40. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,3? Задание выполнить в пакетах *Excel* и *Statistica*.

Вариант 8. Для исследования системы массового обслуживания измерялись интервалы времени между поступлением заявок в систему: 1, 6, 12, 7, 1, 12, 1, 2, 8, 4, 3, 13, 1, 5, 5, 10, 2, 2, 2, 4, 3, 11, 2, 11, 3, 4, 5, 7, 6, 6, 9, 10, 1, 3, 1, 2. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,5? Задание выполнить в пакетах *Excel* и *Statistica.*

Вариант 9. В течение недели регистрировались пропуски занятий студентами одной группы. В результате регистрации получили статистические данные: 2, 1, 3, 1, 2, 1, 2, 4, 3, 5, 3, 2, 2, 2, 1, 2, 3, 1, 0, 0, 0, 2, 3, 1, 4. Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 0,3? Задание выполнить в пакетах *Excel* и *Statistica*.

Вариант 10. Среднее время сборки изделия составляло 90 минут. Инженер изобрел новый метод сборки этого изделия, и продолжительность сборки 10 изделий новым способом составила 79, 74, 112, 95, 83, 96, 77, 84, 70, 90 (мин). Построить 95 %-й доверительный интервал для математического ожидания и СКО, предполагая, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону. Сколько нужно иметь наблюдений, чтобы точность определения математического ожидания не превышала 5? Задание выполнить в пакетах *Excel* и *Statistica*.

Лабораторная работа № 4

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ИСТИННОСТИ ВЫДВИНУТОЙ ГИПОТЕЗЫ

4.1. Краткие теоретические сведения

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений.

Например, статистическими являются гипотезы:

1) генеральная совокупность распределена по нормальному закону;

2) дисперсии двух нормальных совокупностей равны между собой.

В первой гипотезе сделано предположение о виде неизвестного распределения, во второй – о параметрах двух известных распределений.

Гипотеза называется *параметрической*, если в ней содержится некоторое утверждение о параметрах распределения случайной величины (когда сам закон распределения считается известным), и *непараметрической* – в иных случаях.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (основной) называют выдвинутую гипотезу H_0.

Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т. е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная нулевая гипотеза. Вероятность ошибки первого рода называют *уровнем значимости* и обозначают через α.

Ошибка второго рода состоит в том, что будет принята неправильная нулевая гипотеза. Вероятность ошибки второго рода обозначают через β.

Вероятность принять верную гипотезу называется *уровнем доверия* $\gamma = 1 - \alpha$.

Вероятность принять альтернативную гипотезу, если она верна, называется *мощностью критерия*.

Статистическим критерием (или просто критерием) называют случайную величину K, которая служит для проверки нулевой гипотезы.

Схема построения критерия такова: все выборочное пространство делится на две взаимодополняющие области: *область отклонения* основной гипотезы H_0 и *область принятия* этой гипотезы. Область, при попадании в которую выборочной точки отвергается основная гипотеза, называется *критической.*

Наблюдаемым значением $K_{\text{набл}}$ называют значение критерия, вычисленное по выборкам.

Критическими точками (*границами*) называют точки $k_{\text{кр}}$, отделяющие критическую область от области принятия гипотезы. Критические точки разделяются на правосторонние и левосторонние области. *Правосторонняя* область определяется неравенством $K > k_{\text{кр}}$, *левосторонняя* – $K < k_{\text{кр}}$. Это односторонние области.

Существуют также и двусторонние области, определяемые неравенствами $K < k_{1\text{кр}}$, $K > k_{2\text{кр}}$, где $k_{2\text{кр}} > k_{1\text{кр}}$ ($k_{1\text{кр}}$ и $k_{2\text{кр}}$ – критические точки).

После того как критическая точка найдена, по данным выборки вычисляют наблюдаемое значение критерии. Если $K_{\text{набл}} > k_{\text{кр}}$ (для правосторонней области) нулевую гипотезу отвергают, если наоборот, то принимают.

Проверку нулевой гипотезы можно проводить с помощью так называемой *статистической значимости*. Статистическую значимость находят с помощью *p*-значения, которое соответствует вероятности данного события при предположении, что некоторое утверждение (нулевая гипотеза) истинно. Если *p*-значение меньше заданного уровня статистической значимости (обычно это 0,05) – нулевая гипотеза неверна, поэтому нужно перейти к рассмотрению альтернативной гипотезы.

Проверка гипотезы о распределении. Критерий Пирсона.

При проверке статистических гипотез о соответствии отдельных параметров закона распределения случайных величин предполагалось, что законы распределения этих величин известны. Однако при решении практических задач модель закона распределения в общем случае заранее неизвестна. Поэтому возникает необходимость выбора модели закона распределения, согласующейся с результатами выборочных наблюдений.

Для оценки соответствия экспериментальных данных теоретическому закону распределения используют:

1) графический метод, который состоит в сравнении гистограммы или полигона частот выборочного распределения и графика теоретической функции плотности распределения случайной величины;

2) по опытным данным ранее проведенных исследований;

3) с помощью критериев согласия и т. д.

Пусть $x_1, x_2, ..., x_n$ – выборка наблюдений случайной величины X с неизвестной функцией распределения $F(x)$. Проверяется гипотеза H_0, утверждающая, что X распределена по закону, имеющему функцию распределения $F(x)$, равную функции $F_0(x)$, т. е. проверяется нулевая гипотеза $H_0: F(x) = F_0(x)$. Критерии, с помощью которых проверяется нулевая гипотеза о неизвестном распределении, называются *критериями согласия*. Рассмотрим критерий согласия *Пирсона* (хи-квадрат распределения).

Схема проверки нулевой гипотезы $H_0: F(x) = F_0(x)$:

1. По выборке $x_1, x_2, ..., x_n$ строят вариационный ряд; он может быть как дискретным, так и интервальным.

2. По данным предыдущих исследований или по предварительным данным делают предположение (принимают гипотезу) о модели закона распределения случайной величины X.

3. По выборочным данным проводят оценку параметров выбранной модели закона распределения. Предположим, что закон распределения имеет r параметров (например, биномиальный закон имеет один параметр p; нормальный – два параметра (a, σ) и т. д.).

4. Подставляя выборочные оценки значений параметров распределения, находят *теоретические значения вероятностей* $p_i = P\left(X = x_i\right)$.

5. Рассчитывают *теоретические частоты* $n = n \cdot p_i$, где n – объем выборки.

6. Рассчитывают значение критерия согласия Пирсона:

$$\chi^2 = \sum_{i=1}^{l} \frac{\left(n_i - n_i'\right)^2}{n_i'}. \tag{4.1}$$

Здесь n_i – частоты данного статистического распределения; n_i' – теоретические частоты, найденные с помощью функции распределения предполагаемого закона;

Эта величина при $n \to \infty$ стремится к распределению χ^2 с $k = l - r - 1$ степенями свободы, где l – число интервалов для интервального вариационного ряда или число групп для дискретного ряда, r – число параметров предполагаемого распределения. В частности, если предполагаемое распределение является нормальным, то оценивается два параметра, поэтому число степеней свободы $k = l - 3$. В дальнейшем для расчетов используют таблицы распределения χ^2.

7. Задавая уровень значимости α, находят критическую область: она всегда правосторонняя – $\left(\chi^2_{\text{кр}}; \infty\right)$; значение $\chi^2_{\text{кр}}$ определяют из соотношения $\alpha = P\left(\chi^2 > \chi^2_{\text{кр}}\right)$. Если численное значение $\chi^2_{\text{набл}}$ попадает в интервал $\left(\chi^2_{\text{кр}}; \infty\right)$, то гипотеза $H_0 : F(x) = F_0(x)$ отклоняется и принимается альтернативная гипотеза о том, что выбранная модель закона распределения не подтверждается выборочными данными, при этом допускается ошибка, вероятность которой равна α.

Критерий согласия Пирсона можно использовать только в том случае, когда $n \cdot p_i \geq 5$. Поэтому тот интервал, для которого это условие не выполняется, объединяют с соседним и, соответственно, уменьшают число интервалов.

Замечание 4.1. В практике часто используется *приближенная проверка на нормальность*, в основе которой лежат более простые рекомендации, использующие значения числовых характеристик и свойства нормального распределения – известно, что если случайная величина подчиняется нормальному закону распределения, то ее значения удовлетворяют следующим условиям:

– промежуток $\bar{x} \pm 0,3\sigma_В$ содержит примерно $\frac{1}{4}$ части всей совокупности значений;

– промежуток $\bar{x} \pm 0,7\sigma_В$ – примерно $\frac{1}{2}$ части;

– промежуток $\bar{x} \pm 1,1\sigma_В$ – примерно $\frac{3}{4}$ части;

– промежуток $\bar{x} \pm 3\sigma_В$ – примерно 0,99 всех значений.

Если эти соотношения выполняются одновременно для данной эмпирической совокупности и вычисленных $\bar{x}, \sigma_В$, то гипотеза о нормальном законе распределения может быть принята.

Критерий Колмогорова предназначен для проверки гипотезы о законе распределения только непрерывных случайных величин. Он позволяет сравнить эмпирическую функцию $F^*(x)$ и теоретическую функцию распределения $F(x)$.

Схема применения критерия Колмогорова:

1) для предполагаемого закона распределения нужно определить $F(x)$ для значений аргументов, соответствующих правым концам интервалов;

2) вычислить значение статистики:

$$\lambda = \sqrt{n} \cdot \max_{x_i} \left| F(x_i) - F^*(x_i) \right|;$$

3) по уровню значимости α из таблицы 4.1 найти критическую точку $\lambda_{кр}$. Если $\lambda < \lambda_{кр}$, то различия между эмпирическим и предполагаемым теоретическим распределениями несущественны. Если $\lambda > \lambda_{кр}$, то различия между эмпирическим и предполагаемым теоретическим распределениями существенны.

Таблица 4.1

α	0,15	0,1	0,05	0,025	0,02	0,01	0,005	0,001
$\lambda_{кр}$	1,138	1,2238	1,3581	1,4802	1,5174	1,6276	1,738	1,9495

Проверка гипотез о параметрах распределения.

Проверка гипотезы о значении математического ожидания нормальной случайной величины с неизвестной дисперсией. Для проверки гипотезы $H_0: a = a_0$ о том, что среднее a нормальной случайной величины X равно заданному числу a_0, используется статистика

$$T = \frac{\overline{x} - a_0}{s} \cdot \sqrt{n}, \tag{4.2}$$

называемая *отношением Стьюдента.* Здесь n – объем выборки, $\overline{x}, s$ – выборочные оценки среднего a и стандартного отклонения σ нормальной случайной величины X.

Если проверяемая гипотеза верна, то статистика T имеет распределение Стьюдента (t-распределение) с $n-1$ степенями свободы.

Расчетное значение t статистики T вычисляется по формуле (4.2) подстановкой в нее гипотетического значения a_0 математического ожидания исследуемой случайной величины X и числовых значений выборочных оценок $\overline{x}, s,$ найденных по данным конкретной выборки.

Если H_1: $a \neq a_0$, то вычисляется наблюдаемое значение критерия и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $k = n-1$ находится критическая точка двусторонней критической области $t_{\text{двуст.крит}}(\alpha; k)$.

Проверка гипотезы о численном значении дисперсии нормального распределения. Пусть X – случайная величина, имеющая нормальный закон распределения $X \in N\left(a; \sigma^2\right)$, причем числовое значение дисперсии σ^2 неизвестно.

К гипотезе о значении генеральной дисперсии приходим, если требуется проверить предположение о точности станка или устройства. В качестве критерия проверки используют величину, которая зависит от выборочных данных $\left(s^2\right)$ и по значению которой можно судить о близости исправленной дисперсии к предполагаемому значению σ_0: $\chi^2 = \dfrac{(n-1)\cdot s^2}{\sigma_0^2}$. Критерий при выполнении гипотезы H_0 подчиняется распределению Пирсона (χ^2-распределению) с числом степеней свободы $k = n-1$. Особенности этапов проверки статистической гипотезы $H_0: \sigma^2 = \sigma_0^2$ приведены в табл. 4.2

Таблица 4.2

Выбор гипотез	$X \in N\left(a; \sigma^2\right)$; $H_0: \sigma^2 = \sigma_0^2$; $H_1: \sigma^2 = \sigma_1^2$		
Критическая область	$\sigma_1^2 < \sigma_0^2$ *Левосторонняя* $\chi^2_{\text{кр\_лев}}(1-\alpha;\ k)$	$\sigma_1^2 \neq \sigma_0^2$ *Двусторонняя* $\chi^2_{\text{кр\_пр}}(\alpha/2;\ k)$ *и* $\chi^2_{\text{кр\_лев}}(1-\alpha;\ k)$	$\sigma_1^2 > \sigma_0^2$ *Правосторонняя* $\chi^2_{\text{кр\_пр}}(\alpha;\ k)$
Критерий отклонения H_0	$\chi^2_{\text{набл}} < \chi^2_{\text{кр\_лев}}$	$\chi^2_{\text{набл}} < \chi^2_{\text{кр\_лев}}$ *или* $\chi^2_{\text{набл}} > \chi^2_{\text{кр\_пр}}$	$\chi^2_{\text{набл}} > \chi^2_{\text{кр\_пр}}$

4.2. Практическая часть

Контрольный пример 4.1.

На основании данных о массе тела студентов (Лабораторная работа № 2, Контрольный пример 2.1), проверить гипотезу о нормальном распределении исследуемой случайной величины, используя приближенную проверку на нормальность, а также с помощью критериев Пирсона (в пакетах *Excel* и *Statistica*), Колмогорова (в пакете *Excel*). Принять $\alpha = 0{,}05$.

1. *Приближенная проверка с использованием* $\sigma_{\text{В}}$.

Искомая выборка (вариационный ряд):

57	57	58	58	58	58	58	58	59	59	59
59	59	59	59	60	60	60	60	60	60	60
60	60	60	61	61	61	61	61	61	61	61
61	62	62	62	62	62	62	62	62	63	63
63	63	63	63	63	64	64	64	64	65	65

Выборочные числовые характеристики, вычисленные в лабораторной работе № 3: $\bar{x} = 60,855; \sigma = 2,031$.

Проведем приближенную проверку с использованием оценки σ.

Вычислим значения:

$$0,3 \cdot \sigma = 0,3 \cdot 2,031 = 0,6093; \quad 0,7 \cdot \sigma = 0,7 \cdot 2,031 = 1,4217;$$
$$1,1 \cdot \sigma = 1,1 \cdot 2,031 = 2,2341; \quad 3 \cdot \sigma = 3 \cdot 2,031 = 6,093.$$

Вычислим границы интервалов:

$$\bar{x} \pm 0,3\sigma = 60,855 \pm 0,6093 = (60,3;\ 61,5);$$
$$\bar{x} \pm 0,7\sigma = 60,855 \pm 1,4217 = (59,4;\ 62,3);$$
$$\bar{x} \pm 1,1\ \sigma = 60,855 \pm 2,2341 = (58,6;\ 63,1);$$
$$\bar{x} \pm 3\ \sigma = 60,855 \pm 6,093 = (54,8;\ 66,95).$$

Подсчитаем число значений (из общей совокупности), попавших в вычисленные интервалы:

$$n_1 = 9;\ n_2 = 27;\ n_3 = 41;\ n_4 = 55.$$

Вычисляем относительные частоты:

$$n_1 = \frac{9}{55} \approx 0,2; \quad n_2 = \frac{27}{55} \approx 0,5; \quad n_3 = \frac{41}{55} \approx 0,75; \quad n_4 = \frac{55}{55} = 1.$$

Убеждаемся, что во втором, третьем и четвертом интервалах содержится количество случайных чисел не менее рекомендуемого. А в первом количество чисел близко к рекомендуемому.

Данное эмпирическое распределение скорее всего подчиняется нормальному закону распределения, но нужно провести проверку, используя более точные критерии.

Проверим гипотезу о нормальном распределении с помощью критерия Пирсона.

В пакете *Excel.*

Введем исходные данные и оформим их так, как показано на рис. 4.1 (интервальный ряд был получен в лабораторной работе № 2 –

табл. 2.7). Вычислим выборочные характеристики, используя стандартные функции пакета. В ячейке G9 найдем общее число наблюдений, просуммировав фактические частоты.

	A	B	C	D	E	F	G
1	**Наблюдения**		**Числовые характеристики**		**xi-1**	**xi**	**Эмпирические частоты**
2	64		**максимум**	**65**	**57**	**58,15**	**8**
3	57		**минимум**	**57**	**58,15**	**59,3**	**7**
4	63		**среднее**	**60,8545**	**59,3**	**60,45**	**10**
5	62		**станд. откл**	**2,0496**	**60,45**	**61,6**	**9**
6	58		**асиммерия**	**0,0963**	**61,6**	**62,75**	**8**
7	61		**эксцесс**	**-0,7437**	**62,75**	**63,9**	**7**
8	63				**63,9**	**65,05**	**6**
9	60				**Всего наблюдений**		**55**
10	60						

Рис. 4.1. Вычисление выборочных характеристик исходной выборки

В ячейку H2 внесем формулу для вычисления значения функции нормального распределения $F\left(x_1 = 58{,}15\right)$. В *Excel* эту величину можно вычислить, воспользовавшись функцией НОРМ.РАСП (рис. 4.2).

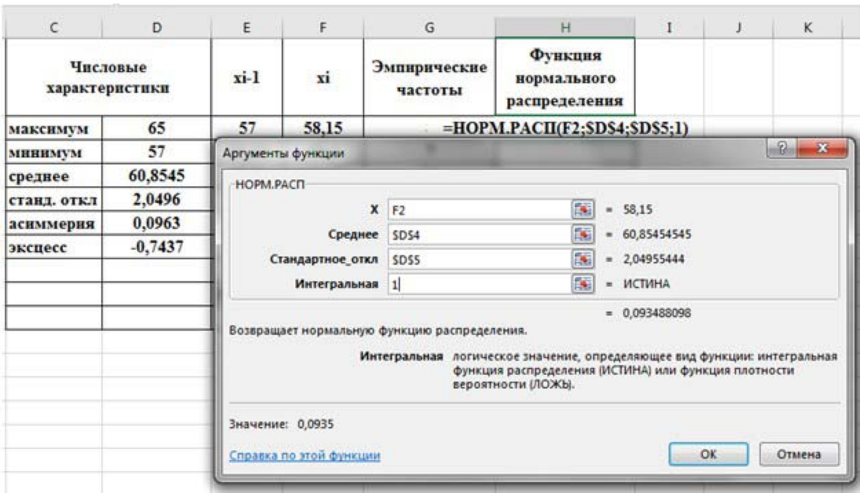


Рис. 4.2. Диалоговое окно функции НОРМРАСП с заполненными полями ввода

В поле *X* введен адрес ячейки, в которой находится граница первого интервала группировки.

В поле *Среднее* введен адрес ячейки, в которой находится среднее значение выборки.

В поле *Стандартное_откл* введен адрес ячейки, в которой находится значение стандартного отклонения выборки.

В поле *Интегральная* введена единица 1. Единица в поле Интегральная означает вычисление функции распределения *F(x)*.

Далее с помощью маркера автозаполнения протянем эту формулу до ячейки I8.

В ячейку I2 введем формулу =H2, в ячейку I8 – формулу =1–H7, а в ячейку I3 – формулу =H3–H2, протянув ее до ячейки I7 (рис. 4.3).

В результате этих действий в диапазоне H2:H8 появятся значения теоретических вероятностей $p_1, p_2, ..., p_7$, причем $p_1 + p_2 + ... + p_7 = 1$.

В ячейку J2 введем формулу =I2*G9 и протянем ее до ячейки J8. С помощью этой формулы вычисляются теоретические частоты n_i'. Их сумма равна объему выборки $n = 55$. На этом заканчивается первый этап проверки (рис. 4.3).

E	F	G	H	I	J
xi-1	**xi**	**Эмпирические частоты**	**Функция нормального распределения**	**Теоретические вероятности**	**Теоретические частоты**
57	58,15	8	0,0935	0,0935	5,1418
58,15	59,3	7	0,2241	0,1306	7,1827
59,3	60,45	10	0,4218	0,1977	10,8725
60,45	61,6	9	0,6420	0,2202	12,1110
61,6	62,75	8	0,8225	0,1805	9,9277
62,75	63,9	7	0,9313	0,1089	5,9884
63,9	65,05	6	0,9797	0,0687	3,7758
Всего наблюдений		55		1	55

Рис. 4.3. Первый этап проверки гипотезы по критерию хи-квадрат

Результаты заключительного этапа проверки приведены в диапазоне L1:N13 (рис. 4.4). В столбце U в ячейке N11 будет вычисляться наблюдаемое значение критерия.

В диапазоне L2:L8 находятся эмпирические частоты n_i, в диапазоне M2:M8 – теоретические частоты n_i'.

В ячейку N2 введена формула =(L2-M2)^2/M2, реализующая вычисления по формуле $\frac{\left(n_i - n_i'\right)^2}{n_i'}$. Размножим эту формулу в диапазоне ячеек N3:N8. В ячейке N11 получим сумму содержимого ячеек N2:N8 по формуле =СУММ(N2:N8).

Критическое значение статистики U, которая имеет распределение с $k = 7 - 3 = 4$ степенями свободы, определяется при помощи функции ХИ2.ОБР.ПХ (0,05; 4).

Расчетное значение $\chi^2_{\text{набл}} = 4{,}3179$ статистики U меньше ее критического значения $\chi^2_{\text{крит}} = 9{,}4877$ (рис. 4.4), поэтому можно сказать, что проверяемая гипотеза, состоящая в том, что генеральная совокупность подчиняется нормальному закону распределения, не противоречит данным эксперимента.

E	F	G	H	I	J	K	L	M	N
xi-1	xi	Эмпирические частоты	Функция нормального распределения	Теоретические вероятности	Теоретические частоты		Эмп. частоты	Теорет. частоты	U
57	58,15	8	0,0935	0,0935	5,1418		8	5,1418	1,5887
58,15	59,3	7	0,2241	0,1306	7,1827		7	7,1827	0,0046
59,3	60,45	10	0,4218	0,1977	10,8725		10	10,8725	0,0700
60,45	61,6	9	0,6420	0,2202	12,1110		9	12,1110	0,7991
61,6	62,75	8	0,8225	0,1805	9,9277		8	9,9277	0,3743
62,75	63,9	7	0,9313	0,1089	5,9884		7	5,9884	0,1709
63,9	65,05	6	0,9797	0,0687	3,7758		6	3,7758	1,3102
Всего наблюдений		55		1	55		55	55	
							Наблюдаемое значение Хи-квадрат		
									4,3179
							Критическое значение Хи-квадрат		
									9,4877

Рис. 4.4. Таблица с окончательными результатами вычисления статистики

2. *Проверка истинности гипотезы* H_0 *по критерию Колмогорова.*

Скопируем данные наблюдений на чистый лист *Excel* и построим интервальный вариационный ряд (рис. 4.5).

Рассчитаем выборочную среднюю и «исправленное» стандартное отклонение, используя функции СРЗНАЧ(A2:A56) и СТАНДОТКЛОН.В(A2:A56) – в ячейках D4 и D5.

Вычислим статистику λ в граничных точках интервального ряда (см. рис. 4.5)

A	B	C	D	E	F	G	H	I	J	K	L
Наблюдения		**Числовые характеристики**		**xi-1**	**xi**	**Частоты**	**Относительные частоты**	**Накопленные частоты (F*(x))**	**(xi-xs)/s**	**F(xi)**	**|F*(x)-F(x)|**
64		**максимум**	**65**	менее 57		0	0	0	-29,6916	0	0,0000
57		**минимум**	**57**	57	58,15	8	0,1455	0,1455	-1,31958	0,093488	0,0520
63		**среднее xs**	**60,85455**	58,15	59,3	7	0,1273	0,2727	-0,75848	0,224082	0,0486
62		**станд_откл s**	**2,049554**	59,3	60,45	10	0,1818	0,4545	-0,19738	0,421764	0,0328
58				60,45	61,6	9	0,1636	0,6182	0,363715	0,641965	0,0238
61				61,6	62,75	8	0,1455	0,7636	0,924813	0,822468	0,0588
63				62,75	63,9	7	0,1273	0,8909	1,485911	0,931349	0,0404
60				63,9	65,05	6	0,1091	1,0000	2,047008	0,979671	0,0203
60				**Всего наблюдений**		**55**	**1**				**0,0588**

Рис. 4.5. Расчеты для критерия Колмогорова

В столбцах H и I рассчитаем относительные и накопленные частоты (см. лабораторную работу № 2). Накопленные частоты (точки кумулятивной кривой) – левые концы «ступенек» эмпирической функции распределения, т. е. $F^*(x)$.

Значения $F(x_i)$ вычисляются с учетом того, что была выдвинута гипотеза о нормальном распределении. В пакете *Excel* значения $F(x_i)$ можно вычислить с помощью функции НОРМ.СТ.РАСП.

Из последнего столбца таблицы ясно, что $\max\limits_i \left|F^*(x_i) - F(x_i)\right| = 0{,}0588$. Тогда значение статистики Колмогорова:

$$\lambda = \sqrt{55} \cdot 0{,}0588 \approx 0{,}436.$$

По заданному уровню значимости $\alpha = 0{,}05$ определим границу критической области (табл. 4.1) $\lambda_{\text{крит}} = 1{,}3581$. Поскольку $\lambda = 0{,}436 < \lambda_{\text{крит}} = 1{,}3581$, то основная гипотеза принимается, т. е. генеральное распределение считается нормальным.

3. Проверим гипотезу о нормальном распределении случайной величины в пакете *Statistica*.

После ввода исходных данных будем использовать процедуру *Distribution Fitting* (подбор распределения), как показано на рис. 4.6.

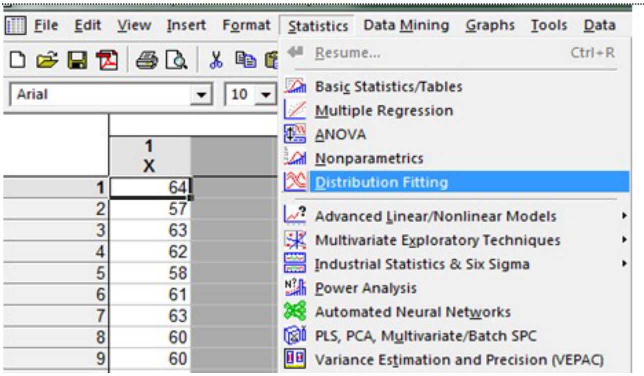


Рис. 4.6. Вызов процедуры *Distribution Fitting*

На вкладке *Quick* выбираем непрерывные распределения (*Continuos Distributions*) – *Normal* (если выдвигается гипотеза об экспоненциальном распределении – *Exponential*; если о равномерном – *Rectangular*).

Зададим диапазон исходных данных, нажав на кнопку *Variable* и выбрав там X . Далее нажмем кнопку *OK*.

Во вкладке *Parameters* установим количество интервалов разбиения (*Number of categories*), равное 7 (рис. 4.7), а также минимальное (*Lower limit*) и максимальное (*Upper limit*) значения выборки.

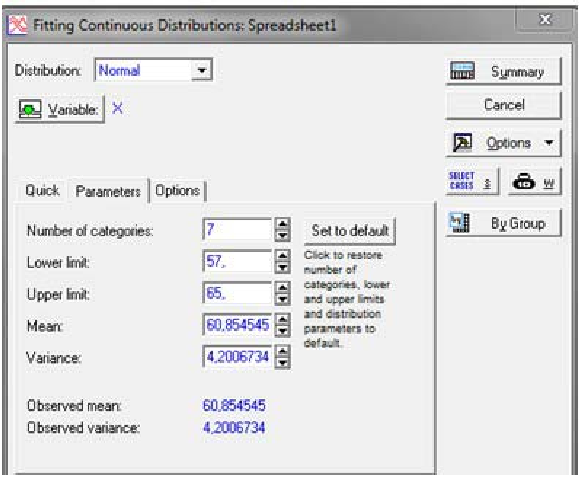


Рис. 4.7. Вкладка *Parameters*

Нажав кнопку *Summary*, получаем таблицу частот (рис. 4.8).

Upper Boundary	Variable: X, Distribution: Normal (Spreadsheet1) Chi-Square = 3,73766, df = 3 (adjusted) , p = 0,29122 Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 58,14286	8	8	14,54545	14,5455	5,10990	5,10990	9,29073	9,2907	2,89010
59,28571	7	15	12,72727	27,2727	7,10020	12,21010	12,90945	22,2002	-0,10020
60,42857	10	25	18,18182	45,4545	10,76219	22,97229	19,56761	41,7678	-0,76219
61,57143	9	34	16,36364	61,8182	12,04876	35,02104	21,90683	63,6746	-3,04876
62,71429	8	42	14,54545	76,3636	9,96340	44,98445	18,11528	81,7899	-1,96340
63,85714	7	49	12,72727	89,0909	6,08523	51,06968	11,06405	92,8540	0,91477
< Infinity	6	55	10,90909	100,0000	3,93032	55,00000	7,14604	100,0000	2,06968

Рис. 4.8. Таблица частот

Если гипотеза верна, вероятность получить 3,73766 или больше равна 0,29122 (больше 0,05 – уровня значимости) – достаточна, чтобы поверить в нормальность распределения исходных данных. Следовательно, гипотезу о нормальном распределении случайной величины принимаем.

Вывод. На основе критериев Пирсона и Колмогорова, которые дали аналогичные результаты, гипотезу о нормальном распределении генеральной случайной величины следует принять.

Контрольный пример 4.2. Согласно технической документации, среднее время срабатывания взрывателя ручной гранаты равно 4 секунды. При проверке 12 взрывателей зафиксированы следующие значения времени срабатывания:

4,02; 3,92; 4,07; 4,18; 4,17; 4,23; 3,83; 4,03; 4,16; 3,94; 3,98; 3,88.

1. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : a = 4$ о том, что среднее время срабатывания взрывателя равно 4 секунды. Принять в качестве $H_1 : a \neq 4$.

2. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу о том, что стандартное отклонение σ времени X срабатывания ручной гранаты равно 0,1 с.

Задание выполнить в пакетах *Statistica* и *Excel*.

Решение.

1. Проверим гипотезу в пакете *Excel*.

В диапазон A1:A12 листа *Excel* введем исходные данные наблюдения (рис. 4.9).

D9 | f_x | =СТЬЮДЕНТ.ОБР(1-0,05;11)

	A	B	C	D	E
1	4,02		Выб. Среднее =	4,034167	
2	3,92		S =	0,129927	
3	4,07		t =	0,910948	
4	4,18		**Критическая точка для двусторонней критич. области**		
5	4,17		t(0.025, 11) =	2,200985	
6	4,23		**Нахождение критической точки с помощью функции**		
7	3,83		СТЬЮДЕНТ.ОБР(1-0,025;11)=	2,200985	
8	4,03		**Критическая точка для односторонней критич. области**		
9	4,16		t_прав_крит(0.05;11)=	1,795885	
10	3,94				
11	3,98				
12	3,88				

Рис. 4.9. Проверка гипотезы о среднем времени срабатывания взрывателя гранаты

В ячейки C1:C5 введем информационные метки: Выб. среднее =, $\mathrm{S} =$, $\mathrm{t} =$, $\mathrm{t}(0{,}025{,}11) =$.

В ячейку D1 введем формулу = СРЗНАЧ(A1:A12) и нажмем клавишу *Enter*. В ячейке D1 появится выборочное среднее времени срабатывания взрывателя. Ручная граната непригодна для боевого использования, если среднее время срабатывания взрывателя слишком мало или слишком велико (противник успеет бросить гранату обратно). Поэтому в качестве H_1 используется гипотеза $a \neq a_0$. Такой альтернативе соответствует двусторонняя критическая область $K_{кр}(0{,}05) = (|T_{11}| \geq t(0{,}025;11))$, где $t(0{,}025;11)$ – критическое значение распределения Стьюдента с 11 степенями свободы.

В ячейку D2 введем формулу = СТАНДОТКЛ.В(A1:A12) и нажмем клавишу *Enter*. В ячейке появится выборочное стандартное отклонение $S = 0{,}129927$ времени срабатывания взрывателя.

В ячейку D3 введем формулу =(D1–4)*КОРЕНЬ(12)/D2 и нажмем клавишу *Enter*.

В ячейку D5 введем формулу =СТЬЮДЕНТ.ОБР.2Х(0,05;11) и нажмем клавишу *Enter*. В ячейке появится критическое значение $t_{кр}(0{,}025{,}11) = 2{,}2$ порядка 0,025 распределения Стьюдента с 11 степенями свободы.

Полученный результат $\left(|t| < t\left(\frac{\alpha}{2}, n-1\right)\right)$ свидетельствует о том, что гипотеза $H_0\colon a = 4$ не противоречит данным эксперимента.

Замечание. Критическое значение можно найти с помощью стандартной функции =СТЬЮДЕНТ.ОБР(1–0,025;11) – результат будет аналогичным.

В пакете *Statistica.*

Введем исходные данные.

Для проверки гипотезы будем использовать процедуру Basic *Statistics/Tables*, которая находится в меню *Statistics*.

Выбираем *t-test*, *single sample* (рис. 4.10).

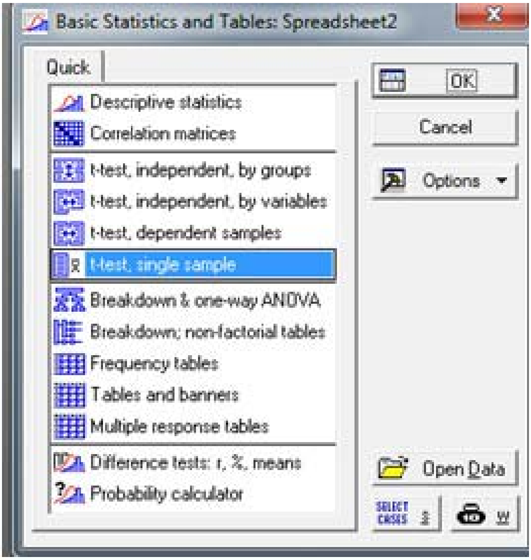


Рис. 4.10. Процедура *Basic Statistics/Tables*

Зададим диапазон исходных данных, нажав на кнопку *Variables.* Нажимаем кнопку *OK*. Далее на вкладке *Advanced* зададим исходные данные задачи (рис. 4.11).

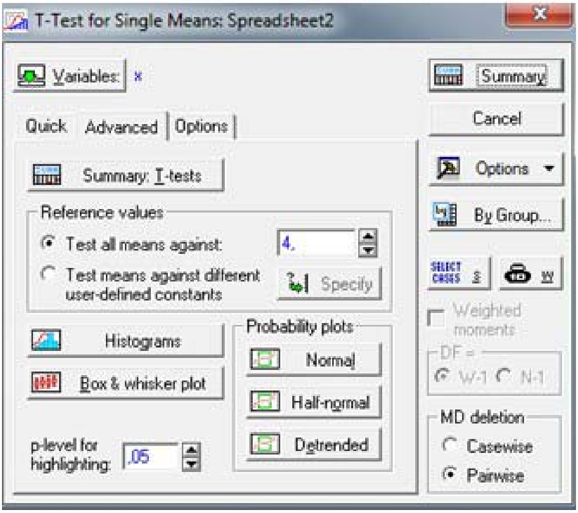
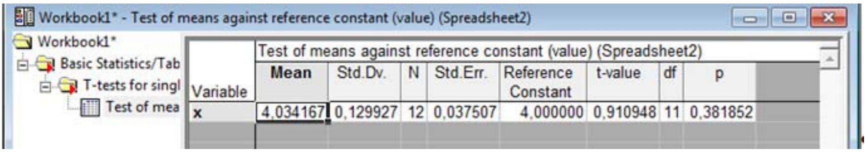


Рис. 4.11. Исходные данные задачи на вкладке *Advanced*

Нажав кнопку *Summary*, получаем таблицу (рис. 4.12).



Workbook1* - Test of means against reference constant (value) (Spreadsheet2)

Test of means against reference constant (value) (Spreadsheet2)

Variable	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
x	4,034167	0,129927	12	0,037507	4,000000	0,910948	11	0,381852

Рис. 4.12. Проверка гипотезы о математическом ожидании
при неизвестной дисперсии в пакете *Statistica*

Так как $p = 0{,}381852 > 0{,}05,$ то приходим к выводу, что проверяемая гипотеза не противоречит данным эксперимента.

2. Проверяемая гипотеза эквивалентна гипотезе $H_0\colon \sigma^2 = 0{,}1^2 = 0{,}01.$ В данном случае объем выборки $n = 12$ и исправленная дисперсия $s^2 = 0{,}13^2 \approx 0{,}017$ с^2.

Увеличение разброса времени срабатывания взрывателя влечет за собой возрастание опасности самоподрыва и опасности использования гранаты противником. В связи с этим в качестве конкурирующей надо выбрать гипотезу $H_1\colon \sigma^2 > \sigma_0^2.$

На рис. 4.13 представлено решение данной задачи в пакете Excel.

C	D	E	F	G	H
Выб. Среднее =	4,034167		$\chi^2_{набл}$ =	18,56917	
S =	0,129927				
t =	0,910948		$\chi^2_{кр}$ =	19,67514	
t(0.025, 1) =	2,200985				

Рис. 4.13. Проверка гипотезы о значении дисперсии

В ячейку G1 введена формула =D2^2*11/0,01, в ячейку G3 – функция =ХИ2.ОБР.ПХ(0,05;11).

Сравнивая расчетное значение статистики χ^2 с ее критическим значением порядка 0,05, приходим к выводу, что проверяемая гипотеза не противоречит данным наблюдения.

4.3. Задание для самостоятельной работы

Задание 1. Используя данные своего варианта из лабораторной работы № 2:

– провести приближенную проверку на нормальность;

– проверить, согласуются ли выборочные данные с гипотезой о нормальном распределении с помощью критериев Пирсона (в пакетах Excel, Statistica) и Колмогорова (в пакете Excel).

Принять $\alpha = 0,05$.

Задание 2. Проектный, контролируемый размер изделий, изготовляемых станком автоматом $a = a_0$ мм. Измерения 20 случайно отобранных изделий дали результаты, приведенные в табл. 4.3.

1. Требуется при уровне значимости $\alpha = 0,05$ проверить гипотезу $H_0 : a = a_0$ при конкурирующей гипотезе $H_1 : a \neq a_0$. Задачу решить с применением пакетов *Statistica* и *Excel*.

2. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Можно ли принять партию при уровне значимости а) 0,01; б) 0,05? Задачу решить с применением пакета Excel.

Таблица 4.3

Вариант 1 $a_0 = 24$									
24,5	21,8	23,2	26,4	23,5	26,1	23,1	24,2	25,8	21,7
23,6	28,1	26,2	22,2	25	24,9	23,9	24,5	26,1	24,6
Вариант 2 $a_0 = 35$									
34,6	35,4	34,1	35,3	36,1	33,5	36,2	35,1	35,2	35,5
34,4	33,9	34,8	34,6	34,6	36,1	34,9	35	35,6	34,8
Вариант 3 $a_0 = 25$									
24	28	29	21	266	32	19	27	32	25
23	25	26	21	25	21	22	22	21	27
Вариант 4 $a_0 = 25$									
25,8	38,9	24,5	26,9	27,7	24,1	27,9	35	29,3	39,1
34,3	28,5	22,2	26,4	26,7	30	30,4	32,3	28,4	35,6
Вариант 5 $a_0 = 22$									
18,6	21,4	23	22,5	21,4	22,4	22,1	19,7	22,5	22,9
21,5	21,3	20,6	23,1	23,2	23,1	20,5	22,1	19,1	22,5
Вариант 6 $a_0 = 30$									
22	24	38	31	22	24	38	31	22	29
30	25	30	27	28	33	28	27	24	24
Вариант 7 $a_0 = 26$									
29	24	26	23	28	25	26	25	30	27
26	27	25	31	29	30	24	21	29	27
Вариант 8 $a_0 = 33$									
31,5	34,4	32,4	30,1	34,4	31,7	25	30,2	33,5	31,7
29,2	26,8	26,4	26,7	33,2	27,2	28,8	30,4	31,6	26
Вариант 9 $a_0 = 30$									
31,3	32,6	29,3	30,6	33,2	28,6	30,5	31,8	29,2	29,6
29,5	28,6	31,7	33,4	36	32,1	28,1	29	32,2	34,4
Вариант 10 $a_0 = 34$									
33,5	30,5	32,4	34,6	32,4	36,7	34,8	35,6	40	30,3
34	32,9	40	32,1	33,2	31,7	29,2	31,7	32,5	30,4

Вопросы к зачету по курсу

1. Законы распределения непрерывной случайной величины: равномерный, показательный.

2. Нормальный закон распределения непрерывной случайной величины.

3. Распределения, связанные с нормальным: χ^2, Фишера, Стьюдента.

4. Основы работы в пакете Statistica. Вероятностный калькулятор.

5. Формы, виды и способы статистического наблюдения

6. Генеральная совокупность, выборка, репрезентативность. Построение дискретного и интервального вариационного ряда.

7. Графическое изображение вариационного ряда.

8. Точечные оценки характеристик случайной величины: средние величины.

9. Структурные характеристики выборочной совокупности. Мода и медиана.

10. Точечные оценки характеристик случайной величины: показатели вариации.

11. Дисперсия и среднее квадратическое отклонение.

12. Точечные оценки характеристик случайной величины: асимметрия и эксцесс. Предварительная проверка на нормальность.

13. Интервальные оценки. Доверительная вероятность и доверительный интервал.

14. Проверка статистических гипотез. Нулевая, конкурирующая, простая и сложная гипотезы. Ошибки первого и второго рода.

15. Критерии проверки гипотез. Уровень значимости и критические области. Схема проверки статистических гипотез.

16. Приближенная проверка на нормальность (графическая, с помощью коэффициентов асимметрии и эксцесса, с использованием σ).

17. Критерий χ^2. Проверка гипотезы о нормальном распределении случайной величины.

18. Критерий Колмогорова.

19. Проверка гипотезы о значении математического ожидания нормально распределенной случайной величины.

20. Проверка гипотезы о значении дисперсии нормально распределенной случайной величины.

Литература

1. Лисьев, В. П. Теория вероятностей и математическая статистика : учебное пособие / В. П. Лисьев. – М. : Московский государственный университет экономики, статистики и информатики, 2006. – 199 с.

2. Маталыцкий, М. А. Теория вероятностей и математическая статистика : пособие / М. А. Маталыцкий, Т. В. Русилко. – 2-е изд., перераб. и доп. – Гродно : ГрГУ, 2009. – 219 с.

3. Математическая статистика в примерах и задачах : учебное пособие / Г. С. Евдокимова. – Смоленск : Изд-во СмолГУ, 2014. – 98 с.

4. Гмурман, В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высшая школа, 2003. – 479 с.

5. Вадзинский, Р. Статистические вычисления в среде Excel. Библиотека пользователя / Р. Вадзинский. – СПб. : Питер, 2008. – 608 с.: ил.

6. Аверьянова, С. Ю. Лабораторный практикум по математической статистике в среде ЭТ MS Excel : учебное пособие / С. Ю. Аверьянова, Н. В. Растеряев. – Ростов н/Д : Издательство Южного федерального университета, 2014. – 64 с.

7. Еськова, О. И. Основы статистической обработки информации : пособие / О. И. Еськова, Л. П. Авдашкова, М. А. Грибовская. – Мн. : Беларусь, 2011. – 175 с.: ил.

8. Математическая статистика : учеб-метод. пособие / авт-сост.: С. Е. Демин, Е. Л. Демина; М-во образования и науки РФ; ФГАОУ ВО «УрФУ им. Первого президента России Б. Н. Ельцина», нижнетагил. технол. ин-т (фил.). – Нижний Тагил: НТИ (филиал) УрФУ, 2016. – 284 с.

9. Решение задач теории вероятностей, математической статистики и математического программирования средствами EXCEL: метод. указания к выполнению самост. работы для бакалавров техн. и экон. направлений подготовки очной формы обучения / сост.: Ю. Г. Кошкин, Е. П. Погодина, Н. Г. Тетерина. – Красноярск : Сиб. гос. аэрокосмич. ун-т, 2014 – 88 с.

10. Лунгу, К. Н. Высшая математика. Руководство к решению задач / К. Н. Лунгу, Е. В. Макаров. – М. : ФИЗМАТЛИТ, 2007. – Ч. 2. – 384 с.

11. Стукач, О. В. Программный комплекс Statistica в решении задач управления качеством: учебное пособие / О. В. Стукач. – Томск : изд-во Томского политехнического университета, 2011. – 163 с.

12. Дубровина, О. В. Прикладная математика : метод. пособие по выполнению практических и лабораторных работ для студентов заочного отделения специальности 1-54 01 01 «Метрология, стандартизация и сертификация» / О. В. Дубровина, Н. К. Прихач, В. М. Романчак. – Мн. : БНТУ, 2009. – 70 с.

13. Теория вероятности и математическая статистика : учебно-методическое пособие для студентов специальностей 1-38 01 01, 1-38-01 02, 1-52 01 01, 1-38 02 02, 1-54 01 01: в 2 ч. / сост. Н. К. Прихач [и др.]; под ред. М. А. Князева. – Мн. : БНТУ, 2020. – Ч. 2. – 72 с.

14. Прихач, Н. К. Прикладная математика [Электронный ресурс]: учебно-методическое пособие для студентов специальности 1-54 01 01 «Метрология, стандартизация и сертификация (по направлениям)» / Н. К. Прихач, И. В. Прусова. – Мн. : БНТУ, 2020. – Режим доступа : http://rep.bntu.by/handle/data/9383. – Дата доступа : 06.02.2022.

15. Лабораторный практикум по информатике для студентов инженерных специальностей приборостроительного факультет : в 3 ч. / сост.: Л. В. Бокуть [и др.]; под ред. В. А. Ибрагимова. – Мн. : БНТУ, 2002. – Ч. 2. – 98 с.

Учебное издание

ПРИХАЧ Наталия Константиновна
ПРУСОВА Ирина Васильевна
РОМАНЧАК Василий Михайлович

**ПРИКЛАДНАЯ МАТЕМАТИКА.
ВЫБОРКА И ЕЕ АНАЛИЗ**

Учебно-методическое пособие
для студентов специальности
1-54 01 01 «Метрология, стандартизация и сертификация
(машиностроение и приборостроение)»

Редактор *Н. А. Костешева*
Компьютерная верстка *Н. А. Школьниковой*

Подписано в печать 13.06.2022. Формат 60×84 $^{1}/_{16}$. Бумага офсетная. Ризография.
Усл. печ. л. 4,48. Уч.-изд. л. 3,50. Тираж 100. Заказ 251.

Издатель и полиграфическое исполнение: Белорусский национальный технический университет.
Свидетельство о государственной регистрации издателя, изготовителя, распространителя
печатных изданий № 1/173 от 12.02.2014. Пр. Независимости, 65. 220013, г. Минск.