МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

(ДГТУ)

Кафедра «Математики и информатики»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
для самостоятельной работы студентов
по дисциплине
«Информационные системы и технологии в научных исследованиях»

для магистрантов направления 09.04.02
«Информационные системы и технологии»

Ростов-на-Дону
ДГТУ
2022

УДК 004.8(075.8)

Составитель: В.В. Мисюра

Методические указания для самостоятельной работы студентов по дисциплине «Информационные системы и технологии в научных исследованиях» – Ростов-на-Дону : Донской гос. техн. ун-т, 2022. – 25 с.

Методические указания предназначены для обучающихся по направлению подготовки магистратуры направления 09.04.02 «Информационные системы и технологии».

УДК 004.8(075.8)

Печатается по решению редакционно-издательского совета
Донского государственного технического университета

Ответственный за выпуск зав. кафедрой «Математика и информатика»
д-р физ.-мат. наук, профессор А.И. Сухинов

_____

В печать _____ г.
Формат 60×84/16. Объем _____усл. п. л.
Тираж _____экз. Заказ № _____

_____

Издательский центр ДГТУ
Адрес университета и полиграфического предприятия:
344000, г. Ростов-на-Дону, пл. Гагарина, 1

Структура выполнения самостоятельной работы магистрами в семестре включает в себя:

1. Усвоение текущего учебного материала на лекциях и лабораторных работах (очная форма обучения– 48 ч; заочная форма обучения – 6 ч)

В процессе лекции студент ведет конспектирование излагаемого преподавателем материала с выделением основных понятий, фактов, формул, правил и т.д. Лекционные, лабораторные занятий проводятся в компьютерном классе с проектором или интерактивной доской, а также наличием пакетов численного моделирования, пакетов MATLAB, MATCAD, средой обработки статистических данных R.

Конспект лекций следует вести аккуратно, выделяя разделы, подразделы, отдельные темы. При последующей самостоятельной внеаудиторной работе студенту необходимо отметить непонятные выражения и положения, закончить (вставить) слова, пропущенные (упущенные) при конспектировании с использованием рекомендуемой к лекции литературы.
Усвоение текущего учебного материала предусматривает:
- повторение ранее изученного материала;
- изучение текущих тем лекций с использованием основной и дополнительной литературы из рабочей программы;
- выявление наиболее трудного для понимания материала;
- подготовку вопросов по материалу лекции.

В течение семестра преподавателем проводятся консультации. В период сессии проводятся индивидуальные и групповые консультации, в том числе перед экзаменом.

2. Самостоятельная работа (очная форма обучения - 92 ч; заочная форма обучения – 161 ч)

В процессе изучения курса «Информационные системы и технологии в научных исследованиях» студентам необходимо обратить особое внимание на

самостоятельное изучение дополнительных разделов дисциплины и тем лекций, вынесенных на самостоятельную проработку с использованием рекомендованной учебной (а также научной и популярной) литературы, для чего предусмотрена работа в библиотеке и в компьютерном классе по изучению материалов в Интернете, а так же на подготовку к лабораторным работам.

Самостоятельная работа с учебными пособиями, научной и популярной литературой по материалам периодики и Интернета является наиболее эффективным методом получения знаний по предмету, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала.

При работе с литературой по конкретным темам курса основное внимание следует уделять важнейшим понятиям, терминам, определениям, для скорейшего усвоения которых целесообразно вести краткий конспект.

Самостоятельная работа студентов с литературой не отделена от лекций и лабораторных работ, однако вдумчивое чтение источников, составление тезисов, обобщение прочитанных материалов способствует гораздо более глубокому пониманию изучаемой проблемы. Данная работа также предполагает обращение студентов к справочной литературе для уяснения конкретных терминов и понятий, введенных в курс, что способствует пониманию и закреплению пройденного лекционного материала и подготовке к лабораторным занятиям.

Самостоятельное изучение дополнительных разделов дисциплины должно выполняться, в том числе, на основе технических средств в компьютерных классах при использовании соответствующих обучающих программ.

Подготовка к лабораторным занятиям осуществляется на основании тематики, представленной в рабочей программе дисциплины, материалов к лабораторным занятиям, а также вопросов предлагаемых для подготовки к занятию преподавателем при изучении предыдущей темы.

Студент перед лабораторной работой должен изучить основные вопросы, теоретический материал, необходимый для понимания сущности процессов протекающих при ее выполнении.

Подготовка и защита лабораторных работ осуществляется на основании предварительно оформленных отчетов, а также вопросов для самоконтроля, приведенных в материалах по выполнению соответствующих работ или предложенных преподавателем в процессе занятий.

Студент должен уточнить цель работы, изучить теоретический материал, необходимый для понимания сущности процессов протекающих при ее выполнении, выводы, сформулированные по результатам работы и ответы на контрольные вопросы.

Прием лабораторных работ преподавателем производится в течение семестра, как правило, на занятиях, либо на плановых консультациях, назначаемых преподавателем в течение семестра.

Для студентов заочной формы обучения предполагается контроль самостоятельной работы в семестре в виде индивидуального задания (контрольной работы), приведенной ниже.

Контрольная работа для магистрантов заочной формы обучения

В среде статистической обработки данных R:

1. Построить график заданной функции плотности вероятности

$$f(x)=\begin{cases}\frac{1}{2}\lambda_1 e^{-\lambda_1(b-x)}, x\in[a,b)\\ 0{,}15, x\in[b,c)\\ \frac{1}{3}\lambda_2 e^{-\lambda_2(x-c)}, x\in[c,d)\\ 0, x\notin(a,d]\end{cases}$$

где a=9.945, b=13, c=16, d=19.083, λ_1=0.3, λ_2=0.45

2. Используя метод обратной функции и композиции, предложить алгоритм получения случайных величин в соответствии с заданными параметрами закона распределения.

3. Получить выборку размером 10000 случайных чисел в соответствии с заданными параметрами закона распределения.

4. Провести разведывательный анализ, основные статистические характеристики полученных данных.

5. Проверить соответствие полученных данных теоретическому закону распределения по критерию Пирсона, Колмогорова, Шапиро-Уилка.

Примеры выполнения разведывательного анализа, получения основных статистических характеристик приведены в приложениях 1,2.

Таблица 1.

Номер варианта	Пуассоновское распределение	Экспоненциальное распределение	Нормальное распределение	
	λ	μ	Mx	σ^2
1	1,0	0,33	0,15	2,55
2	1,1	0,44	0,22	2,22
3	5,2	0,45	0,33	2,07
4	1,4	0,55	0,45	1,91
5	1,5	0,66	0,51	2,11
6	4,6	0,75	0,62	1,88
7	1,8	0,72	0,77	2,33
8	2,0	0,83	0,83	1,77
9	5,5	0,94	0,91	2,44
10	3,0	1,08	1,05	0,66

По контрольной работе проводится устный опрос (зачет контрольной работы), после которого магистрант приступает к сдаче промежуточной аттестации в форме экзамена. По результатам устного опроса по контрольной

работе обучающемуся выставляется оценка «зачтено», или «не зачтено».

Оценка «зачтено» выставляется обучающемуся, если:

- обучающийся знает и воспроизводит основные положения дисциплины в соответствии с заданием, применяет их для выполнения типового задания, в котором очевиден способ решения;

- обучающийся демонстрирует базовые знания, умения и навыки, примененные при выполнении заданий контрольной работы;

- у обучающегося не имеется затруднений в использовании научно-понятийного аппарата в терминологии курса, а если затруднения имеются, то они незначительные;

- на дополнительные вопросы преподавателя обучающийся дал правильные или частично правильные ответы.

Оценка «не зачтено» ставится обучающемуся, если:

- обучающийся имеет представление о содержании дисциплины, но не знает основные положения (темы, раздела, закона и т.д.), к которому относится задание, не способен выполнить задание с очевидным решением, не владеет навыками в области изучаемой дисциплины;

- обучающийся не демонстрирует базовые знания, умения и навыки, необходимые для выполнения заданий контрольной работы;

- в процессе ответа по теоретическому и практическому материалу, содержащемуся в вопросах контрольной работы, допущены принципиальные ошибки при изложении материала.

3. Подготовка к экзамену

Подготовка к экзамену представляет собой обобщение всего материала дисциплины на основании конспекта лекций и рекомендованных литературных источников и заключается во всестороннем рассмотрении всех тем с обязательным повторением материала лабораторных занятий.

Вопросы, выносимые на экзамен, в соответствии с рабочей программой дисциплины доводятся до студентов на последнем лекционном занятии в семестре.

Примерный перечень вопросов к экзамену (вопросы могут использоваться как темы рефератов для студентов заочной формы обучения)

1. Научное исследование: его сущность и особенности. Классификация научных исследований. Методология научного исследования.
2. Методология и научное познание. Метод научного исследования. Метод и теория научного исследования.
3. Теоретический и эмпирический уровни научного исследования. Классификация методов (философские, общенаучные, частнонаучные).
4. Методы междисциплинарного исследования. Системный метод научных исследований, его сущность и основные характеристики.
5. Классификация систем (статические, динамические, детерминистические, стохастические).
6. Понятия «модель» и «моделирование» в научном исследовании. Этапы процесса моделирования. Классификация моделей и формы моделирования.
7. Математические модели и методы.
8. Значение математических моделей в научных исследованиях, их основные типы (описательные, объяснительные, прогнозные, управленческие). Понятие научного знания и определение научных проблем.
9. Анализ и синтез, абстрагирование, индукция и дедукция. Методы моделирования изучаемых объектов.
10. Математическое и физическое моделирование. Выбор направления научного исследования и этапы научно-исследовательской работы.
11. Классификация научно исследовательских работ (НИР). Основные этапы НИР. Критерии актуальности НИР.
12. Сбор и анализ информации по теме исследования. Рабочая гипотеза составление плана исследования. Основные стадии выполнения теоретических исследований. Мат. методы в исследованиях.
13. Типы мат. моделей. Виды уравнений. описывающих динамику объекта. Аналитические методы исследования мат. моделей. Методы стат анализа.
14. Дисперсионный, регрессионный, корреляционный и спектральный

анализы. Основные задачи, виды и основы планирования эксперимента.

15. Метрологическое обеспечение экспериментальных исследований. Государственная система обеспечения единства измерений. Методы измерений прямые и косвенные. Методы оценки.

16. Автоматизированная система, объект исследования, исполнительная, информационная и вычислительная подсистемы. Квантование непрерывного сигнала.

17. Аналого-цифровые и цифро-аналоговые преобразователи. Примеры автоматизированных систем для научных исследований. Основные структуры систем автоматизации научных исследований.

18. Виды обеспечений АСНИ (организационное, информационное, математическое, техническое, программное, лингвистическое, метрологическое, правовое и эргономическое).

19. Технические средства автоматизации эксперимента. Программное обеспечение. Структура управляющей программы.

Список источников для усвоения учебного материала по дисциплине «Информационные системы и технологии в научных исследованиях»

	Авторы,	Заглавие	Издательство, год
Л1.1	Афанасьев, В.Н., Еремеева, Н.С.	Статистическая методология в научных исследованиях: Учебное пособие для	Оренбург: Оренбургский государственный университет, ЭБС АСВ, 2017
Л1.2	Казиев, В.М.	Введение в анализ, синтез и моделирование систем: Учебное пособие	Москва, Саратов: Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020
Л1.3	Бабёнышев, С.В., Матеров, Е.Н.	Математические методы и информационные технологии в научных исследованиях:	Железногорск: Сибирская пожарно-спасательная академия ГПС МЧС России, 2018
Л1.4	Киценко, Т.П., Лахтарина, С.В.	Методология, планирование и обработка результатов эксперимента в научных исследованиях: учебно-	Макеевка: Донбасская национальная академия строительства и архитектуры, ЭБС АСВ, 2020
Л1.5	Голубева, Н.В.	Математическое моделирование систем и процессов	Санкт-Петербург: Лань, 2021
Л1.6	Воскобойников, Ю.Е.	Регрессионный анализ данных в пакете MATHCAD	Санкт-Петербург: Лань, 2021
Л1.7	Агалаков, С.А.	Анализ данных в среде R: практикум	Омск: Омский государственный университет им. Ф.М. Достоевского, 2020
Л1.8	Поспелов, Е.А., Попов, И.С.	Пакеты прикладных программ в научных исследованиях: учебно-методическое пособие	Омск: Омский государственный университет им. Ф.М. Достоевского, 2019
2. Дополнительная литература			
Л2.1	Маккинли, Уэс, Слинкина, А.	Python и анализ данных	Саратов: Профобразование, 2019
Л2.2	Логунова О.С., Романов П.Ю.	Обработка экспериментальных данных на ЭВМ: Учебник	Москва: ООО "Научно-издательский центр ИНФРА-М", 2021

3. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"	
1	Система тематических коллективных блогов https://habr.com/
2	Образовательные курсы ведущих вузов России https://openedu.ru/
3	Образовательный портал ДГТУ
4	Научная электронная библиотека. URL: https://elibrary.ru/
5	Национальная электронная библиотека. URL: https://нэб.рф/

Приложение 1

Пример протокола разведочного анализа данных: проверка на нормальность распределения

Подчиняются ли анализируемые количественные переменные закону нормального распределения вероятностей? Очень многие статистические методы предполагают положительный ответ на этот вопрос, и поэтому проверка исследуемых переменных на нормальность распределения является важной составной частью разведочного анализа данных.

Проверяя условие нормальности распределения данных, необходимо, однако, хорошо представлять себе, в каких случаях его выполнение является критическим для применения конкретного статистического метода. Так, например, метод главных компонент (Principle Components Analysis, PCA) не требует, чтобы данные были распределены нормально. Линейная регрессия (Linear Regression) хотя и предполагает нормальность распределения зависимой переменной, является достаточно робастным методом при незначительных отклонениях от этого условия. В то же время для успешного применения дискриминантного анализа (Discriminant Analysis) нормальность распределениях признаков в каждой группе классифицируемых объектов - условие обязательное.

Существует несколько способов проверки анализируемых данных на нормальность распределения. Все их можно разделить на две рассмотренные ниже группы.

Графические способы

Самый простой графический способ проверки характера распределения данных - построение *гистограммы*. Создать гистограмму можно при помощи R-функции hist(). Если гистограмма имеет колоколообразный симметричный вид, можно сделать заключение о том, что анализируемая переменная имеет примерно нормальное распределение. Однако при интерпретации гистограмм следует соблюдать осторожность, поскольку их внешний вид может сильно

зависеть как от числа наблюдений, так и от шага, выбранного для разбиения данных на классы. Кроме того, достаточно часто при анализе нормально распределенных *смешанных* совокупностей гистограммы приобретают асимметричный вид, вводя исследователя в заблуждение (Рисунок 1).

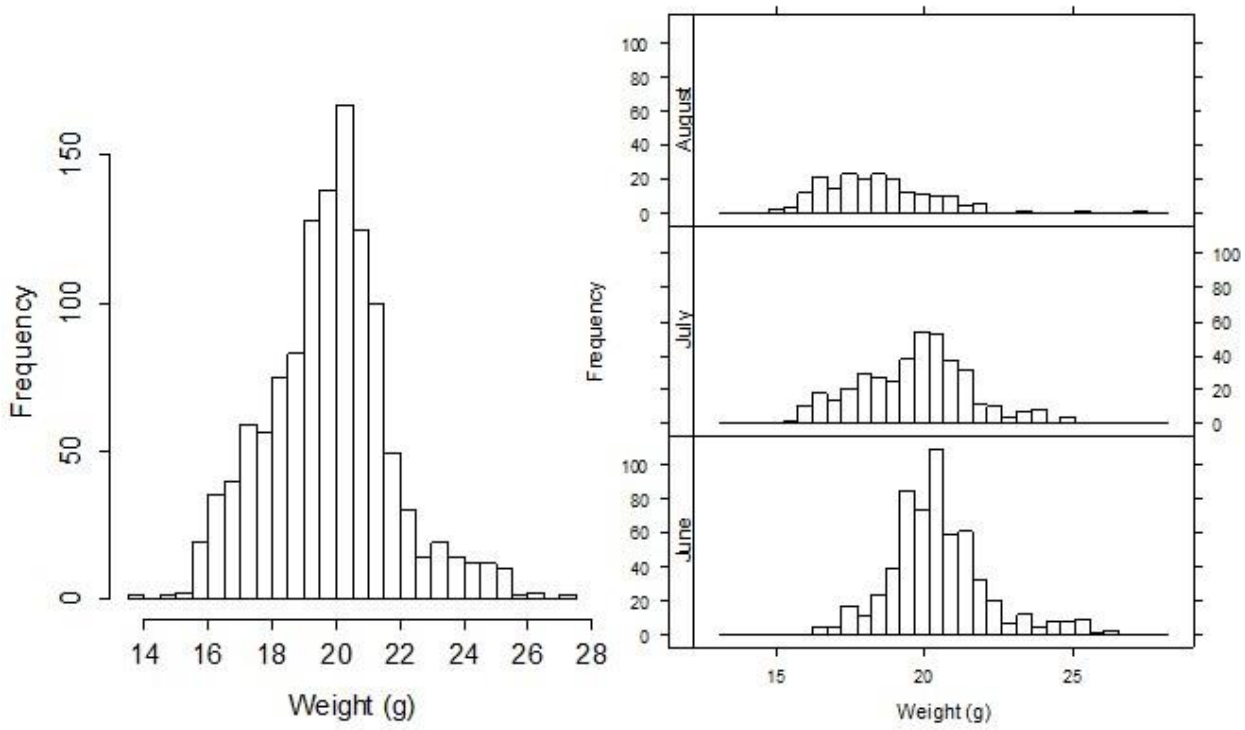

Рисунок 1. Гистограмма распределения веса 1193 воробьев (по: Zuur et al. 2010).

На графике слева приведены объединенные данные для июня, июля и августа. Поскольку вес птиц зависит от времени года, гистограмма приобретает асимметричный вид. На графике справа показаны те же данные, но отдельно по каждому месяцу. Из этого графика хорошо видно, что на самом деле вес воробьев как биологический признак имеет
примерно нормальное распределение.

Другим очень часто используемым графическим способом проверки характера распределения данных является построение т.н. графиков квантилей (Q-Q plots, QuantileQuantile plots). На таких графиках изображаются квантили двух распределений - эмпирического (т.е. построенного по анализируемым данным) и теоретически ожидаемого

стандартного нормального распределения. При нормальном распределении проверяемой переменной точки на графике квантилей должны выстраиваться в прямую линию, исходящую под улом 45 градусов из левого нижнего угла графика. Графики квантилей особенно полезны при работе с небольшими по размеру совокупностями, для которых невозможно построить гистограммы, принимающие какую-либо выраженную форму.

В R для построения графиков квантилей можно использовать базовую функцию qqnorm(), которая в качестве основного аргумента принимает вектор со значениями анализируемой переменной (Рисунок 2):

```
x <- rnorm(500) # генерация нормально распределенной совокупности с n = 500 qqnorm(x)
```

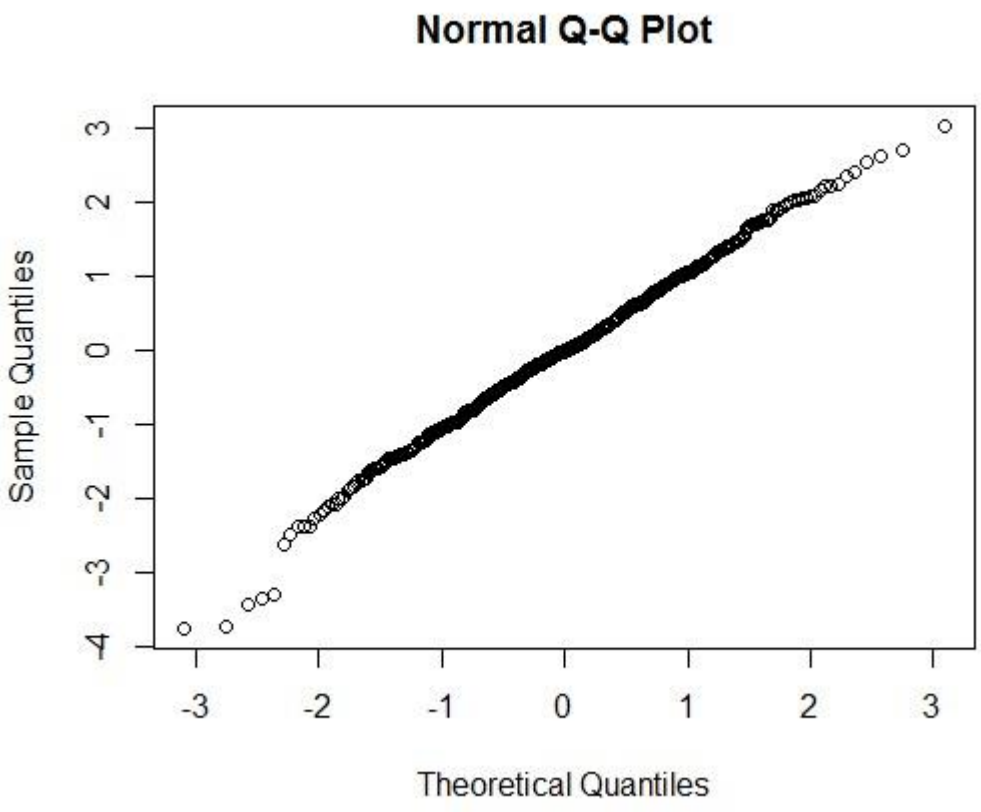

Рисунок 2. Пример графика квантилей для нормально распределенной совокупности, состоящей из 500 наблюдений.

Следует отметить, что интерпретация графиков квантилей при работе с

небольшими выборками, происходящими из нормально распределенных генеральных совокупностей, требует определенного навыка. Дело в том, что при небольшом числе наблюдений точки на графике квантилей могут не всегда образовывать четко выраженную прямую линию.

Формальные тесты

Существует целый ряд статистических тестов, специально разработанных для проверки нормальности распределения данных. В общем виде проверяемую при помощи этих тестов нулевую гипотезу можно сформулиировать так: "Анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение". Если получаемая при помощи того или иного теста вероятность ошибки P оказывается меньше некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется.

В R реализованы практически все имеющиеся тесты на нормальность - либо в виде стандарных функций, либо в виде функций, входящих в состав отдельных пакетов. Примером базовой функции является shapiro.test(), при помощи которой можно выполнить широко используемый тест Шапиро-Уилка:

```
shapiro.test(rnorm(500))

Shapiro-Wilk normality test
    data:  rnorm(500)
W = 0.9978, p-value = 0.7653 # P > 0.05 - нулевая гипотеза не отвергается
```

Ниже перечислены функции из пакета nortest, реализующие другие распространенные тесты на нормальность (установить этот покет можно командой install.packages("nortest")):

- `ad.test()` - тест Андерсона-Дарлинга

- `cvm.test()` - тест Крамера фон Мизеса
- `lillie.test()` - тест Колмогорова-Смирнова в модификации Лиллиефорса
- `pearson.test()` - критерий хи-квадрат Пирсона
- `sf.test()` - тест Шапиро-Франсия (см. *Thode 2002*)

Приложение 2

Расчет параметров описательной статистики в R

Использование специальных функций

Благодаря наличию специально созданных для этого функций, расчет параметров описательной статистики в R не составляет никакого труда. Ниже демонстрируется использование этих функций на примере данных по характеристикам 32 моделей автомобилей (таблица `mtcars`, входящая в стандартный набор данных R):

```
data(mtcars)
mtcars
```

```
                  mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4        21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag    21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710       22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive   21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7  8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant          18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360       14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D        24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230         22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280         19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
...
```

Каждая модель описана по 11 признакам, два из которых (`vs` и `am`) являются номинальными переменными (факторами) с уровнями, закодированными в виде 0 и 1 (подробнее см. `?mtcars`).

Для расчета *арифметической средней*, *медианы*, *дисперсии*, *стандартного отклонения*, а также *минимального и максимального значений* в R служат функции `mean()`, `median()`, `var()`, `sd()`, `min()` и `max()` соответственно.

Используем эти функции в отношении, например, параметра `mpg` (пробег автомобиля (в милях) в расчете на один галлон топлива):

```
# Арифметическая средняя:
```

```
mean(mtcars$mpg)
```

```
[1] 20.1
```

```
# Медиана
median(mtcars$mpg)
[1] 19.2

# Дисперсия:
var(mtcars$mpg)
[1] 36.3

# Стандартное отклонение:
sd(mtcars$mpg)
[1] 6.0
# Минимальное значение:
min(mtcars$mpg)
[1] 10.4

# Максимальное значение:
max(mtcars$mpg)
[1] 33.9
```

Специальной функции для расчета *стандартной ошибки средней* в R нет, однако для этого вполне подойдут уже имеющиеся функции. Как известно, стандартная ошибка средней рассчитывается как отношение стандартного отклонения к квадратному корню из объема выборки:

$$+S_{\bar{x}}+ = +\frac{S}{\sqrt{n}}$$

На языке R мы можем записать это следующим образом:

```
SEmpg = sd(mtcars$mpg)/sqrt(length(mtcars$mpg))
# функция length() возвращает число элементов в векторе mpg
```

Получаем:

```
SEmpg
[1] 1.065
```

Квантили рассчитываются в R при помощи функции `quantile()`:

```
quantile(mtcars$mpg)
```

```
     0%     25%     50%     75%    100% 
10.400 15.425 19.200 22.800 33.900 
```

При настройках, заданных по умолчанию, выполнение указанной команды приведет к расчету минимального (10.4) и максимального (33.9) значений, а также трех *квартилей*, т.е. значений, которые делят совокупность на четыре равные части - 15.4, 19.2 и 22.8.

Разница между первым и третим квартилями носит название *интерквартильный размах* (*ИКР;* англ. *interquartile range*). ИКР является *робастным* аналогом дисперсии и может быть рассчитан в R при помощи функции `IQR()`:

```
IQR(mtcars$mpg)
[1] 7.375
```

Функция `quantile()` позволяет рассчитать и другие квантили. Например, *децили*(т.е. значения, делящие совокупность на десять частей) можно получить следующим образом:

```
quantile(mtcars$mpg, p = seq(0, 1, 0.1))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
10.40 14.34 15.20 15.98 17.92 19.20 21.00 21.47 24.08 30.09 33.90 
```

В приведенной команде важен аргумент `p` (от *probability* - вероятность), при помощи которого был задан вектор чисел от 0 до 1 с шагом 0.1.

Обратите внимание на то, что существует несколько способов оценки квантилей по выборочным данным.

Подробнее об этом можно узнать в справочном файле по функции `quantile()` (доступен по команде `?quantile`). Отсутствующие значения в данных могут несколько усложнить вычисления.

В качестве демонстрации заменим 3-е значение переменной `mpg` на `NA` (от *not available* - не доступно) - обозначение, используемое в R для отсутствующих наблюдений, - а затем попытаемся вычислить среднее

значение:

```
mtcars$mpg[3] <- NA

# Просмотрим результат:
head(mtcars$mpg)
[1] 21.0 21.0   NA 21.4 18.7 18.1

# Попробуем рассчитать среднее значение для mpg:
mean(mtcars$mpg)
[1] NA
```

Ничего не вышло - вместо среднего значения программа выдала `NA`, что вполне логично. R не будет пропускать отсутствующие значения автоматически, если мы не включим соответствующую опцию - `na.rm` (от *not available* и *remove* - удалить):

```
mean(mtcars$mpg, na.rm = TRUE)
[1] 20.0
```

Рассмотренный прием срабатывает в большинстве случаев. Одним из немногих исключений является функция `length()`, используемая для определения размера вектора. Аргумент `na.rm` у этой функции отсутствует, так что подсчитаны будут и имеющиеся, и отсутствующие значения:

```
length(mtcars$mpg)
[1] 32
```

Если все же стоит задача подсчитать количество неотсутствующих значений, то можно воспользоваться следующим приемом:

```
sum(!is.na(mtcars$mpg))
[1] 31
```

Ключом здесь является использование команды `is.na(mtcars$mpg)`, которая проверяет каждое значение `mpg` и возвращает `FALSE`, если это значение *не* равно `NA`, и `TRUE` иначе. В сочетании с логическим оператором `!` ("не"), команда `sum` далее подсчитывает только те

значения mpg, которые не равны NA (логические TRUEздесь конвертируются в 1, которые можно суммировать).

Существуют еще две функции, которые могут оказаться полезными при анализе свойств совокупностей - which.min() и which.max(). Как следует из названий, эти функции позволяют выяснить порядковые номера элементов, обладающих минимальным и максимальным значениями соответственно. Если минимальное/максимальное значение принимают несколько элементов в векторе, то будет возвращен порядковый номер первого элемента с этим значением. В случае с mpg имеем:

```
which.min(mtcars$mpg)
[1] 15
which.max(mtcars$
mpg)
[1] 20
```

Видим, что минимальный и максимальный пробег в расчете на галлон топлива имеют модели под номерами 15 и 20 соответственно. Выяснить названия этих моделей мы можем, совместив команды which.min() и which.max() с командой rownames() (от *row* - строка, и *names* - имена):

```
rownames(mtcars)[which.min(mtcars$mpg)]
[1] "Cadillac Fleetwood"

rownames(mtcars)[which.max(mtcars$mpg
)]
[1] "Toyota Corolla"
```

Использование функции summary()

В системе R имеется возможность и более быстрого расчета основных параметров описательной статистики. Для этого, в частности, служит *функция общего назначения* summary():

```
summary(mtcars$mpg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  10.40   15.35   19.20   20.00   22.15   33.90    1.00
```

Всего одной строки кода оказалось достаточно для получения минимального (`Min`) и максимального (`Max`) значений переменной `mpg`, медианы (`Median`), арифметической средней (`Mean`), первого (`1st Qu.`) и третьего (`3rd Qu.`) квартилей, а также для выяснения количества отсутствующих значений (`NA's`). Более того, подобную сводку мы можем получить сразу для всей таблицы данных:

```
summary(mtcars)
      mpg              cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.35   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.00   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.15   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
NA's   : 1.00                                                  
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

Результат выглядел бы замечательно, если бы не одно "но". Переменные `vs` и `am` являются факторами, но уровни их закодированы при помощи *чисел* 0 и 1. К сожалению, система R не распознала эти две переменные как факторы

и рассчитала соответствующие параметры описательной статистики, как для обычных числовых переменных. Однако мы можем изменить такое поведение R, самостоятельно преобразовав `vs` и `am` в факторы при помощи функции `as.factor()`:

```
mtcars$vs <- as.factor(mtcars$vs)
  mtcars$am <-
as.factor(mtcars$am)

# Проверим, удалась ли конвертация:
is.factor(mtcars$vs)
[1] TRUE
is.factor(mtcars
$am)
[1] TRUE
```

Теперь результат действия функции `summary()` в отношении таблицы `mtcars` будет выглядеть так:

```
summary(mtcars)
      mpg             cyl             disp             hp
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
 1st Qu.:15.35   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
 Median :19.20   Median :6.000   Median :196.3   Median :123.0
 Mean   :20.00   Mean   :6.188   Mean   :230.7   Mean   :146.7
 3rd Qu.:22.15   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
NA's   : 1.00
      drat             wt             qsec       vs      am
 Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   0:19
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   1:13
 Median :3.695   Median :3.325   Median :17.71
 Mean   :3.597   Mean   :3.217   Mean   :17.85
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90                Max.
:4.930   Max.   :5.424   Max.   :22.90
      gear            carb
 Min.   :3.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:2.000
```

```
Median :4.000   Median :2.000
Mean   :3.688   Mean   :2.812
```

Обратите внимание на сводки по `vs` и `am`: поскольку эти переменные теперь распознаны программой как факторы, единственный способ описать их - это подсчитать количество наблюдений для каждого уровня.

Использование функции `tapply()`

Функция `tapply()` принадлежит к важному "`apply`-семейству" R-функций. Эти функции позволяют выполнять математические вычисления над определенными элементами таблиц данных, матриц, или массивов (например, быстро вычислять среднее значение для каждого столбца или строки таблицы, и т.п.).

Предположим, мы хотим выяснить средний объем двигателя (переменная `disp`, в кубических дюймах) у моделей с автоматической и ручной коробкой передач (переменная `am`; `1` - ручная коробка, `0` - автоматическая коробка). Функция `tapply()` позволяет сделать это следующим образом:

```
tapply(X = mtcars$disp, INDEX = mtcars$am, FUN = mean)
     0          1
290.38     143.53
```

Как видно из приведенной команды, основными аргументами функции `tapply()` являются:

- X - числовой вектор
- INDEX - список факторов, для уровней которых рассчитываются значения функции
- FUN - любая, в том числе пользовательская, функция

Поскольку аргумент `INDEX` способен принимать список из нескольких факторов, мы можем усложнить приведенную выше команду:

```
tapply(X = mtcars$disp, INDEX = list(mtcars$am, mtcars$vs), FUN = mean)
       0      1
0 357.62 175.11
1 206.22  89.80
```

Аргумент FUN, как уже было отмечено, может принимать любые, в том числе и пользовательские функции. Рассмотрим, например, расчет стандартных ошибок для средних значений объема двигателя у автомобилей с автоматической и ручной коробкой передач. Для начала создадим функцию SE для расчета стандартных ошибок (см. пример выше):

```
SE <- function(x) {sd(x)/sqrt(length(x))}
```

Теперь совместим эту новую функцию с tapply():

```
tapply(X = mtcars$disp, INDEX = mtcars$am, FUN = SE)
   0    1
25.3 24.2
```

Таким образом, объем двигателя у моделей с автоматической коробкой передач составляет в среднем 290.4 ± 25.3, а у автомобилей с механической коробкой - 143.5 ± 24.2 кубических дюймов.

Использование возможностей дополнительных пакетов

Рассмотренные выше функции позволяют получить достаточно полное представление об анализируемых выборках и таблицах данных. Однако специальные функции для расчета некоторых параметров описательной статистики не входят в базовую версию R. С одним из таких параметров мы уже столкнулись - стандартная ошибка арифметической средней. Другие примеры включают коэффициенты *эксцесса* (англ. *kurtosis*) и *асимметрии* (*skewness*) - параметры, характеризующие форму распределения. Конечно, мы можем рассчитать эти величины по соответствующим формулам или даже написать собственные функции для этих целей. Однако это уже было сделано до нас - достаточно воспользоваться имеющимися дополнительными пакетами для R, например, пакетом moments. Если этот пакет не установлен на Вашем компьютере, выполните следующую команду (естественно, Ваш

компьютер должен быть при этом подключен к *Internet*):

```
install.packages("moments")
```

Рассчитать коэффициенты эксцесса и асимметрии теперь очень просто:

```
library(moments)  #загрузка пакета moments
  kurtosis(mtcars$mpg,  na.rm  =
TRUE)
[1] 2.79    skewness(mtcars$mpg,
na.rm = TRUE)
[1] 0.68
```

Многие R-пакеты имеют собственные функции, аналогичные стандартной `summary()`, для вывода компактных описательных сводок по таблицам данных. Ниже приведены несколько примеров таких пакетов и функций (предполагается, что соответствующий пакет уже установлен на Вашем компьютере и загружен в рабочее пространство R; подробности вывода результатов анализа здесь не обсуждаются - см. справочные материалы по соответствующим командам).

```
# Пакет Hmisc, функция describe():
  describe(mtcars)
 # Пакет pastecs, функция stat.desc()
  stat.desc(mtcars)
# Пакет psych, функция describe()
  describe(mtcars)
# Пакет psych, функция describe.by() - расчет параметров описательной
статистики
# для каждого уровня некоторого фактора:
  describe.by(mtcars, mtcars$am)
# Пакет doBy. Обладает мощным функционалом. Пример приведен ниже:
  summaryBy(mpg + wt ~ cyl + vs, data = mtcars,     FUN = function(x) { c(m
= mean(x), s = sd(x)) } )
```