

МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА  
И ПРОДОВОЛЬСТВИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

ГЛАВНОЕ УПРАВЛЕНИЕ ОБРАЗОВАНИЯ,  
НАУКИ И КАДРОВОЙ ПОЛИТИКИ

Учреждение образования  
«БЕЛОРУССКАЯ ГОСУДАРСТВЕННАЯ  
ОРДЕНОВ ОКТЯБРЬСКОЙ РЕВОЛЮЦИИ  
И ТРУДОВОГО КРАСНОГО ЗНАМЕНИ  
СЕЛЬСКОХОЗЯЙСТВЕННАЯ АКАДЕМИЯ»

Н. В. Барулин, К. Л. Шумский

# **ФУНДАМЕНТАЛЬНЫЕ И ПРИКЛАДНЫЕ НАУЧНЫЕ ИССЛЕДОВАНИЯ В АКВАКУЛЬТУРЕ**

**В трех частях**

**Часть 1**

## **ИСПОЛЬЗОВАНИЕ ПРОГРАММНОЙ СРЕДЫ R ПРИ СТАТИСТИЧЕСКОМ АНАЛИЗЕ**

*Рекомендовано учебно-методическим объединением  
по образованию в области сельского хозяйства в качестве  
учебно-методического пособия для студентов учреждений,  
обеспечивающих получение высшего образования II ступени  
по специальности 1-74 80 03 Зоотехния*

Горки  
БГСХА  
2022

УДК 639.3:519.22(075.8)

ББК 47.2я73

Б24

*Рекомендовано методической комиссией факультета  
биотехнологии и аквакультуры 30.03.2021 (протокол № 7)  
и Научно-методическим советом БГСХА 31.03.2021 (протокол № 7)*

Авторы:

кандидат сельскохозяйственных наук, доцент *Н. В. Барулин*;  
кандидат сельскохозяйственных наук *К. Л. Шумский*

Рецензенты:

кандидат биологических наук, доцент *В. Г. Костоусов*;  
доктор сельскохозяйственных наук, доцент *Е. В. Таразевич*

**Барулин, Н. В.**

Б24

Фундаментальные и прикладные научные исследования в аквакультуре : учебно-методическое пособие. В 3 ч. Ч. 1. Использование программной среды R при статистическом анализе / Н. В. Барулин, К. Л. Шумский. – Горки : БГСХА, 2022. – 102 с.  
ISBN 978-985-882-256-9.

Приведен минимальный набор методов для выполнения статистического анализа при обработке данных рыбохозяйственных исследований.

Для студентов учреждений, обеспечивающих получение высшего образования II ступени по специальности 1-74 80 03 Зоотехния.

УДК 639.3:519.22(075.8)

ББК 47.2я73

ISBN 978-985-882-256-9 (ч. 1)  
ISBN 978-985-882-255-2

© УО «Белорусская государственная  
сельскохозяйственная академия», 2022

## ВВЕДЕНИЕ

Программная среда R является бесплатной альтернативой современным платным статистическим программам. Сегодня R является безусловным лидером среди свободно распространяемых систем статистического анализа, о чем свидетельствует, например, тот факт, что в 2010 г. система R стала победителем ежегодного конкурса открытых программных продуктов Bossie Awards в нескольких номинациях. Ведущие университеты мира, аналитики крупнейших компаний и исследовательских центров постоянно используют R при проведении научно-технических расчетов и создании крупных информационных проектов.

Данное учебно-методическое пособие представляет собой руководство по работе в программе R с использованием пакета R Commander – одного из сотни статистических пакетов R, а также с использованием некоторых других пакетов. В данном издании описываются минимальные базовые статистические методики для обработки биологической информации. Руководство не включает в себя такие классические методы, как критерий хи-квадрат, многофакторный дисперсионный анализ, корреляционные и регрессионные методы и множество других.

Авторы настоятельно рекомендуют ознакомиться более подробно с основами работы в программе R, изложенными в электронной книге С. Э. Мاستицкого, В. К. Шитикова «Статистический анализ и визуализация данных с помощью R» (2014). Адрес доступа: <http://r-analytics.blogspot.com>.

Часть информации, представленной в данном пособии, была взята из вышеназванной книги.

Также авторы рекомендуют ознакомиться с книгой Г. Джеймса, Д. Уиттона, Т. Хасты, Р. Тибширани «Введение в статистическое обучение с примерами на языке R» (2016). Адрес доступа: <http://dmkpress.com/catalog/computer/statistics/978-5-97060-293-5/>.

**Важно!** При написании научных статей обязательно приводите сведения о том, какой тест и для каких целей вы использовали в своих исследованиях. Фразы *«В исследованиях нами использовались стандартные статистические методы»* или *«Статистическая обработка проводилась по общепринятой методике и в программе «X»* недопустимы, так как стандартных статистических методов не существует,

а описанные в данном пособии методы отражают только один из сотен подходов в обработке статистической информации. Игнорирование этого положения дает основание специалисту, при прочтении ваших работ, сомневаться во владении автором методами статистической обработки, а также в достоверности полученных данных.

## 1. ИСПОЛЬЗОВАНИЕ ПРОГРАММНОЙ СРЕДЫ R ПРИ СТАТИСТИЧЕСКОМ АНАЛИЗЕ

Программу R можно скачать по прямой ссылке <https://mirror.truenetwork.ru/CRAN/bin/windows/base/R-4.2.2-win.exe> или установить из файла R-4.2.2.zip (сам файл можно запросить у автора по e-mail: barulin@list.ru). После распаковки файла R-4.2.2.zip необходимо открыть папку bin затем папку i386 (рис. 1.1). Следует отметить, что на программу очень часто появляются обновления. Обновления можно скачать с сайта <https://cran.r-project.org>.

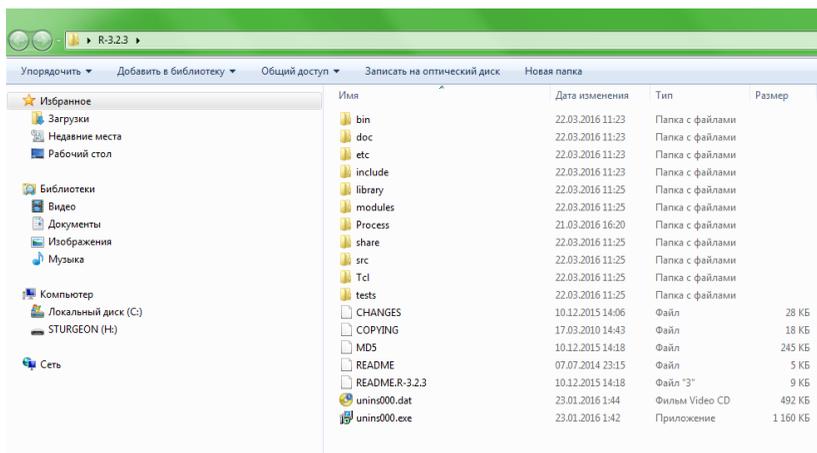


Рис. 1.1. Диалоговое окно папки bin

Запускаем файл Rgui.exe (рис. 1.2).

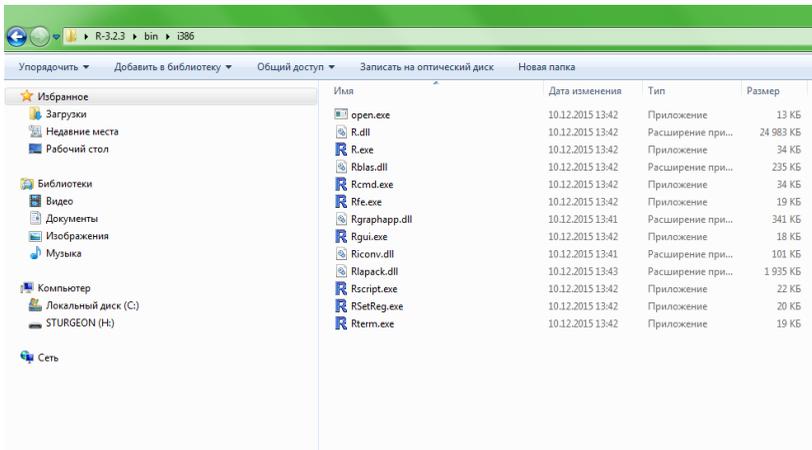


Рис. 1.2. Диалоговое окно папки i386

Откроется диалоговое окно консоли R (рис. 1.3).

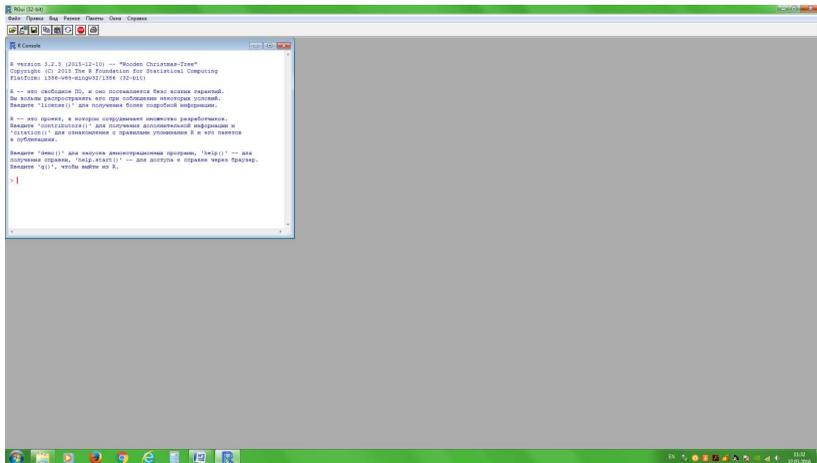


Рис. 1.3. Диалоговое окно консоли R

Удобным средством освоения вычислений в R для начинающего пользователя является R Commander – платформонезависимый графический интерфейс в стиле кнопочного меню, реализованный в

пакете Rcmdr. Он позволяет осуществить большой комплект процедур статистического анализа, не прибегая к предварительному заучиванию функций на командном языке, однако невольно способствует этому, поскольку отображает все выполняемые инструкции в специальном окне.

Если вы скачивали программу напрямую с сайта <https://cran.r-project.org>, то для запуска пакета R Commander вам понадобится дополнительно установить этот пакет, используя меню **Пакеты** → **Установить пакет(ы)** → **Russia (Moscow) [https]** → **Rcmdr**.

Для установки пакета R Commander можно также просто выполнить команду

```
install.packages("Rcmdr", dependencies=TRUE),
```

где включение опции `dependencies` вызовет гарантированную установку полного комплекта остальных пакетов, которые могут потребоваться при обработке данных через меню Rcmdr. После установки пакета его надо запустить, нажав **Включить пакет** (рис. 1.4) и выбрав Rcmdr (рис. 1.5).

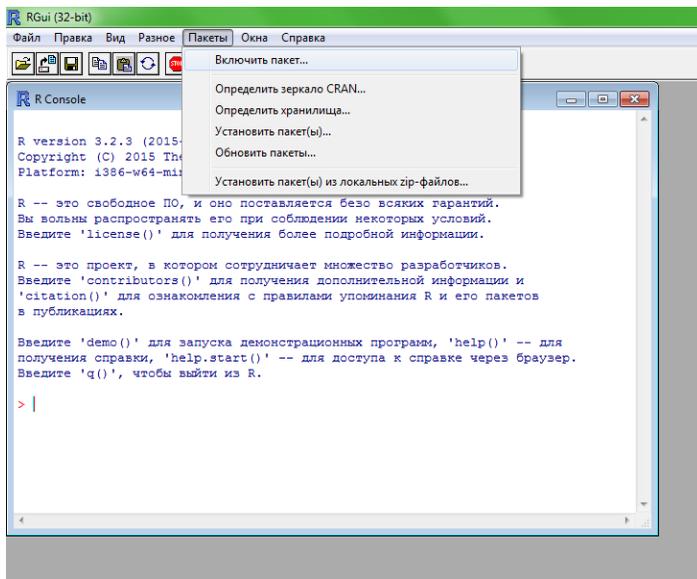


Рис. 1.4. Диалоговое окно, показывающее момент включения пакета

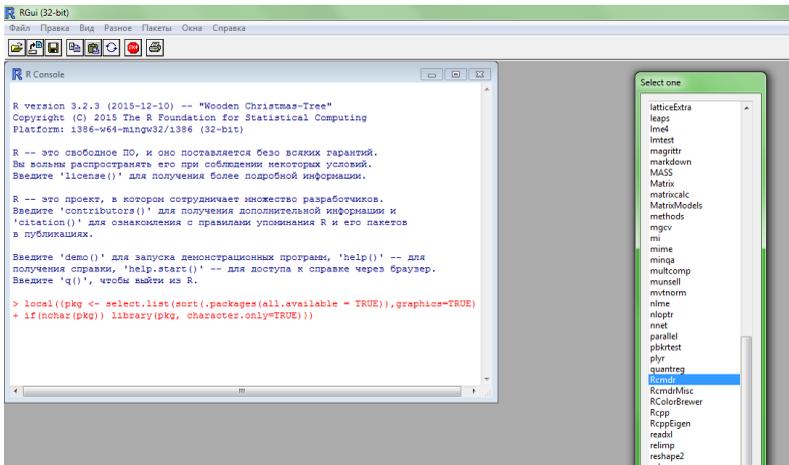


Рис. 1.5. Диалоговое окно, показывающее момент выбора пакета Rcmdr

Откроется диалоговое окно пакета R Commander (рис. 1.6).

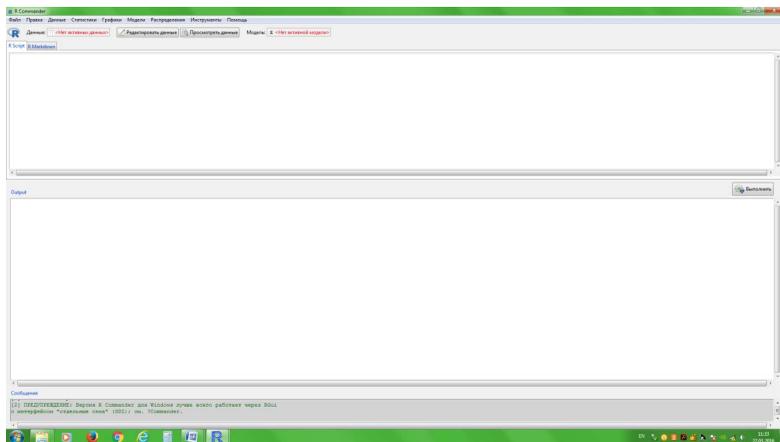
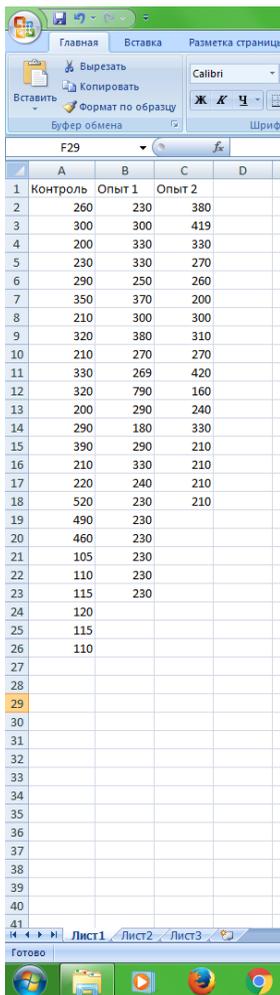


Рис. 1.6. Диалоговое окно пакета R Commander

Работу в R Commander рассмотрим на примере результатов взвешивания массы личинок стерляди контрольной и опытных групп (*данные имитированы*), на которых оказывалось (Опыт1, Опыт2) или не

оказывалось (Контроль) воздействие фактором X. Такие данные можно предварительно подготовить в Excel (рис. 1.7), в блокноте или в другой программе.



The image shows a screenshot of the Microsoft Excel application window. The ribbon at the top includes 'Главная', 'Вставка', and 'Разметка страницы'. The 'Буфер обмена' (Clipboard) group is active, showing 'Вырезать', 'Копировать', 'Вставить', and 'Формат по образцу'. The font settings are set to 'Calibri' with bold, italic, and underline options. The active cell is F29. The spreadsheet contains the following data:

	A	B	C	D
1	Контроль	Опыт 1	Опыт 2	
2		260	230	380
3		300	300	419
4		200	330	330
5		230	330	270
6		290	250	260
7		350	370	200
8		210	300	300
9		320	380	310
10		210	270	270
11		330	269	420
12		320	790	160
13		200	290	240
14		290	180	330
15		390	290	210
16		210	330	210
17		220	240	210
18		520	230	210
19		490	230	
20		460	230	
21		105	230	
22		110	230	
23		115	230	
24		120		
25		115		
26		110		
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				

Рис. 1.7. Диалоговое окно, показывающее принцип построения данных для базовой статистики

Первый этап – загрузка нового набора данных: выбираем из меню **Данные** → **Импорт данных** → **из файла Excel** (рис. 1.8).

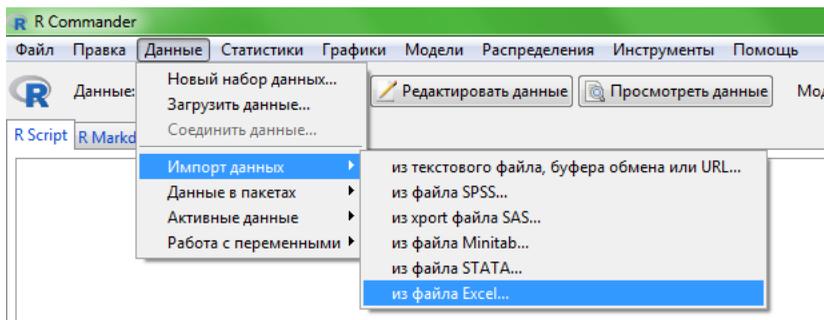


Рис. 1.8. Диалоговое окно, показывающее процесс импорта данных из файла Excel

Определяем во всплывающих окнах режим загрузки данных. Нетрудно заметить, что те же данные можно было легко загрузить из локального текстового файла или таблицы базы данных. В случае наличия имен групп в первой строке таблицы необходимо поставить галочку в соответствующей строке (рис. 1.9).

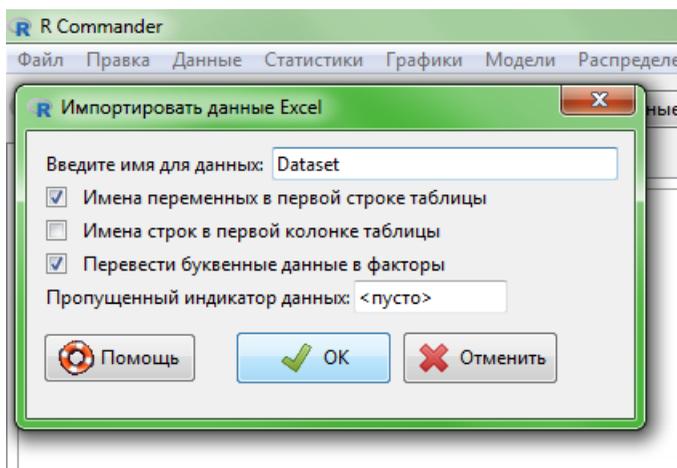


Рис. 1.9. Диалоговое окно, показывающее процесс импорта данных из файла Excel в случае наличия имен групп в первой строке таблицы

Чтобы убедиться в том, что наши данные загружены верно (и при необходимости их отредактировать), нажимаем кнопку **Посмотреть данные**.

Вначале необходимо рассчитать в выборке такие базовые элементы, как среднее (mean), стандартное отклонение (sd), стандартная ошибка среднего (se(mean)), коэффициент вариации (cv). Для этого необходимо нажать кнопки **Итоги** → **Базовые статистики** (рис. 1.10), затем выбрать переменные и в колонке **Статистики** (рис. 1.11) выбрать нужные критерии (рис. 1.12).

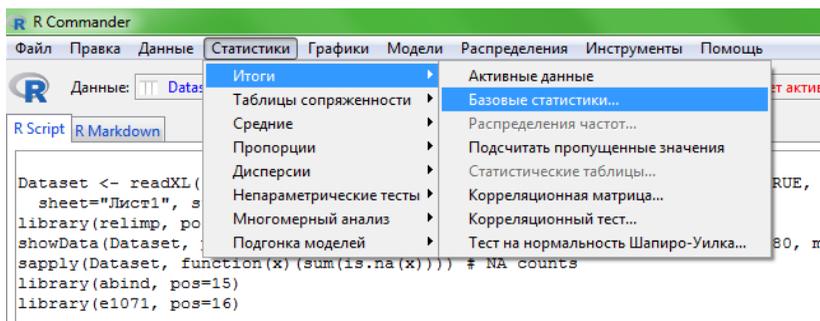


Рис. 1.10. Диалоговое окно, показывающее процесс включения функции «Базовые статистики»

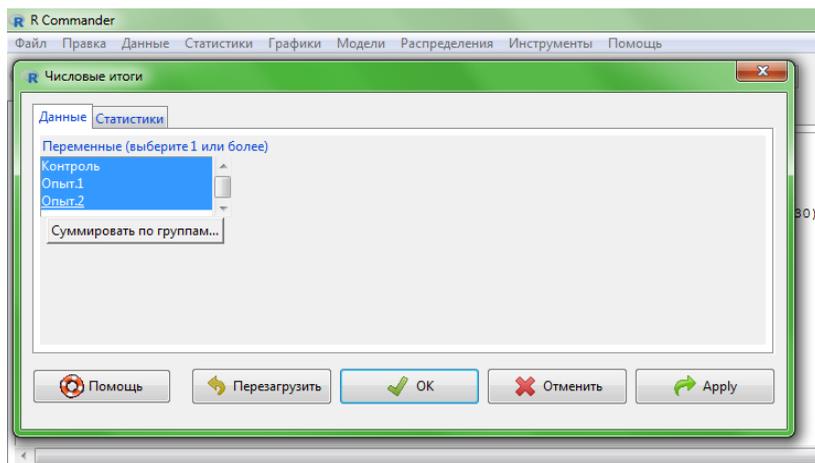


Рис. 1.11. Диалоговое окно, показывающее процесс выбора переменных

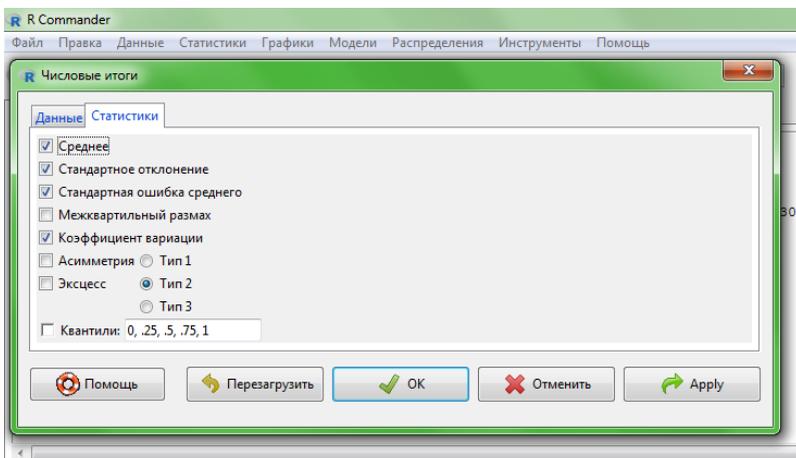


Рис. 1.12. Диалоговое окно, показывающее процесс выбора критериев базовой статистики

Получаем следующие значения (рис. 1.13).

```

              mean      sd se (mean)      cv  n
Контроль 259.0000 120.56292 24.11258 0.4654939 25
Опыт.1   296.7727 121.64191 25.93414 0.4098824 22
Опыт.2   278.1765  78.16828 18.95859 0.2810025 17

```

Рис. 1.13. Пример вывода рассчитанных значений в результате анализа данных при помощи функции «Базовые статистики»

Полученные при исследовании данные представляются, как правило, в научной или дипломной работе, статье, научном отчете в виде таблицы (табл. 1.1).

Таблица 1.1. Пример оформления полученных данных, рассчитанных при помощи функции «Базовые статистики»

Группа	Mean $\pm$ SE, мг	SD	CV, %	n
Контрольная	259,0 $\pm$ 24,1	120,6	0,5	25
Опытная № 1	296,8 $\pm$ 25,9	121,6	0,4	22
Опытная № 2	278,2 $\pm$ 18,9	78,2	0,3	17

Примечание. Mean – среднее значение массы; SE – стандартная ошибка среднего; SD – стандартное отклонение; CV – коэффициент вариации, %; n – объем выборки.

После работы с базовой статистикой необходимо определить, насколько полученные различия между группами достоверны, т. е. насколько им можно доверять.

Вначале необходимо имеющиеся данные перевести в форму двух колонок по нижепредставленному примеру (рис. 1.14).

	А	В	С
1	Контроль	260	
2	Контроль	300	
3	Контроль	200	
4	Контроль	230	
5	Контроль	290	
6	Контроль	350	
7	Контроль	210	
8	Контроль	320	
9	Контроль	210	
10	Контроль	330	
11	Контроль	320	
12	Контроль	200	
13	Контроль	290	
14	Контроль	390	
15	Контроль	210	
16	Контроль	220	
17	Контроль	520	
18	Контроль	490	
19	Контроль	460	
20	Контроль	105	
21	Контроль	110	
22	Контроль	115	
23	Контроль	120	
24	Контроль	115	
25	Контроль	110	
26	Опыт 1	230	
27	Опыт 1	300	
28	Опыт 1	330	
29	Опыт 1	330	
30	Опыт 1	250	
31	Опыт 1	370	
32	Опыт 1	300	
33	Опыт 1	380	
34	Опыт 1	270	
35	Опыт 1	269	
36	Опыт 1	790	
37	Опыт 1	290	
38	Опыт 1	180	
39	Опыт 1	290	
40	Опыт 1	330	
41	Опыт 1	240	

Рис. 1.14. Диалоговое окно, показывающее принцип построения данных для определения критериев достоверности

На рис. 1.15 представлен алгоритм простейшего статистического анализа для выявления различий между исследуемыми группами и определения уровня значимости полученных различий.

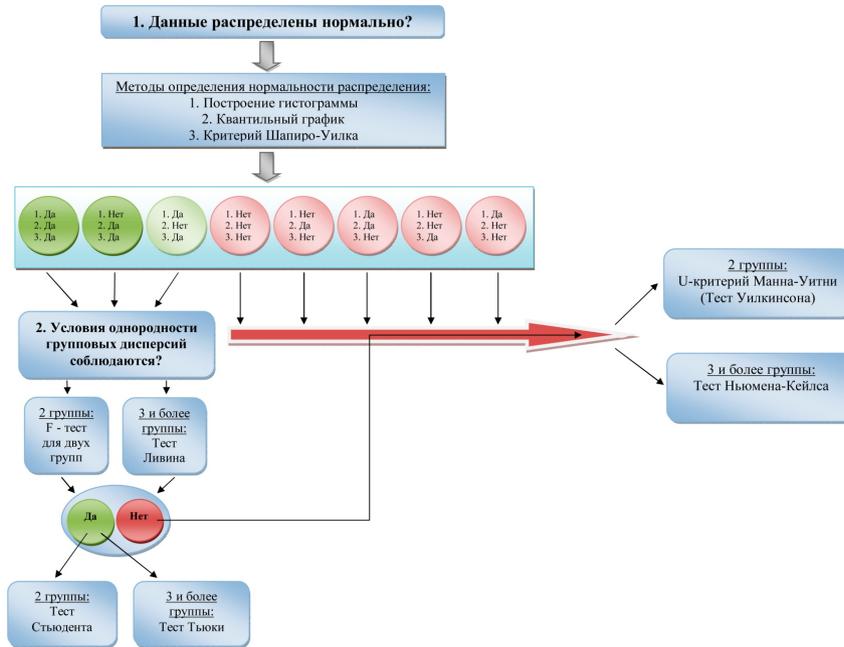


Рис. 1.15. Алгоритм простейшего статистического анализа для выявления различий между исследуемыми группами и определения уровня значимости полученных различий

Прежде чем установить, насколько полученные различия между группами достоверны, т. е. насколько им можно доверять, мы должны выбрать тест, который поможет нам в этом.

Для этого исследователю необходимо ответить на два вопроса:

1. Данные распределены нормально?
2. Условия однородности групповых дисперсий соблюдаются?

Чтобы ответить на первый вопрос, необходимо осуществить проверку на нормальность распределения.

### **1.1. Проверка на нормальность распределения**

Проверка исследуемых переменных на нормальность распределения является важной составной частью статистического анализа данных. Существует несколько способов такой проверки, и их можно разделить на две рассмотренные ниже группы.

#### ***Графические способы.***

Самый простой графический способ проверки характера распределения данных – это построение гистограммы. 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 выборок, извлеченных из смеси нормально распределенных совокупностей, гистограммы приобретают асимметричный вид, вводя исследователя в заблуждение.

Для построения гистограммы полученных данных нажимаем кнопки **Графики** → **Гистограмма** (рис. 1.16), выбираем исследуемую переменную (рис. 1.17) и получаем результат.

Как видно из рис. 1.18, гистограмма имеет колоколообразный, но недостаточно симметричный вид. И мы не можем с достаточной уверенностью утверждать, что наши данные нормально распределены. Ответ по тесту нормальности распределения гистограммой – **нет**.

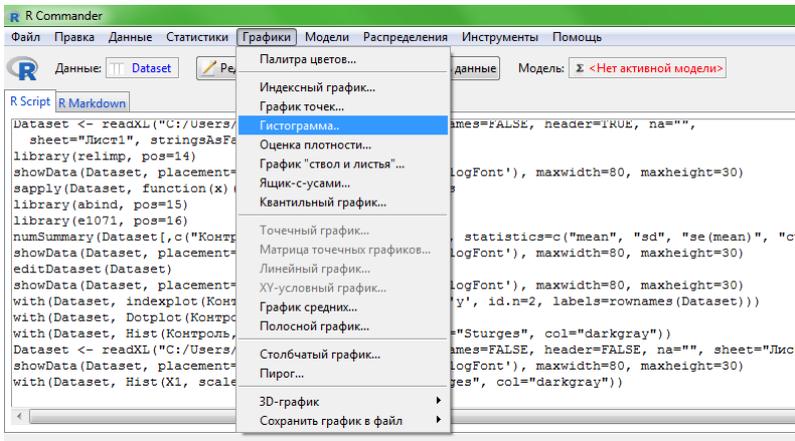


Рис. 1.16. Диалоговое окно, показывающее процесс построения гистограммы

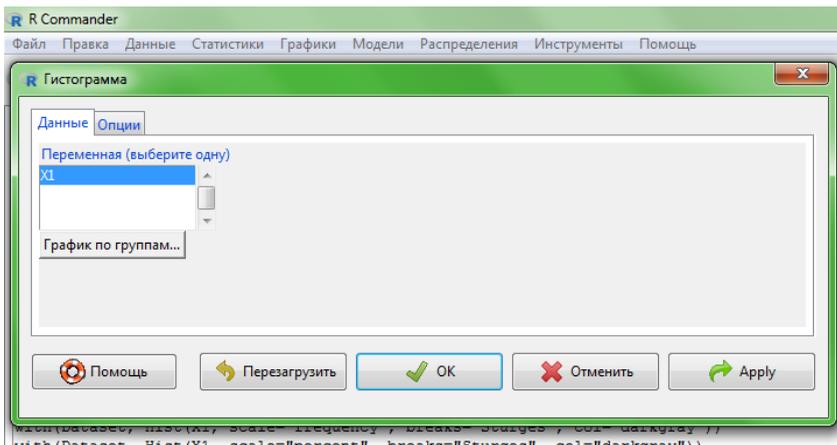


Рис. 1.17. Диалоговое окно, показывающее процесс выбора переменной при построении гистограммы

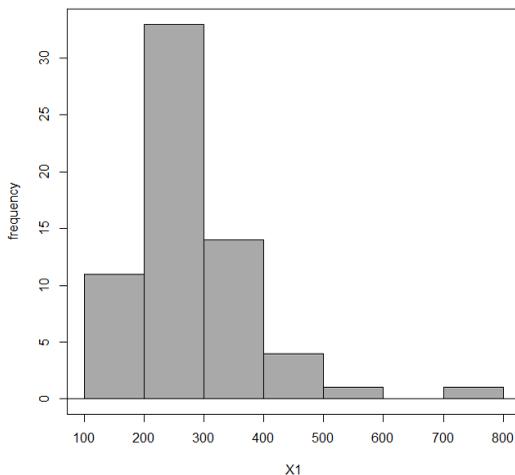


Рис. 1.18. Колоколообразная несимметричная гистограмма

Образец колоколообразной симметричной гистограммы представлен на рис. 1.19.

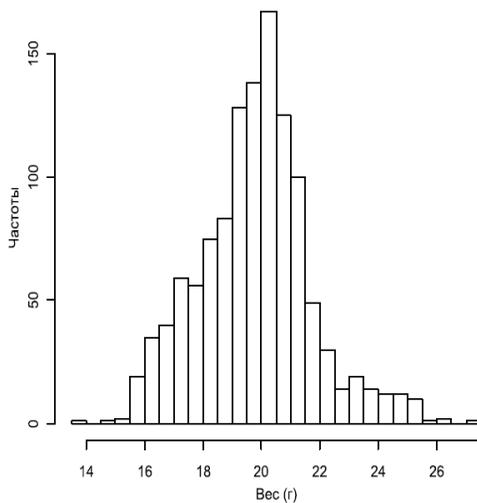


Рис. 1.19. Колоколообразная симметричная гистограмма

Другим очень часто используемым графическим способом проверки характера распределения данных является построение так называемых графиков квантилей (*q-q plots, quantile-quantile 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 углом  $45^\circ$  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.

Интерпретация квантильных графиков носит в значительной мере субъективно-эмоциональный характер. Например, одни исследователи могут полагать, что экспериментальные точки в значительной степени отклоняются от диагонали теоретических квантилей, а другие на том же графике сочтут эти отклонения достаточно приемлемыми для принятия гипотезы о нормальности.

Для построения квантильного графика наших данных нажимаем кнопки **Графики** → **Квантильный график** (рис. 1.20), выбираем исследуемую переменную и получаем результат.

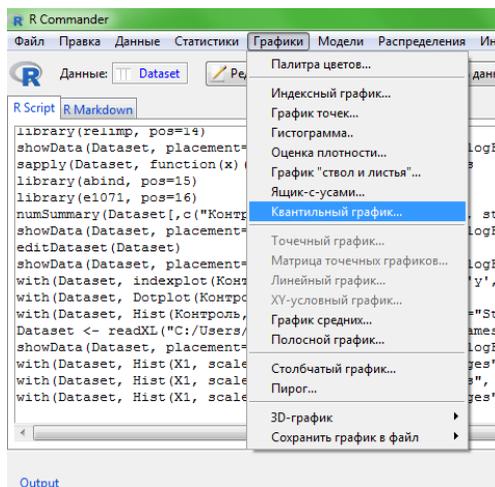


Рис. 1.20. Диалоговое окно, показывающее процесс выбора квантильного графика

На графике, представленном для нашего примера на рис. 1.21, видно, что большинство точек находятся в пределах доверительной полосы, что позволяет нам сделать вывод о том, что данные распределены нормально. Наш ответ по тесту нормальности распределения квантильным графиком – да.

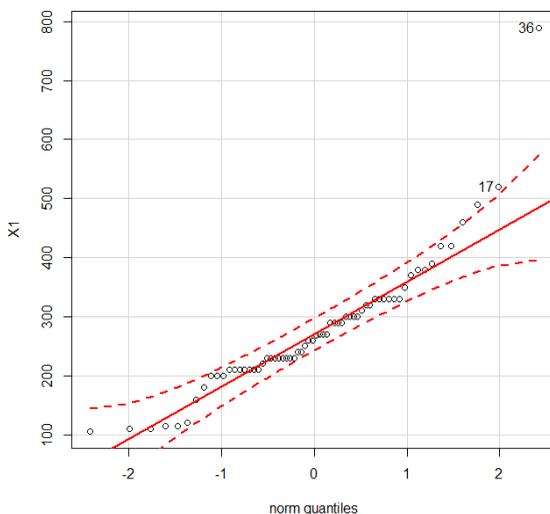


Рис. 1.21. Квантильный график с нормальным распределением

Пример квантильного графика с ненормальным распределением представлен на рис. 1.22.

### **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 сформулировать 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 стандартных функций, либо в виде функций, входящих в состав подгружаемых пакетов.

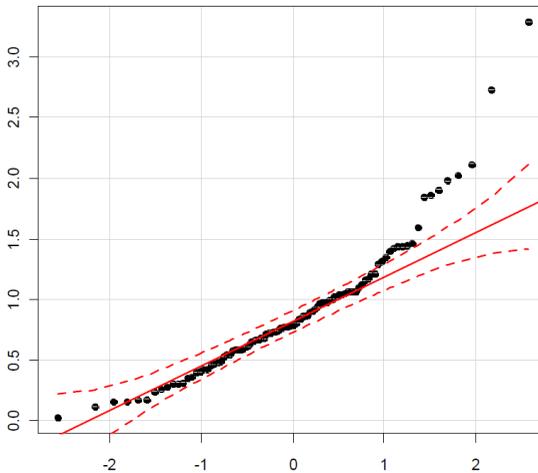


Рис. 1.22. Квантильный график с ненормальным распределением

Примером базовой функции является `shapiro.test()`, при помощи которой можно выполнить широко используемый тест Шапиро – Уилка (Статистики → Итоги → Тест на нормальность Шапиро – Уилка) (рис. 1.23).

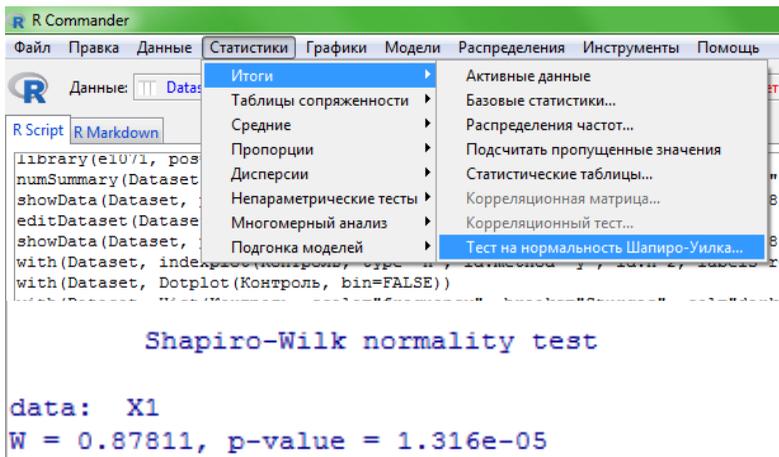


Рис. 1.23. Диалоговые окна, показывающие процесс выбора теста Шапиро – Уилка и вид представленных значений

Как видно из представленных данных, вероятность ошибки  $p$  ( $p$ -value) значительно больше некоторого заранее принятого уровня значимости (например, 0,05). Наш ответ на проверку нормальности распределения тестом Шапиро – Уилка – **нет**. Если бы  $p$ -value было больше 0,05 – ответ был бы **да**.

## 1.2. Проверка на однородность групповых дисперсий

Проведенные тесты на проверку нормальности распределения данных позволяют заключить, что наши данные распределены ненормально.

На данном этапе можно сразу переходить к непараметрическим критериям (U-критерий Манна – Уитни (Тест Уилкоксона) и критерий парных множественных сравнений средних рангов (тест Ньюмена – Кейлса)).

В случае если бы данные были распределены нормально, необходимо было бы перейти к ответу на второй вышестоящий вопрос: «Условия однородности групповых дисперсий соблюдаются?».

В качестве примера продолжим дальнейшее определение однородности групповых дисперсий. Для ответа на вышестоящий вопрос можно задействовать следующие тесты: F-тест для двух дисперсий (используется, если имеются две исследуемые группы, например контрольная и опытная) и тест Ливина (используется, если имеются три исследуемые группы и более).

Представим, что изначально было только две исследуемые группы – контрольная и опытная группа № 1 (*проводить попарные тесты для реальных исследований, в которых имеется множество исследуемых групп, недопустимо*).

Нажимаем последовательно **Статистики** → **Дисперсии** → **F-тест для двух дисперсий** (рис. 1.24).

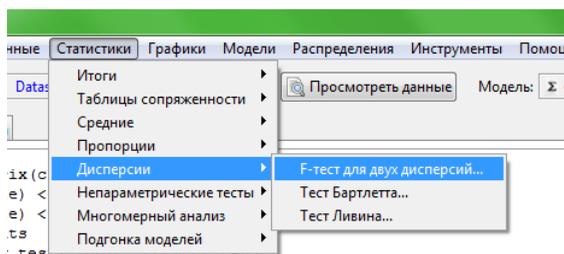


Рис. 1.24. Диалоговое окно, показывающее процесс выбора F-теста для двух дисперсий

Указываем группирующую и зависимую переменные (рис. 1.25).

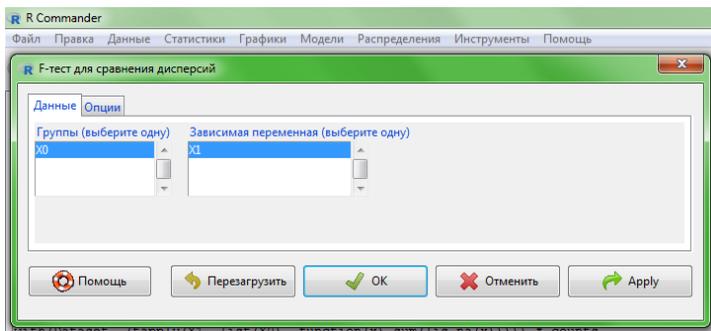


Рис. 1.25. Диалоговое окно, показывающее процесс выбора группирующих и зависимых переменных

Получаем результат (рис. 1.26).

```
F test to compare two variances

data: X1 by X0
F = 0.91775, num df = 24, denom df = 19, p-value = 0.8317
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3742385 2.1522717
sample estimates:
ratio of variances
 0.9177529
```

Рис. 1.26. Диалоговое окно, показывающее результаты расчета F-теста для двух дисперсий

В итоге получилось значение  $p$ -value, равное 0,8317, что значительно превышает уровень значимости 0,05, т. е. значительно превышает 5%-ный уровень значимости, на основании чего мы не можем отклонить нулевую гипотезу о равенстве дисперсий в исследованных совокупностях, т. е. наш ответ на вопрос: «Условия однородности групповых дисперсий соблюдаются?» – **да**. Если бы  $p$ -value было меньше 0,05 – ответ был бы **нет**.

Проанализируем выборку, в которой имеется три исследуемые группы (контрольная группа, опытная группа № 1, опытная группа № 2). Для этого будем использовать тест Ливина.

Нажимаем последовательно **Статистики** → **Дисперсии** → **Тест Ливина** (рис. 1.27).

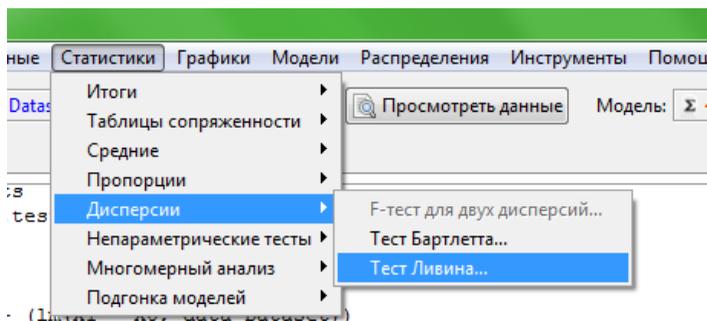


Рис. 1.27. Диалоговое окно, показывающее процесс выбора теста Ливина

Получаем результат (рис. 1.28).

```
Levene's Test for Homogeneity of Variance (center = "median")
  Df F value Pr(>F)
group 2      1.17 0.3172
     61
```

Рис. 1.28. Диалоговое окно, показывающее результаты расчета теста Ливина

В итоге получается значение  $P_r$ , равное 0,3172, что превышает уровень значимости 0,05, т. е. превышает 5%-ный уровень значимости, на основании чего мы не можем отклонить нулевую гипотезу о равенстве дисперсий в исследованных совокупностях, т. е. наш ответ на вопрос: «Условия однородности групповых дисперсий соблюдаются?» – **да**. Если бы  $P_r$  было меньше 0,05 – ответ был бы **нет**.

На основании проведенной оценки однородности групповых дисперсий можно сделать вывод о том, что дисперсии однородны. Это позволяет, наконец, перейти к установлению того, насколько полученные различия между группами достоверны, т. е. насколько им можно доверять. Читатель должен помнить о том, что анализируемые данные показали отрицательные результаты при проверке на нормальность распределения и к ним должны быть применены непараметрические критерии. Но в качестве примера мы также проанализируем наши данные на параметрические критерии. **Однако в реальных исследованиях это делать недопустимо!**

### 1.3. Параметрические критерии

Для установления того, насколько полученные различия между группами достоверны, можно обратиться к параметрическим тестам, которые включают в себя тест Стьюдента (только для двух исследуемых групп) и тест Тьюки (для трех исследуемых групп и более).

В случае неоднородности данных необходимо переходить к непараметрическим критериям (U-критерий Манна – Уитни (Тест Уилкоксона) и критерий парных множественных сравнений средних рангов (тест Ньюмена)).

Снова представим, что изначально у нас было только две исследуемые группы – контрольная и опытная группа № 1 (*еще раз отметим, что проводятся попарные тесты для реальных исследований, в которых имеется множество исследуемых групп, недопустимо*). Воспользуемся тестом Стьюдента.

Для этого последовательно нажимаем **Статистики** → **Средние** → **t-тест для независимых выборок** (рис. 1.29).

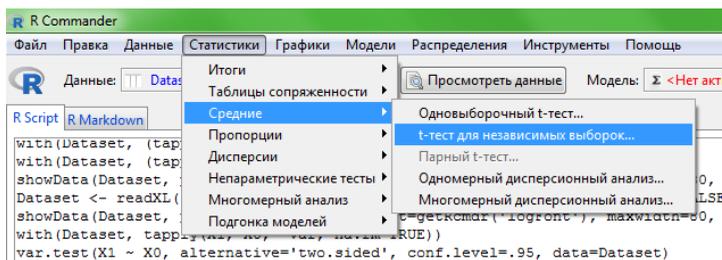


Рис. 1.29. Диалоговое окно, показывающее процесс выбора t-теста для независимых выборок

Получаем результат (рис. 1.30).

```
Welch Two Sample t-test

data: X1 by X0
t = -1.1995, df = 40.05, p-value = 0.2374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -119.34473  30.44473
sample estimates:
mean in group Контроль   mean in group Опыт 1
      259.00             303.45
```

Рис. 1.30. Диалоговое окно, показывающее результаты расчета t-теста для независимых выборок

Как видно из представленных данных, вероятность ошибки  $p$  ( $p$ -value) больше заранее принятого уровня значимости (0,05), т. е. существенных различий между исследуемыми выборками нет. И, несмотря на более высокую среднюю массу в опытной группе (303,4 мг) по сравнению с контрольной (259,0 мг), эти различия являются недостоверными.

Проанализируем выборку, в которой имеется три исследуемые группы (контрольная группа, опытная группа № 1, опытная группа № 2). Для этого будем использовать тест Тьюки.

Нажимаем последовательно **Статистики** → **Средние** → **Одномерный дисперсионный анализ** (рис. 1.31).

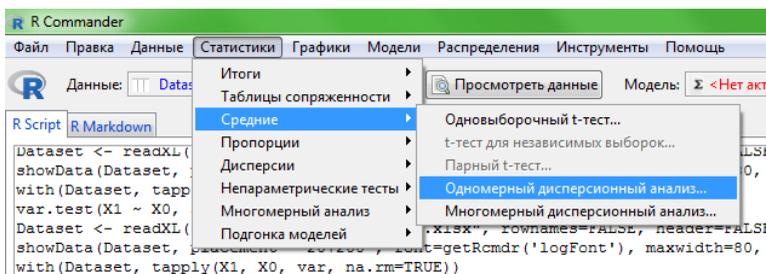


Рис. 1.31. Диалоговое окно, показывающее процесс выбора одномерного дисперсионного анализа

Необходимо поставить галочку в колонке **Попарные сравнения средних** (рис. 1.32).

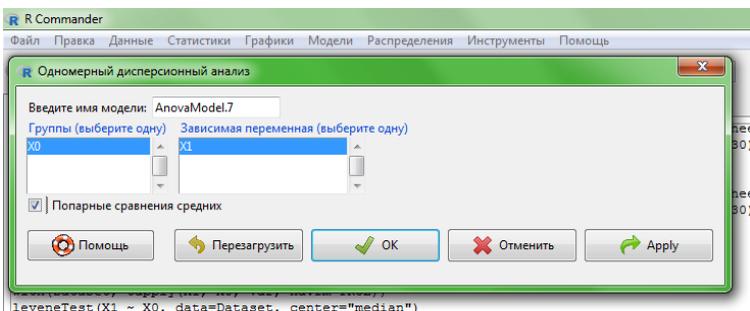


Рис. 1.32. Диалоговое окно, показывающее процесс выбора функции попарных сравнений средних

Получаем результат (рис. 1.33).

```
Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = X1 ~ X0, data = Dataset)

Linear Hypotheses:
      Estimate Std. Error t value Pr(>|t|)
Опыт 1 - Контроль == 0   37.77    32.57   1.160   0.481
Опыт 2 - Контроль == 0   19.18    35.03   0.547   0.848
Опыт 2 - Опыт 1 == 0    -18.60    35.98  -0.517   0.863
(Adjusted p values reported -- single-step method)
```

Рис. 1.33. Диалоговое окно, показывающее результаты расчета одномерного дисперсионного анализа

Как видно из представленных данных, значение  $Pr(t)$  в опытных группах выше принятого уровня значимости (0,05) по отношению к контрольной группе и в сравнении между собой, т. е. существенных различий между исследуемыми выборками нет, и различия по массе являются недостоверными.

Результаты парных сравнений групповых средних также изображаются в виде рисунка (рис. 1.34).

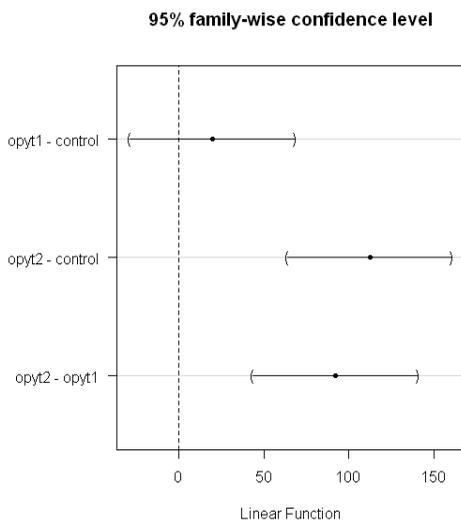


Рис. 1.34. Результаты парных сравнений групповых средних

На представленном рисунке приведены разности между групповыми средними (differences in mean levels of water) и их доверительные интервалы, рассчитанные с учетом контроля над групповой вероятностью ошибки (95 % family-wise confidence level). Критерий Тьюки имеет те же условия применимости, что и собственно дисперсионный анализ, т. е. нормальность распределения данных и (особенно важно!) однородность групповых дисперсий. Устойчивость к отклонению от этих условий, равно как и статистическая мощность критерия Тьюки, возрастают при одинаковом числе наблюдений во всех сравниваемых группах.

#### 1.4. Непараметрические критерии

Как было указано выше, если наши данные не отвечают требованиям нормальности распределения и однородности групповых дисперсий, то необходимо использовать непараметрические критерии:

1. U-критерий Манна – Уитни (Тест Уилкоксона) – используется для сравнения двух исследуемых групп между собой. *Здесь необходимо сразу пояснить, что создатели системы R под названием «критерий Уилкоксона» (или «тест Уилкоксона») объединяют как метод, предложенный собственно Ф. Уилкоксоном (F. Wilcoxon) в 1945 г., так и опубликованный несколько позднее (1947 г.) метод Манна – Уитни.*

Для вызова данного теста необходимо последовательно нажать **Статистики** → **Непараметрические тесты** → **Двухвыборочный тест Уилкоксона**.

Различия будут достоверны, если значение p-value < 0,05.

2. Критерий парных множественных сравнений средних рангов (тест Ньюмена) – используется для сравнения трех и более исследуемых групп.

Данный критерий в пакете R Commander не установлен.

Для того чтобы запустить критерий парных множественных сравнений, необходимо, чтобы данные были уже загружены (с помощью пакета R Commander, как это было описано выше).

При этом следует обратить внимание на то, чтобы группирующая колонка и колонка с данными имели подписи в верхней строке (например, **Группа** и **Масса**), как это указано на рис. 1.35.

	А	В
1	Группа	Масса
2	Контроль	260
3	Контроль	300
4	Контроль	200
5	Контроль	230
6	Контроль	290
7	Контроль	350
8	Контроль	210
9	Контроль	320
10	Контроль	210
11	Контроль	330
12	Контроль	320
13	Контроль	200
14	Контроль	290
15	Контроль	390
16	Контроль	210
17	Контроль	220
18	Контроль	520
19	Контроль	490
20	Контроль	460
21	Контроль	105
22	Контроль	110
23	Контроль	115
24	Контроль	120
25	Контроль	115
26	Контроль	110
27	Опыт 1	230
28	Опыт 1	300
29	Опыт 1	330
30	Опыт 1	330
31	Опыт 1	250
32	Опыт 1	370
33	Опыт 1	300
34	Опыт 1	380
35	Опыт 1	270
36	Опыт 1	269
37	Опыт 1	790
38	Опыт 1	290
39	Опыт 1	180
40	Опыт 1	290
41	Опыт 1	330

Рис. 1.35. Диалоговое окно, показывающее принцип построения данных для определения критериев достоверности при помощи теста Ньюмена

Затем необходимо зайти в основное рабочее окно системы R (рис. 1.36).

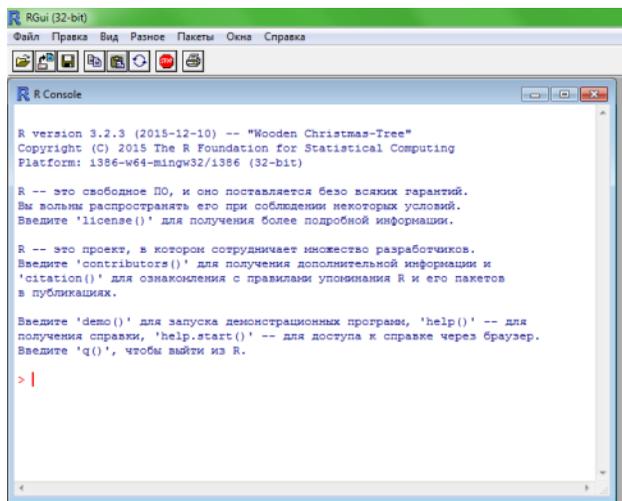


Рис. 1.36. Диалоговое окно консоли R

Для работы с критерием парных множественных сравнений средних рангов необходимо установить и запустить пакет `PMCMR` по аналогии с установкой и запуском пакета `R Commander`, описанными выше.

Затем в строку основного рабочего окна консоли R необходимо ввести скрипт

```
posthoc.kruskal.nemenyi.test(Масса ~ Группа, data=Dataset).
```

Получим следующие результаты (рис. 1.37).

```
> local({pkg <- select.list(sort(.packages(all.available = TRUE)), graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
> posthoc.kruskal.nemenyi.test(Масса ~ Группа, data=Dataset)
Предупреждение в posthoc.kruskal.nemenyi.test.default(c(260, 300, 200, 230, 290),
Ties are present, p-values are not corrected.

Pairwise comparisons using Tukey and Kramer (Nemenyi) test
with Tukey-Dist approximation for independent samples

data:  Масса by Группа

      Контроль Опыт 1 Опыт 2
Опыт 1 0.57    -      -
Опыт 2 0.86    0.98   -
Опыт 3 0.68    1.00   0.99
```

Рис. 1.37. Диалоговое окно, показывающее результаты обработки данных при помощи теста Ньюмена

Как видно из полученных данных, при сравнении контрольной группы с опытной № 1  $p\text{-value} = 0,57$ ; с опытной группой № 2 –  $p\text{-value} = 0,86$ ; с опытной группой № 3 –  $p\text{-value} = 0,68$ . При сравнении опытных групп между собой получили следующие значения: опытная группа № 1 с опытной группой № 2 –  $p\text{-value} = 0,98$ ; опытная группа № 1 с опытной группой № 3 –  $p\text{-value} = 1,00$ ; опытная группа № 2 с опытной группой № 3 –  $p\text{-value} = 0,99$ . Согласно представленным данным, вероятность ошибки  $p$  ( $p\text{-value}$ ) во всех вариантах сравнения значительно больше принятого уровня значимости (0,05), т. е. существенных различий между исследуемыми выборками нет. Следовательно, заключить, что фактор X оказал достоверный стимулирующий эффект, нельзя.

### Табличное исполнение полученных результатов

В результате проведенных статистических исследований можно составить табл. 1.2–1.6 (составлены произвольно из имитированных данных).

Таблица 1.2. Пример для двух исследуемых групп (данные не прошли проверку на нормальность распределения)

Группа	Mean $\pm$ SE, мг	SD	CV, %	n	Тест Шапиро – Уилка	U-критерий Манна – Уитни
Контрольная	259,0 $\pm$ 24,1	120,6	0,5	25	$p < 0,05$	–
Опытная № 1	296,8 $\pm$ 25,9	121,6	0,4	22		$p > 0,05$

Таблица 1.3. Пример для двух исследуемых групп (данные прошли проверку на нормальность распределения, но не прошли проверку на однородность дисперсий)

Группа	Mean $\pm$ SE, мг	SD	CV, %	n	Тест Шапиро – Уилка	F-тест	U-критерий Манна – Уитни
Контрольная	259,0 $\pm$ 24,1	120,6	0,5	25	$p > 0,05$	$p < 0,05$	–
Опытная № 1	296,8 $\pm$ 25,9	121,6	0,4	22			$p > 0,05$

Таблица 1.4. Пример для двух исследуемых групп (данные прошли проверку на нормальность распределения и на однородность дисперсий)

Группа	Mean $\pm$ SE, мг	SD	CV, %	n	Тест Шапиро – Уилка	F-тест	Тест Стьюдента
Контрольная	259,0 $\pm$ 24,1	120,6	0,5	25	$p > 0,05$	$p > 0,05$	–
Опытная № 1	296,8 $\pm$ 25,9	121,6	0,4	22			$p > 0,05$

Таблица 1.5. Пример для трех и более исследуемых групп  
(данные прошли проверку на нормальность распределения,  
но не прошли проверку на однородность дисперсий)

Группа	Mean $\pm$ SE, мг	SD	CV, %	n	Тест Шапиро – Уилка	Тест Ливина	Тест Ньюмена
Контрольная	259,0 $\pm$ 24,1	120,6	0,5	25	p > 0,05	p < 0,05	–
Опытная № 1	296,8 $\pm$ 25,9	121,6	0,4	22			p > 0,05
Опытная № 2	278,2 $\pm$ 18,9	78,2	0,3	17			p > 0,05

Таблица 1.6. Пример для трех и более исследуемых групп  
(данные прошли проверку на нормальность распределения  
и на однородность дисперсий)

Группа	Mean $\pm$ SE, мг	SD	CV, %	n	Тест Шапиро – Уилка	Тест Ливина	Тест Тьюки
Контрольная	259,0 $\pm$ 24,1	120,6	0,5	25	p > 0,05	p > 0,05	–
Опытная № 1	296,8 $\pm$ 25,9	121,6	0,4	22			p > 0,05
Опытная № 2	278,2 $\pm$ 18,9	78,2	0,3	17			p > 0,05

### 1.5. Оценка полученных результатов на соответствие нормативным значениям

Для проверки соответствия полученных данных какой-то определенной норме (нормативному значению) обратимся к одновыборочному t-тесту.

Предположим, у нас имеются данные гидрохимических исследований концентрации нитратов в воде. Зададимся вопросом: «Отличается ли выборочное среднее значение от установленной нормы в УЗВ (100 мг/л)?».

Вначале сформируем и загрузим данные концентрации нитратов в воде на различных участках технологического процесса (рис. 1.38).

	X0
1	100
2	120
3	150
4	120
5	120
6	100
7	90
8	80
9	70
10	60
11	40
12	30

Рис. 1.38. Диалоговое окно со значениями концентрации нитратов в воде

Затем в меню выберем последовательно **Статистики** → **Среднее** → **Одновыборочный t-тест** (рис. 1.39).

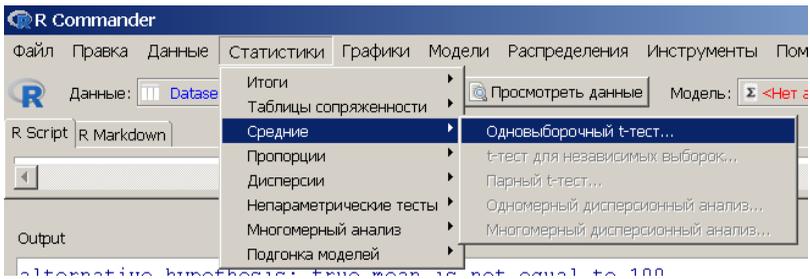


Рис. 1.39. Диалоговое окно, показывающее принцип выбора одновыборочного t-теста

В ячейке **Нулевая гипотеза** устанавливаем нормативное значение 100 (мг/л) (рис. 1.40).

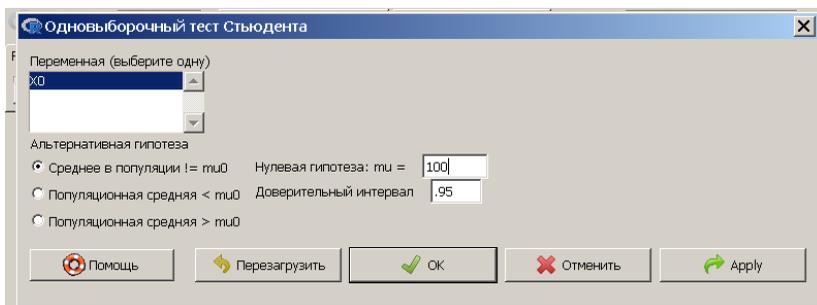


Рис. 1.40. Диалоговое окно, показывающее принцип установления нормативного значения

Нажимаем **OK** и получаем результат статистического анализа (рис. 1.41).

```

One Sample t-test

data: X0
t = -0.97101, df = 11, p-value = 0.3524
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 67.33299 112.66701
sample estimates:
mean of x
 90

```

Рис. 1.41. Диалоговое окно, показывающее результаты применения одновыборочного t-теста при оценке полученных результатов на соответствие нормативным значениям

Как видно из полученных данных,  $p\text{-value} > 0,05$ . Следовательно, по правилу мы не можем отклонить проверяемую нулевую гипотезу о равенстве выборочного среднего значения нормативу и принять альтернативную гипотезу (alternative hypothesis: true mean is not equal to 100). То есть на вопрос, поставленный выше: «Отличается ли выборочное среднее значение от установленной нормы в УЗВ (100 мг/л)?» – мы с уверенностью можем ответить: «Нет, не отличается».

## 1.6. Базовые графические возможности R

Графическое представление данных играет очень важную роль в статистике. Читатель, интересующийся всем спектром графических возможностей R, может посетить сайт R Graph Gallery, на котором представлены не только примеры всевозможных графиков, но и исходный R-код, использованный для их построения.

Как правило, создание графика начинается с функции высокого уровня, которая определяет его общую структуру: размерность (1D, 2D, 3D), масштабы осей, названия и др. Наиболее часто используемые графические функции высокого уровня – `plot()`, `hist()`, `boxplot()`, `scatterplot()` и `pairs()`.

### Функция `plot()` и ее параметры

**Функция `plot()`** – главная «рабочая лошадка», используемая для построения графиков в R. Поведение этой функции высокого уровня определяется классом объектов, указываемых в качестве ее аргументов. Соответственно, с помощью `plot()` можно создать очень большой набор разнотипных графиков.

В качестве примера используем данные по изменению процента подвижных сперматозоидов осетровых рыб в зависимости от времени подвижности (табл. 1.7).

Таблица 1.7. **Изменение процента подвижности сперматозоидов осетровых рыб в зависимости от времени подвижности**

	time	conc
1	2	3
1	1.00	99.0
2	30.00	90.0
3	60.00	80.0
4	90.00	70.0
5	120.00	60.0
6	150.00	50.0
7	180.00	30.0
8	210.00	20.0
9	1.00	99.0
10	30.00	87.0
11	60.00	82.0

Окончание табл. 1.7

1	2	3
12	90.00	71.0
13	120.00	62.0
14	150.00	53.0
15	180.00	33.0
16	210.00	25.0
17	1.00	98.0
18	30.00	88.0
19	60.00	79.0
20	90.00	70.0
21	120.00	63.0
22	150.00	51.0
23	180.00	35.0
24	210.00	21.0
25	1.00	97.0
26	30.00	86.0
27	60.00	76.0
28	90.00	73.0
29	120.00	62.0
30	150.00	50.0
31	180.00	32.0
32	210.00	20.0

После загрузки данных (например, через пакет R Commander, как это было указано выше) в строку основного рабочего окна консоли R необходимо ввести скрипт

```
data(Dataset)
attach(Dataset)
plot(time, X.conc)
```

или

```
plot(time, conc).
```

Получаем график зависимости процента подвижных сперматозоидов осетровых рыб от времени подвижности, представленный на рис. 1.42.

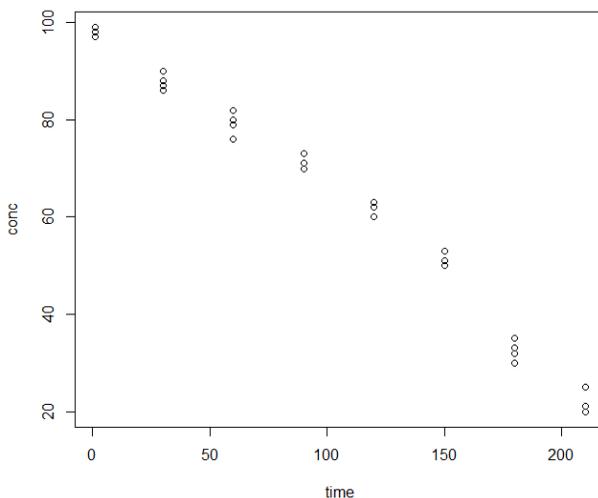


Рис. 1.42. График зависимости процента подвижных сперматозоидов осетровых рыб от времени подвижности

Предположим, что перед нами стоит задача отобразить на графике не все исходные данные, а только средние значения для каждой временной точки. Рассчитать средние значения (или любые другие количественные величины) для отдельных групп данных позволяет функция `tapply()`:

`(means <- tapply(X.conc, time, mean))` (рис. 1.43).

```

1    30    60    90    120    150    180    210
98.25 87.75 79.25 71.00 61.75 51.00 32.50 21.50

```

Рис. 1.43. Диалоговое окно со средними значениями для каждой временной точки

Обратите внимание на то, что при создании вектора `means` функция `tapply()` автоматически присвоила каждому из рассчитанных средних величин имя, соответствующее времени подвижности сперматозоидов.

Мы можем воспользоваться этим обстоятельством при построении графика и создать числовой вектор со значениями времени подвижности сперматозоидов:

```
indo.times <- as.numeric(names(means)).
```

Далее строим график типа «точки с линиями» (рис. 1.44):

```
plot(indo.times, means, type="b").
```

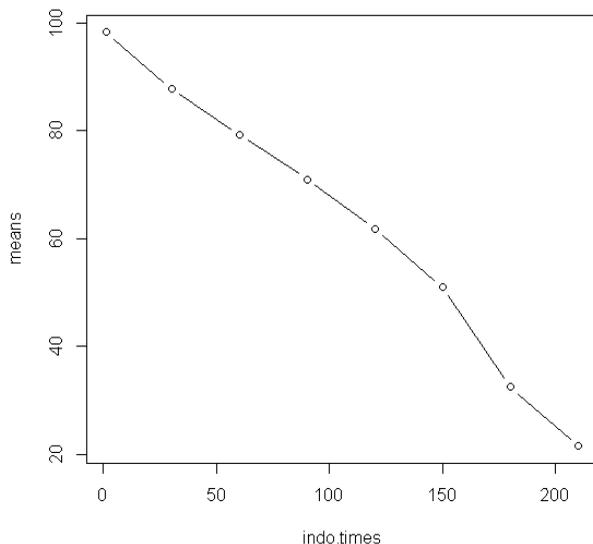


Рис. 1.44. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности

### ***Управляющие параметры функции plot().***

Функция `plot()` имеет большое количество управляющих параметров, которые позволяют осуществить тонкую настройку внешнего вида графика. Ниже рассмотрены лишь некоторые из них.

Параметры `xlab` и `ylab` служат для изменения названий осей X и Y соответственно:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов,  
с", ylab = "Процент подвижных сперматозоидов", type = "b") (рис. 1.45).
```

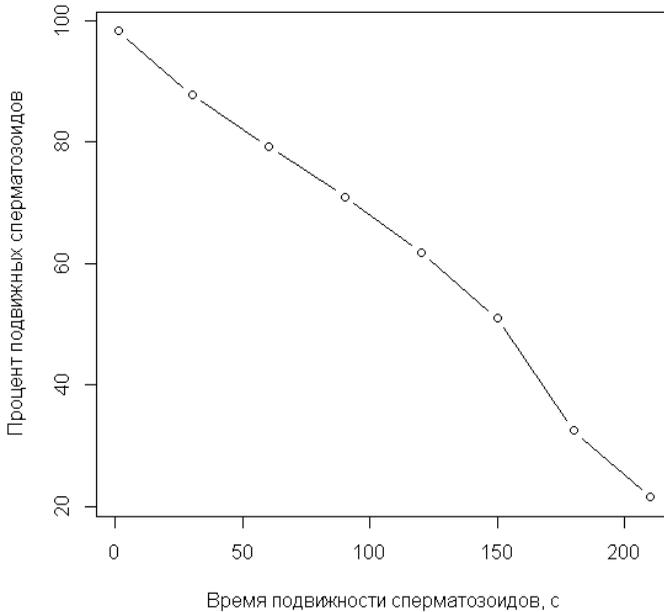


Рис. 1.45. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с подписанными осями)

Параметр `type` позволяет изменять внешний вид точек на графике. Он принимает одно из следующих значений:

- "p" – точки (points; используется по умолчанию);
- "l" – линии (lines);
- "b" – изображаются и точки, и линии (both points and lines);
- "o" – точки изображаются поверх линий (points over lines);
- "h" – гистограмма (histogram);
- "s" – ступенчатая кривая (steps);
- "n" – данные не отображаются (no points).

Например:

`plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="s")` (рис. 1.46).

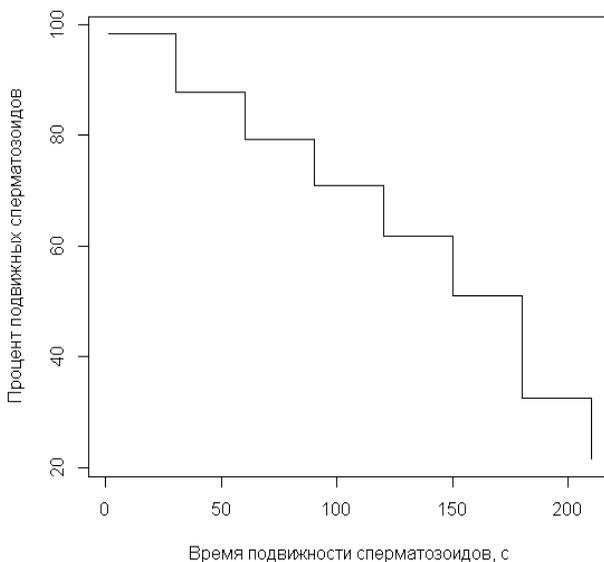


Рис. 1.46. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (ступенчатая кривая)

Параметры `xlim` и `ylim` контролируют размах значений на каждой из осей графика. По умолчанию они оба принимают значение `NULL` – в этом случае размах выбирается программой автоматически. Для отмены автоматических настроек соответствующему параметру необходимо присвоить значение в виде числового вектора, содержащего минимальное и максимальное значения, которые должны отображаться на оси.

Например:

`plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", xlim=c(0, 150), ylim=c(0, 90))` (рис. 1.47).

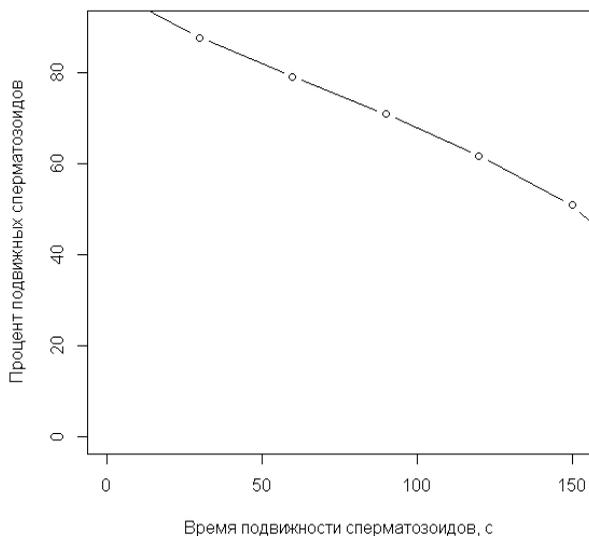


Рис. 1.47. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с изменением параметров xlim и ylim)

Аргумент main служит для создания заголовка графика. По умолчанию название размещается в верхней части рисунка.

plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", main = "Динамика изменения процента подвижных сперматозоидов") (рис. 1.48).

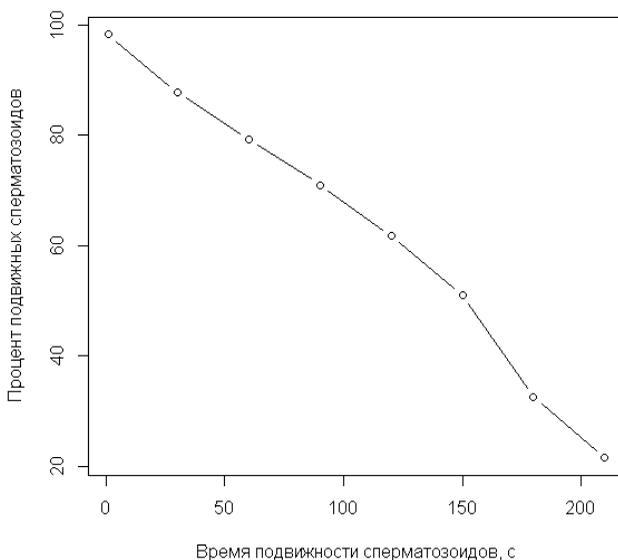


Рис. 1.48. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с заголовком)

Далее будут рассмотрены графические параметры, контролирующие внешний вид графиков, например: тип, размер и цвет символов и линий, тип и размер шрифта в названиях графика и его осей, использование математических символов в названиях, размещение легенды и т. п. Они применяются в качестве аргументов при вызове не только `plot()`, но и многих других функций.

Изменить тип символов, используемых для отображения наблюдений, позволяет аргумент `pch` (*plotting character* – символ изображения).

Таблица 25 стандартных маркеров и соответствующих им числовых кодов представлена на рис. 1.49.



Рис. 1.49. Стандартные маркеры с числовыми кодами

Например:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов,
с", ylab = "Процент подвижных сперматозоидов", type="b", main =
"Динамика изменения процента подвижных сперматозоидов",
pch = 25) (рис. 1.50).
```

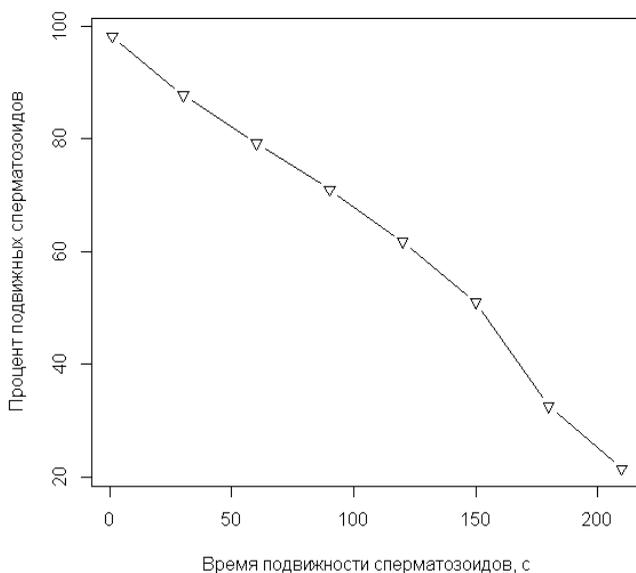


Рис. 1.50. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененным маркером)

Набор стандартных маркеров может быть значительно расширен в случае, когда аргумент `pch` используется в комбинации с другим аргументом – `font`, задающим шрифт символов. Параметр `pch` может при этом принимать любое целое число от 1 до 128 и от 160 до 254. Например, при `font = 5` маркеру в виде сердечка соответствует код 169:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", main = "Динамика изменения процента подвижных сперматозоидов", pch = 169, font = 5) (рис. 1.51).
```

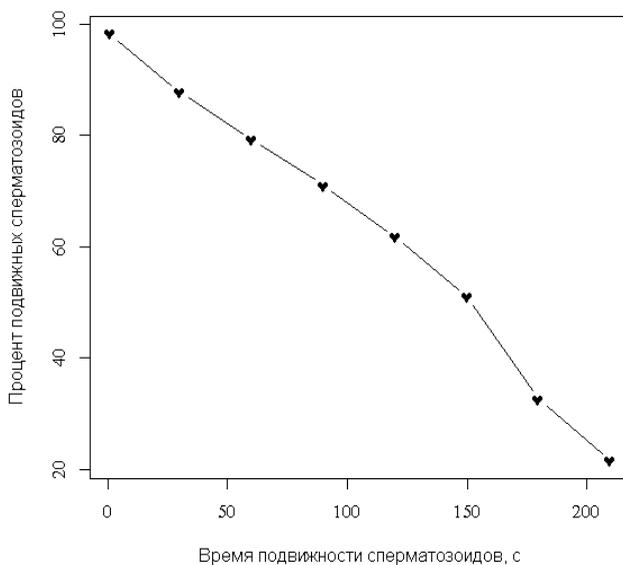


Рис. 1.51. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененным маркером)

Размер маркеров задается при помощи аргумента `sex` (`character extension` – размер символа), который по умолчанию равен 1. Уменьшение или увеличение данного параметра приводит к соответствующим пропорциональным изменениям размера символов.

Например:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", main = "Динамика изменения процента подвижных сперматозоидов", pch = 5, sex=5) (рис. 1.52).
```

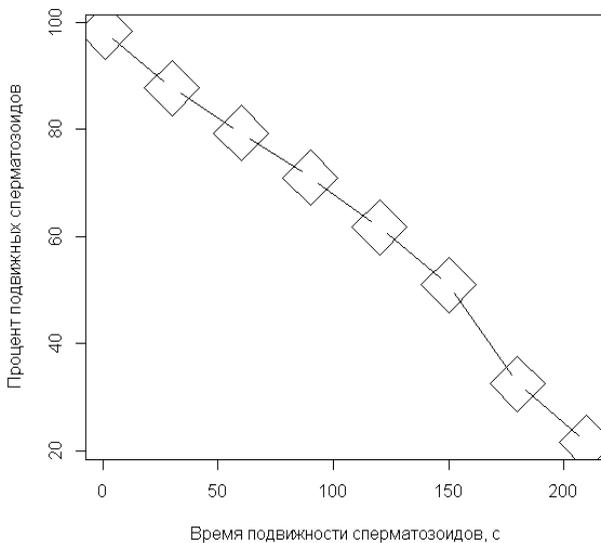


Рис. 1.52. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененным размером маркера)

При необходимости можно также изменить ширину линии обводки символа. Для этого служит параметр `lwd` (line width – ширина линии).

Например:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", main = "Динамика изменения процента подвижных сперматозоидов", pch = 5, sex=2, lwd=10) (рис. 1.53).
```

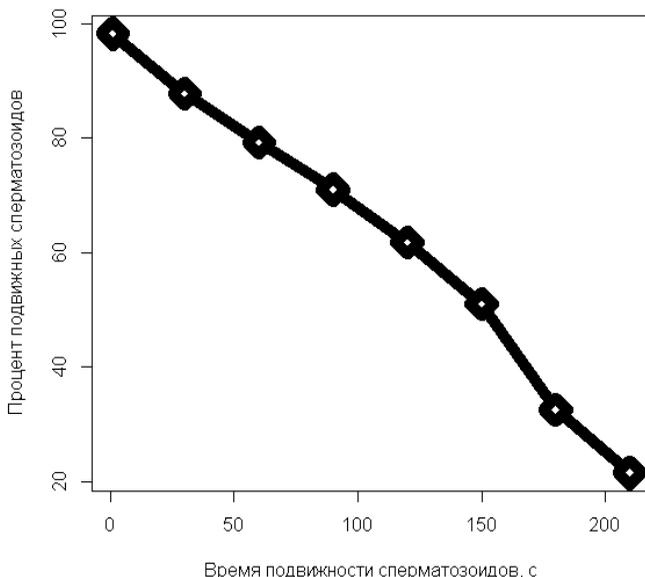


Рис. 1.53. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененной шириной линии)

Цвет любого графического объекта может быть задан несколькими способами:

- по названию цвета: например, `col = "red"` (красный), `col = "green"` (зеленый) или `col = "black"` (черный). Всего в R имеется 675 стандартных цветов. Их названия доступны по команде `colors()`;
- путем непосредственного указания красного, зеленого и синего компонентов RGB спектра: например, `"#RRGGBB"` (подробнее см. [ru.wikipedia.org](http://ru.wikipedia.org) и [stm.dp.ua](http://stm.dp.ua));
- по численному коду: например, `col = 2` (красный), `col = 3` (зеленый) или `col = 1` (черный).

Например:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", main = "Динамика изменения процента подвижных сперматозоидов", pch = 5, sex=2, lwd=3, col = 3) (рис. 1.54).
```

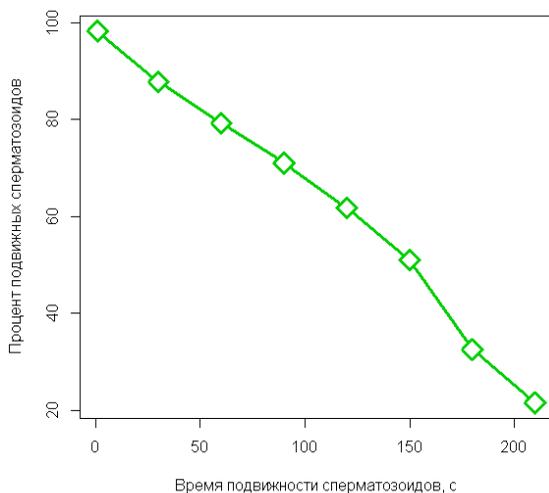


Рис. 1.54. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененным цветом линии)

Цвет маркеров можно подобрать следующим образом (рис. 1.55).

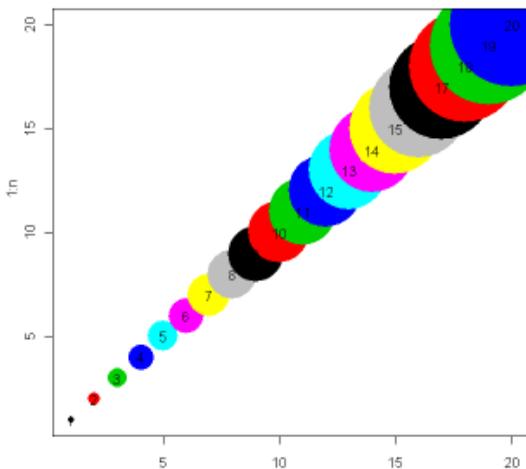


Рис. 1.55. Подбор цвета маркеров

Имеются также отдельные параметры для настройки цвета других элементов графика (например, заголовка – `col.main`, названий осей – `col.lab`, меток осей – `col.axes` и др.).

Например:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов,  
с", ylab = "Процент подвижных сперматозоидов", type="b", main =  
"Динамика изменения процента подвижных сперматозоидов", pch = 5,  
sex=2, lwd=3, col = 3, col.main=3);
```

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов,  
с", ylab = "Процент подвижных сперматозоидов", type="b", main =  
"Динамика изменения процента подвижных сперматозоидов", pch = 5,  
sex=2, lwd=3, col = 3, col.main=3, col.lab=3).
```

Тип линии настраивается при помощи аргумента `lty` (от *line* – линия и *type* – тип) функции `plot()`. Существует шесть предустановленных типов линий, которые задаются числами от 1 до 6 соответственно.

При необходимости можно создать пользовательские типы линий. В таких случаях в качестве значения аргумента `lty` выступает текстовая последовательность из четырех цифр. Эти цифры (от 1 до 9) определяют размер четырех элементов, составляющих повторяющийся паттерн: "штрих – пробел – штрих – пробел". Например, при `lty = "4241"` линия будет состоять из повторяющегося паттерна, в котором имеется штрих длиной 4 ед., пробел длиной 2 ед., опять штрих длиной 4 ед. и пробел в 1 ед.

Например:

```
plot(indo.times, means, xlab = "Время подвижности сперматозоидов,  
с", ylab = "Процент подвижных сперматозоидов", type="b", main =  
"Динамика изменения процента подвижных сперматозоидов", pch = 5,  
sex=2, lwd=3, col = 3, col.main=3, col.lab=3, lty=4) (рис. 1.56).
```

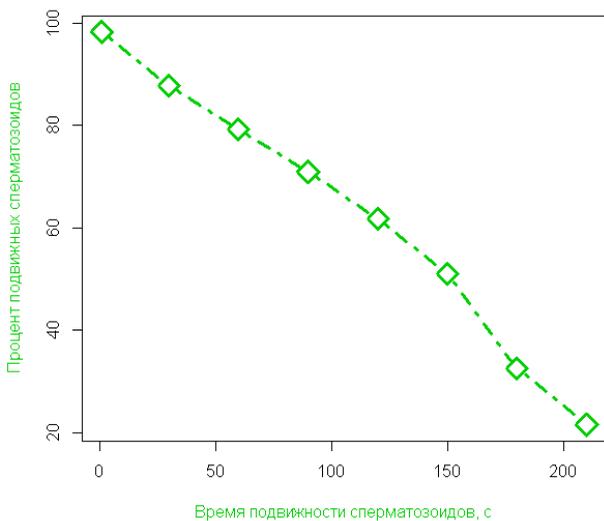


Рис. 1.56. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененными цветами отдельных элементов графика)

Примеры пользовательских типов линий приведены на рис. 1.57.

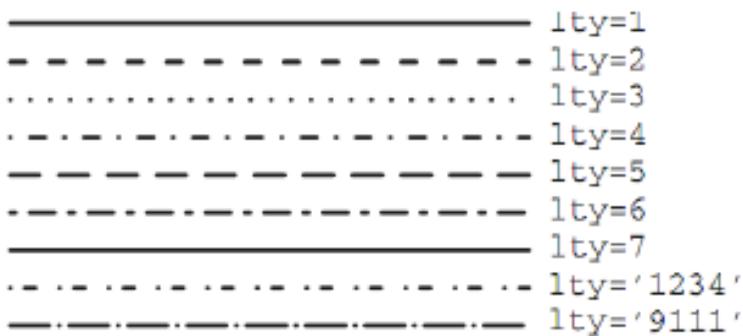


Рис. 1.57. Примеры пользовательских типов линий

Для настройки внешнего вида рамки графика служит аргумент `btu` (от `box` – коробка и `type` – тип) функции `plot()`. Этот аргумент принимает одно из следующих шести текстовых значений:

"O" "L" "7" "C" "U" "[".

Рамка будет принимать вид в соответствии с формой указанного символа (допускается также использование строчных букв `o`, `l`, `c`, и `u`). Ниже приведен пример использования различных перечисленных опций.

`plot(indo.times, means, xlab = "Время подвижности сперматозоидов, с", ylab = "Процент подвижных сперматозоидов", type="b", main = "Динамика изменения процента подвижных сперматозоидов", pch = 5, cex=2, lwd=3, col = 3, col.main=3, col.lab=3, lty=4, btu="L")` (рис. 1.58).

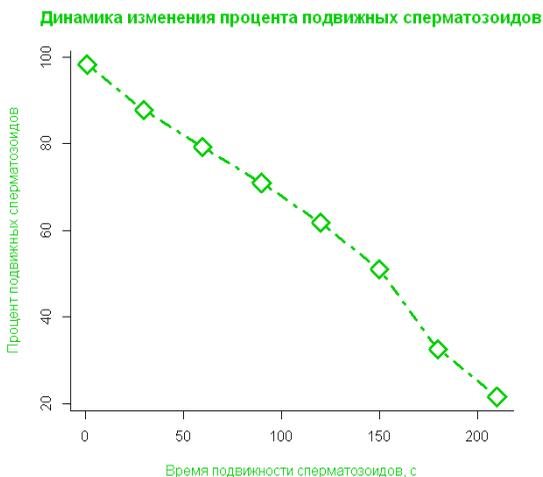


Рис. 1.58. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (с измененным внешним видом рамки)

Функция `plot()` имеет множество различных производных.

Например, проанализируем данные, полученные в ходе выполнения исследований влияния фактора  $X$  при различных времени экспо-

зиции (от 0 до 600 с) и дозировке воздействия на контрольную, опытную № 1 и опытную № 2 группы (табл. 1.8).

Таблица 1.8. Результаты исследований влияния фактора X

time	effect	
0	70	опыт1
30	75	опыт1
60	80	опыт1
90	85	опыт1
180	90	опыт1
300	95	опыт1
600	100	опыт1
0	70	опыт2
30	65	опыт2
60	60	опыт2
90	55	опыт2
180	50	опыт2
300	45	опыт2
600	40	опыт2
0	70	опыт3
30	71	опыт3
60	72	опыт3
90	73	опыт3
180	74	опыт3
300	75	опыт3
600	76	опыт3

После загрузки данных (например, через пакет R Commander, как это было указано выше) в строку основного рабочего окна консоли R необходимо ввести скрипт

```
plot_ly(Dataset, x = time, y = effect, mode = "markers", color = X,
marker=list( size=30, opacity=10)).
```

Предварительно необходимо установить и запустить пакет plotly по аналогии с установкой и запуском пакета R Commander, описанными выше.

В результате таких действий получится график, представленный на рис. 1.59.

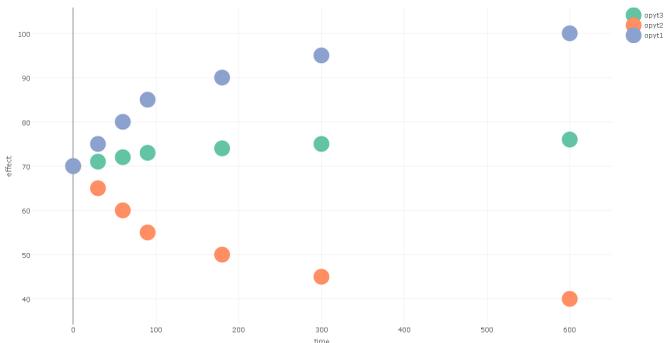


Рис. 1.59. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности с учетом исследовательских групп (функция `plot_ly`)

Также результаты полученных исследований можно представить в виде следующего графика (рис. 1.60).

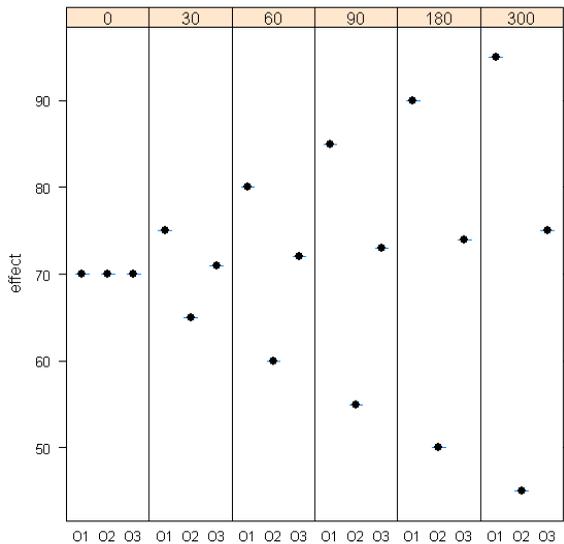


Рис. 1.60. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности с учетом исследовательских групп (функция `bwplot`)

Для этого необходимо немного поменять структуру таблицы с данными (табл. 1.9).

Таблица 1.9. Таблица с данными

groups	time	effect
o1	0	70
o1	30	75
o1	60	80
o1	90	85
o1	180	90
o1	300	95
o2	0	70
o2	30	65
o2	60	60
o2	90	55
o2	180	50
o2	300	45
o3	0	70
o3	30	71
o3	60	72
o3	90	73
o3	180	74
o3	300	75

Далее ввести следующий скрипт:

```
bwplot(effect ~ groups | factor(time), Dataset, varwidth = TRUE, layout = c(6, 1), ylab = "effect").
```

### ***Функция ядерной плотности.***

Для того чтобы оценить взаимодействие между двумя переменными, желательно построить график поверхности ядерной плотности распределения двумерной случайной величины  $z = f(x,y)$ . Это можно сделать с использованием функции `kde2d()` из популярного пакета MASS. В качестве примера используем предыдущие данные по динамике изменения процента подвижных сперматозоидов в зависимости от времени подвижности:

```

data(Dataset)
attach(Dataset)
library(MASS)
f <- kde2d(time, X.conc) или f <- kde2d(time, conc)
image(f,xlab="Время подвижности сперматозоидов",ylab="Процент
подвижности сперматозоидов") (рис. 1.61).

```

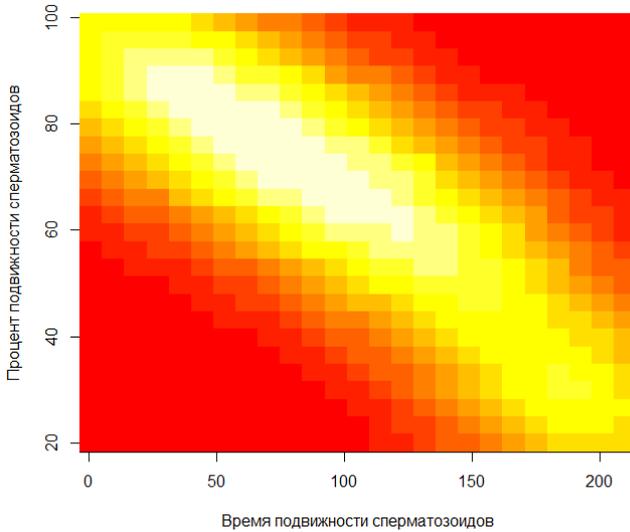


Рис. 1.61. График зависимости среднего процента подвижных сперматозоидов осетровых рыб от среднего времени подвижности (функция ядерной плотности)

**Функция *cdplot()*.**

Для визуализации связи между двумя переменными, одна из которых является количественной, а другая качественной (фактором), можно использовать диаграмму размахов (*boxplot*). При таком подходе ось ординат обычно соответствует значениям количественной переменной. Но что если зависимой является качественная переменная? Полезной здесь может оказаться еще одна базовая графическая функция R – *cdplot()*, позволяющая совмещать на одном графике плотности вероятностей для каждого уровня интересующей исследователя качественной переменной (англ. *conditional density plot*).

В качестве примера проанализируем данные исследований концентрации аспаратаминотрансферазы (AST) и аланинаминотрансферазы (ALT) сыворотки крови самок ленского осетра в зависимости от положительного (+) и отрицательного (-) ответа (response) на гормональное инъектирование (табл. 1.10).

**Таблица 1.10. Концентрации аспаратаминотрансферазы (AST) и аланинаминотрансферазы (ALT) сыворотки крови самок ленского осетра**

	AST	ALT	response
1	40	14	+
2	50	20	+
3	60	15	+
4	30	15	+
5	40	13	+
6	34	16	+
7	40	10	+
8	50	11	+
9	60	12	+
10	70	20	+
11	120	36	-
12	130	29	-
16	120	36	-
18	110	34	-
19	130	38	-
20	120	37	-
21	130	30	-
22	120	46	-
24	100	36	-
25	130	39	-
26	20	9	+
27	30	7	+
28	20	14	+
30	20	23	+
31	30	22	+
32	40	21	+
13	100	37	-
14	120	32	-
15	120	37	-
17	120	46	-
23	140	44	-
29	120	32	-

После загрузки данных (например, через пакет R Commander, как это было указано выше) в строку основного рабочего окна консоли R необходимо ввести скрипт:

```
layout(matrix(1:2, ncol = 2))
cdplot(response ~ AST, data = Dataset)
cdplot(response ~ ALT, data = Dataset) (рис. 1.62).
```

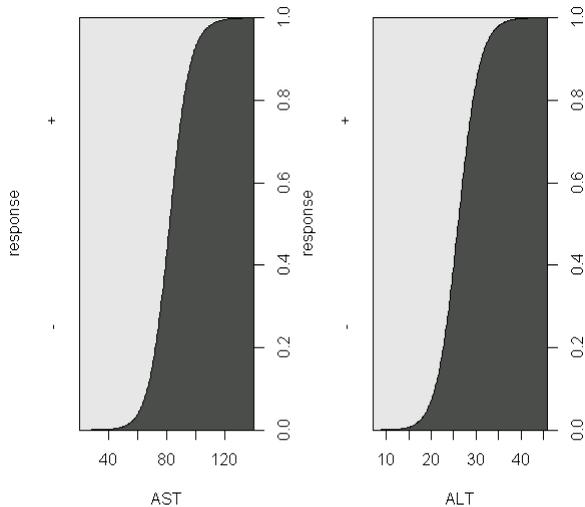


Рис. 1.62. График плотности вероятностей

При помощи аргумента `col` можно задать текстовый вектор с названиями цветов, которые будут использованы для заливки полигонов на графике:

```
layout(matrix(1:2, ncol = 2))
cdplot(response ~ AST, data = Dataset, col = c("coral", "skyblue"))
cdplot(response ~ ALT, data = Dataset, col = c("coral", "skyblue"))
(рис. 1.63).
```

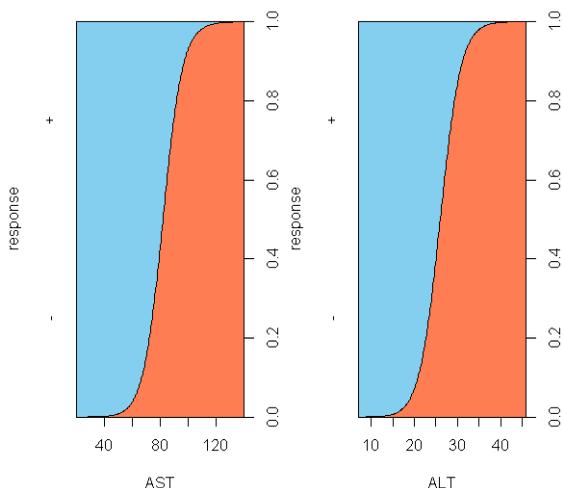


Рис. 1.63. График плотности вероятностей (цветной)

Из полученных графиков видно, что самки, положительно ответившие на гормональное инъектирование, характеризуются более высокой концентрацией аспартатаминотрансферазы и аланинаминотрансферазы в отличие от самок, отрицательно ответивших на гормональное инъектирование.

#### ***Диаграммы размахов.***

Диаграммы размахов, или «ящики с усами» (англ. box-whisker plots), получили свое название за характерный вид: точку или линию, соответствующую среднему положению совокупности данных, окружает прямоугольник («ящик»), длина которого соответствует одному из показателей разброса или точности оценки генерального параметра; дополнительно от этого прямоугольника отходят «усы», также соответствующие по длине одному из показателей разброса или точности. Графики данного типа очень популярны, поскольку позволяют дать полную статистическую характеристику анализируемой совокупности. Кроме того, диаграммы размахов можно использовать для визуальной экспресс-оценки разницы между двумя и более группами (например, между датами отбора проб, экспериментальными группами, участками пространства и т. п.).

В R для построения диаграмм размахов служит функция `boxplot()`. Здесь, в отличие от других статистических программ, при построении

диаграмм размахов используются устойчивые (робастные) оценки центральной тенденции (медиана) и разброса (интерквартильный размах – ИКР). Верхний «ус» простирается от верхней границы «ящика» до наибольшего выборочного значения, находящегося в пределах расстояния 1,5 ИКР от этой границы. Аналогично нижний «ус» простирается от нижней границы «ящика» до наименьшего выборочного значения, находящегося в пределах расстояния 1,5 ИКР от этой границы. Длину данного интервала (1,5 ИКР) можно изменить при помощи аргумента `range` функции `boxplot()`.

Строение получаемых при помощи этой функции «ящичков с усами» представлено на рис. 1.64.

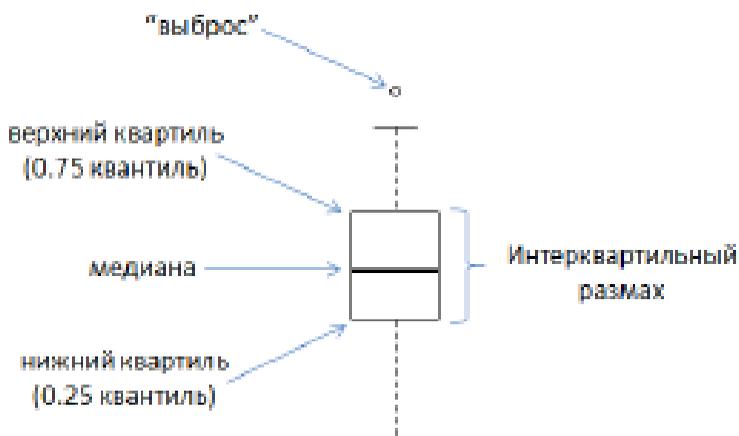


Рис. 1.64. Ключ для расшифровки значений диаграммы размахов

Наблюдения, находящиеся за пределами «усов», потенциально могут быть выбросами. Однако всегда следует внимательно относиться к такого рода нестандартным наблюдениям – они вполне могут оказаться «нормальными» для исследуемой совокупности и поэтому не должны удаляться из анализа без дополнительного расследования причин их появления.

Особенности использования функции `boxplot()` рассмотрим на примере данных, описывающих эксперимент по изучению влияния фактора  $X$  на активность концентрации аспаратаминотрансферазы в сыворотке крови стерляди (рис. 1.65).

	group	AST
1	control	40
2	control	50
3	control	60
4	control	30
5	control	40
6	control	34
7	control	40
8	control	50
9	control	60
10	control	70
11	control	120
12	control	130
13	control	120
14	control	110
15	control	130

	group	AST
16	control	120
17	control	130
18	control	120
19	control	100
20	control	130
21	opyt1	20
22	opyt1	30
23	opyt1	20
24	opyt1	20
25	opyt1	30
26	opyt1	40
27	opyt1	100
28	opyt1	120
29	opyt1	120
30	opyt1	120

	group	AST
31	opyt1	140
32	opyt1	120
33	opyt1	123
34	opyt1	100
35	opyt1	130
36	opyt1	120
37	opyt1	150
38	opyt1	190
39	opyt1	190
40	opyt1	200
41	opyt2	300
42	opyt2	250
43	opyt2	230
44	opyt2	230
45	opyt2	200

	group	AST
46	opyt2	250
47	opyt2	230
48	opyt2	200
49	opyt2	190
50	opyt2	189
51	opyt2	189
52	opyt2	203
53	opyt2	210
54	opyt2	250
55	opyt2	340
56	opyt2	200
57	opyt2	199
58	opyt2	23
59	opyt2	24
60	opyt2	20

Рис. 1.65. Активность концентрации аспаратаминотрансферазы в сыворотке крови стерляди

Для построения графика, на котором будут представлены «ящики с усами» для каждой исследуемой группы, достаточно выполнить команду

`boxplot(AST ~ group, data = Dataset)` (рис. 1.66).

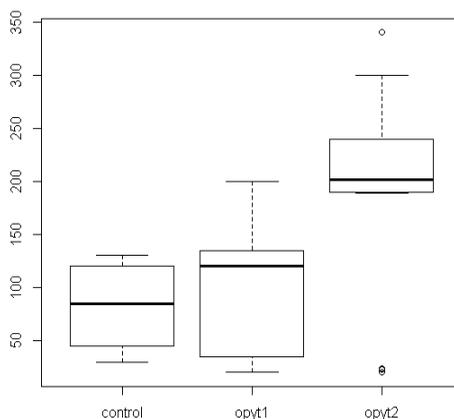


Рис. 1.66. Диаграмма размахов

Как всегда, мы можем поработать над автоматически построенным графиком и несколько улучшить его внешний вид. Например, можно добавить заголовки осей и самого рисунка (аргументы `xlab`, `ylab` и `main`), а также залить «ящики» каким-нибудь цветом (аргумент `col`):

```
boxplot(AST ~ group, data = Dataset, xlab = " группы ", ylab = " МЕ/л ",  
main = "Активность аминотрансфераз", col = "coral") (рис. 1.67).
```

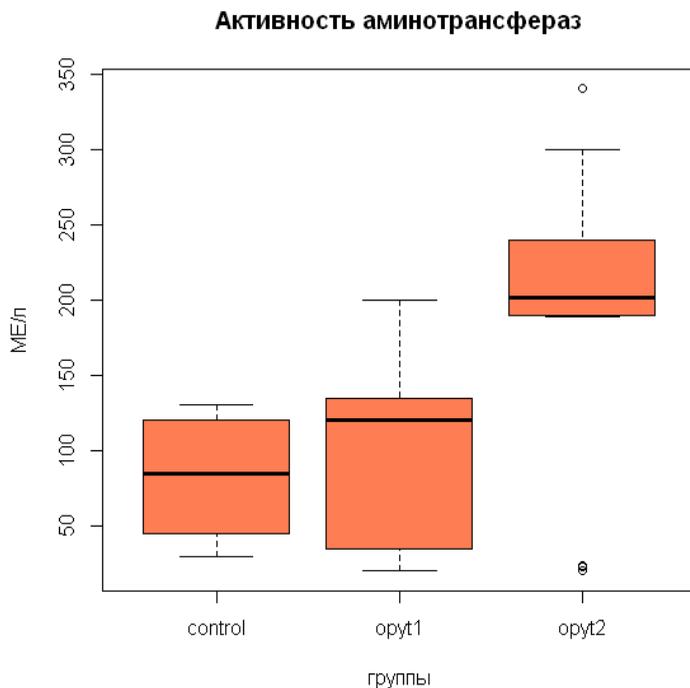


Рис. 1.67. Диаграмма размахов с заливкой цветом, подписями осей и заголовком рисунка

Для создания столбиковых (или столбчатых, реже линейчатых; англ. `bar plots` или `bar charts`) диаграмм в системе R служит функция `barplot()`.

```

library(MASS)
data(Dataset)
means = with(Dataset, tapply(AST, list(group), mean))
sds = with(Dataset, tapply(AST, list(group), sd))
barplot(means, beside = TRUE, col = topo.colors(4), legend.text =
rownames(means), xlab = "Группы", ylab = "ACT, ME/л", ylim = c(0,
400))
arrows(b, means+sds, b, means-sds, angle = 90, code = 3, length = 0.1)
(рис. 1.68).

```

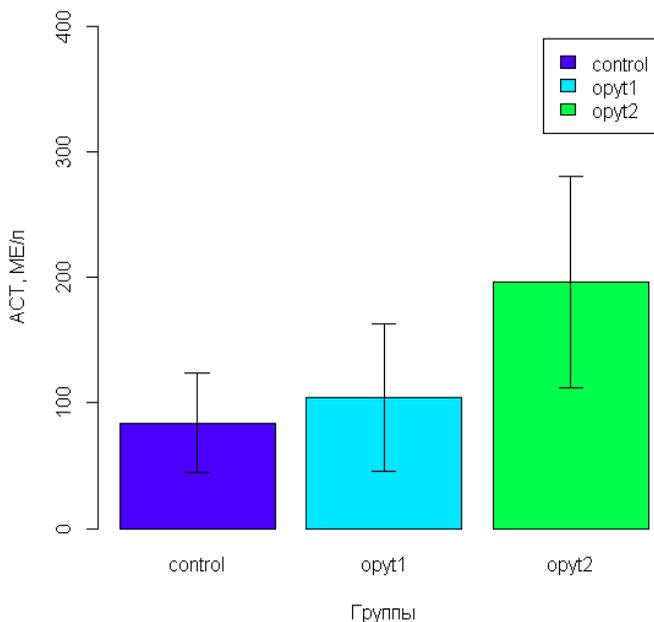


Рис. 1.68. Столбиковая диаграмма

По аналогии с ядерной функцией, описывающей плотность распределения вероятности двумерной случайной величины, при наличии двух измерений «ящик с усами» превращается в «мешок в мешке» (англ. bag plots). Как и в одномерном случае, внешний мешок ограничивается экстремальными выборочными значениями, а внутренний мешок содержит 50 % наблюдений. За пределами внешнего мешка

могут находиться выбросы. В центре диаграммы размещается область аппроксимации двумерной медианы. В качестве примера покажем диаграмму совместного распределения показателей time и conc в ранее рассмотренном эксперименте по динамике изменения процента подвижных сперматозоидов в зависимости от времени их подвижности.

```
data(Dataset)
attach(Dataset)
library(aplpack)
bagplot(time, conc, xlab="Время подвижности
сперматозоидов", ylab="Процент подвижных сперматозоидов", main="
Мешок в мешке") (рис. 1.69).
```

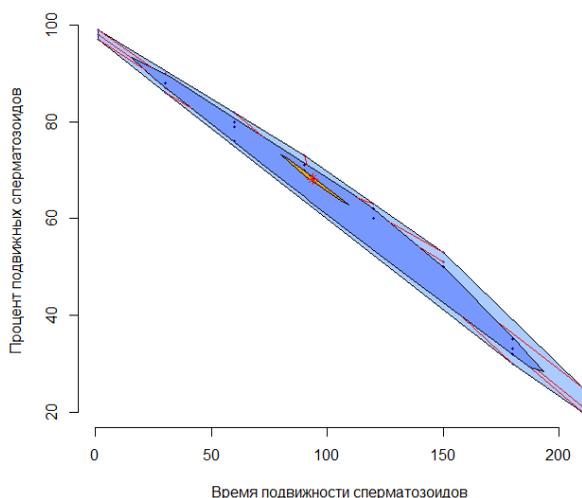


Рис. 1.69. График «Мешок в мешке»

### ***Одномерные диаграммы рассеяния.***

Одномерные диаграммы рассеяния (1-D scatter plots) представляют собой один из вариантов графического представления распределений количественных переменных. Точки, соответствующие значениям исследуемой переменной, изображаются на таких графиках вдоль единственной числовой оси. При необходимости визуализации свойств небольших выборок одномерные диаграммы рассеяния будут отличной альтернативой диаграммам размахов. В англоязычной литературе

одномерные диаграммы рассеяния называют также *strip charts* или *strip plots*, что можно перевести как «ленточные диаграммы». Это название происходит от характера расположения точек на графике – они как бы выстраиваются в «ленты». Реже такие графики называют еще точечными диаграммами Уилкоксона.

В системе R для построения одномерных диаграмм рассеяния служит функция `stripchart()`.

В качестве примера используем ранее рассмотренные данные по влиянию фактора X на активность концентрации аспаргатаминотрансферазы в сыворотке крови стерляди. Загрузим таблицу с данными в рабочую среду R и введем следующий скрипт:

```
data(Dataset)
stripchart(Dataset$AST ~ Dataset$group, ylab = "АСТ, МЕ/л", xlab = "Группы", vertical = TRUE, method = "stack").
```

Получим рис. 1.70.

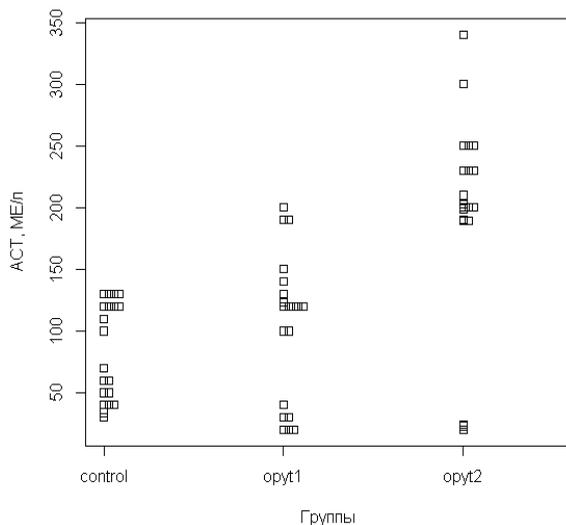


Рис. 1.70. Одномерная диаграмма рассеяния

Также при помощи функции `arrows()` можно добавить к линиям «усы» (рис. 1.71):

```

means <- tapply(Dataset$AST, Dataset$group, FUN = mean)
(SDs <- tapply(Dataset$AST, Dataset$group, FUN = sd))
arrows(c(0.9, 1.9, 2.9, 3.9, 4.9, 5.9), means + SDs, c(0.9, 1.9, 2.9, 3.9,
4.9, 5.9), means-SDs, angle = 90, code = 3, length = 0.05, lwd = 1, lend =
"square").

```

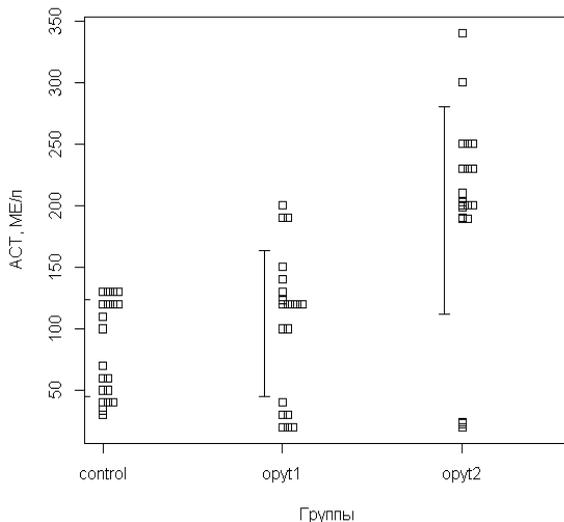


Рис. 1.71. Одномерная диаграмма рассеяния с «усами»

Новый график очень информативен – на нем изображены как все исходные наблюдения для каждой группы, так и сводная статистическая информация в виде соответствующих средних значений и стандартных отклонений.

Еще более полную картину можно получить, совместив на одном графике одномерные диаграммы рассеяния и диаграммы размахов. Сделать это можно всего несколькими строками R-кода, используя функции `boxplot()` и `stripchart()`.

```

boxplot(AST ~ group, outline = FALSE, ylab = "ACT, ME/л", xlab =
"Группы", data = Dataset)

```

```

stripchart(AST ~ group, method="stack", data = Dataset, add = TRUE,
pch = 1, col = "gray60", vertical = TRUE) (рис. 1.72).

```

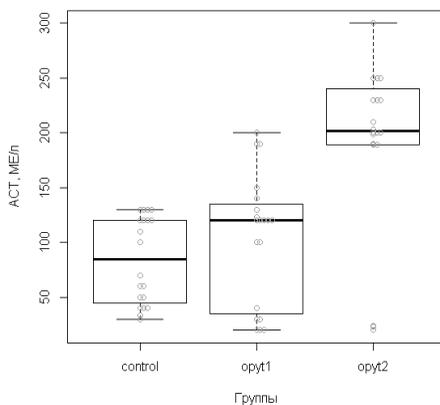


Рис. 1.72. Совмещенная одномерная диаграмма рассеяния и диаграмма размахов

При необходимости можно легко повернуть весь график на 90 °(аргумент `horizontal = TRUE` функции `boxplot()`).

`boxplot(AST ~ group, outline = FALSE, ylab = "ACT, ME/л", xlab = "Группы", data = Dataset, horizontal = TRUE)`

`stripchart(AST ~ group, method="stack", data = Dataset, add = TRUE, pch = 1, col = "gray60",)` (рис. 1.73).

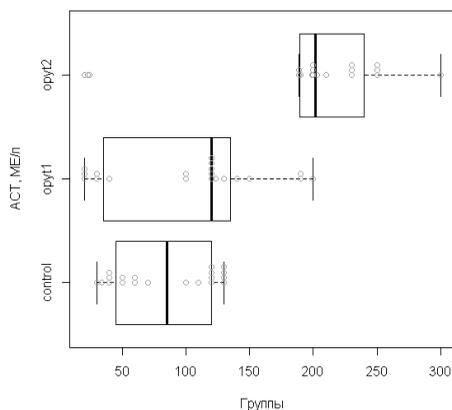


Рис. 1.73. Совмещенная одномерная диаграмма рассеяния и диаграмма размахов (горизонтальное исполнение)

### ***Вариограммы.***

Эффективным способом визуальной оценки пространственной анизотропии изучаемого явления является построение поверхности вариограммы.

Для построения поверхности вариограммы необходимо установить и запустить пакет lattice по аналогии с установкой и запуском пакета R Commander, описанными выше.

В качестве примера используем данные о влиянии оптического излучения на выживаемость личинок ленского осетра при переходе на активное питание в зависимости от частоты модуляции (0, 1, 2, 5, 10, 50 Гц – по вертикали) и времени экспозиции (0, 30, 60, 90, 180, 300 с – по горизонтали) (табл. 1.11).

Таблица 1.11. Влияние оптического излучения на выживаемость личинок ленского осетра в зависимости от частоты модуляции

	0	30	60	90	180	300
0	70	75	80	85	90	95
1	70	76	81	86	91	96
2	70	77	82	87	92	97
5	70	78	83	88	93	98
10	70	79	84	89	94	99
50	70	80	85	90	95	100

После загрузки данных введем в строку основного рабочего окна консоли R следующий скрипт:

```
data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ])).
```

Получим вариограмму, представленную на рис. 1.74.

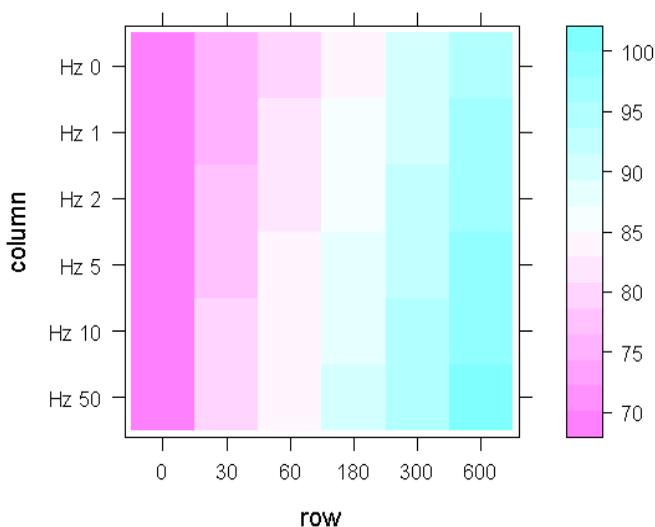


Рис. 1.74. Вариограмма влияния оптического излучения на выживаемость личинок ленского осетра в зависимости от частоты модуляции

При загрузке данных через R Commander необходимо поставить галочку во всех трех строках (рис. 1.75).

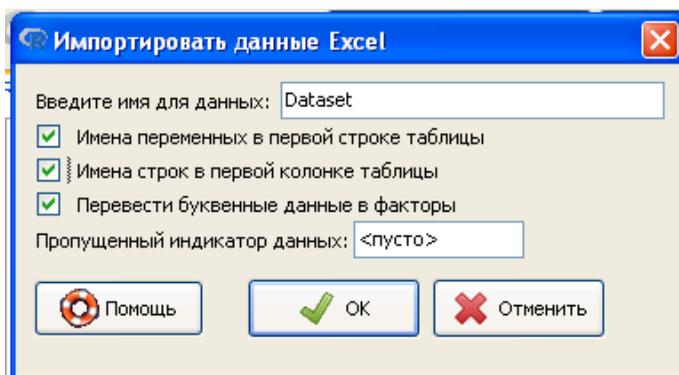


Рис. 1.75. Диалоговое окно, показывающие принцип загрузки данных для построения вариограммы

В приведенном выше скрипте функция `data=matrix` регулирует количество квадратов в вариограмме (36) и количество их по горизонтали (6) и вертикали (6). Функции `colnames(Dataset)` и `rownames(Dataset)` дают название горизонтальным и вертикальным строкам.

Если добавить в вышеприведенный скрипт функцию `col.regions` можно регулировать цветовую матрицу вариограммы.

В следующих примерах приведены скрипты и вариограммы различных цветовых матриц.

```
data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions = terrain.colors)
(рис. 1.76).
```

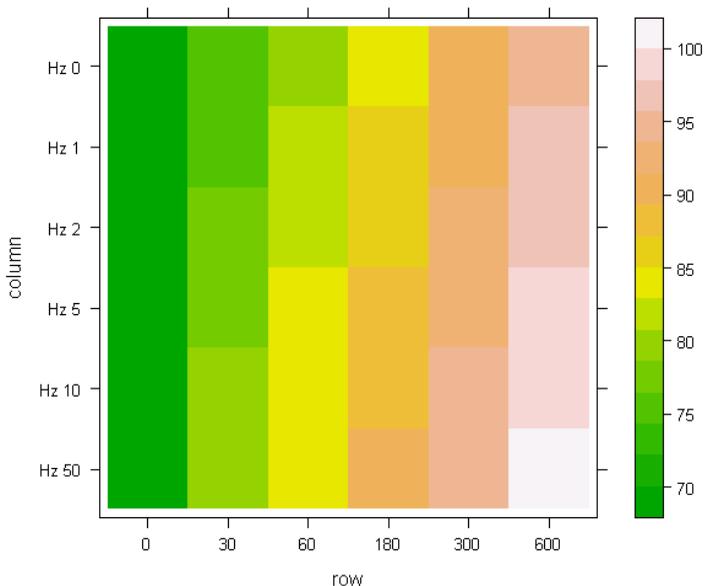


Рис. 1.76. Вариограмма (`col.regions = terrain.colors`)

```

data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions =
gray(0:100/100)) (рис. 1.77).

```

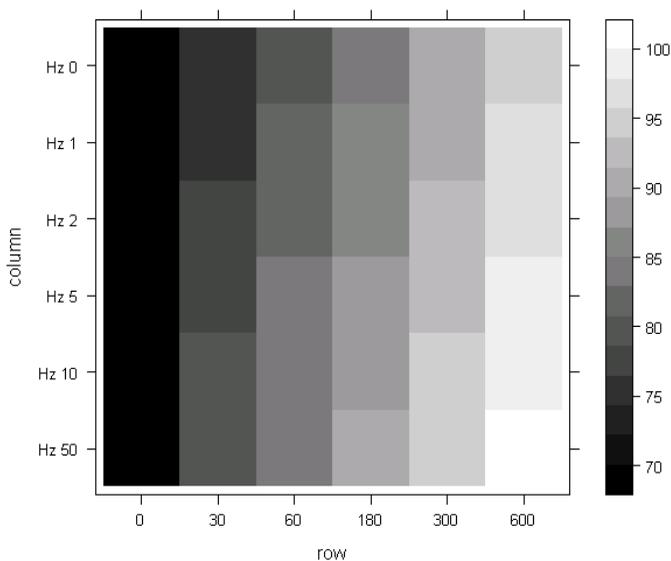


Рис. 1.77. Вариограмма (col.regions = gray(0:100/100))

```

data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions=colors)
(рис. 1.78).

```

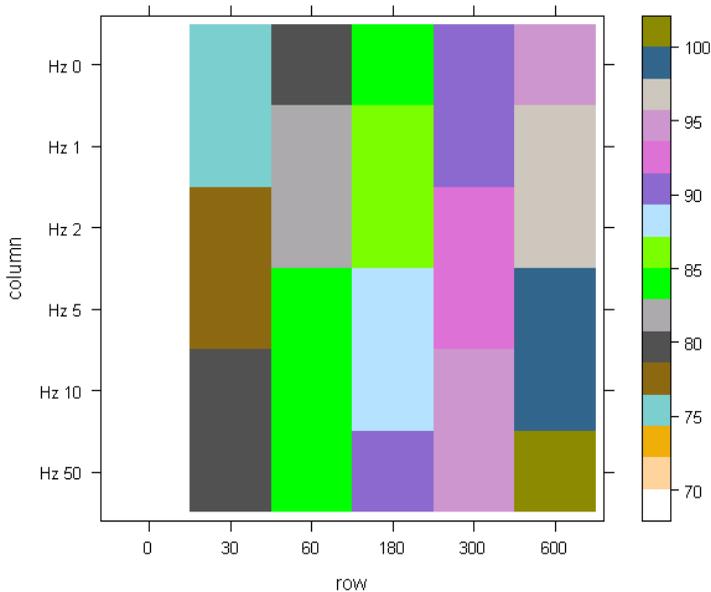


Рис. 1.78. Вариограмма (col.regions=colors)

```

data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions =
grey(100:0/100)) (рис. 1.79).

```

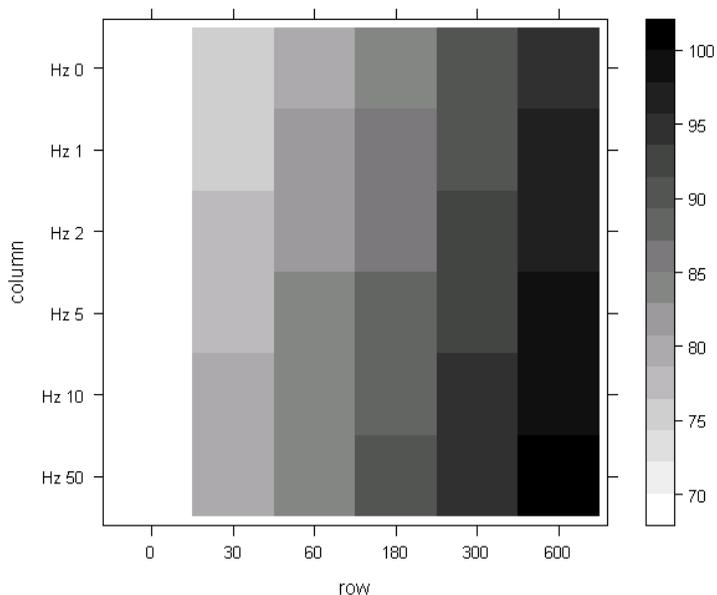


Рис. 1.79. Вариограмма (grey(100:0/100))

```

data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions = c(rgb(50:0/50,
0, 0),rgb(0,0:50/50,0) )) (рис. 1.79).

```

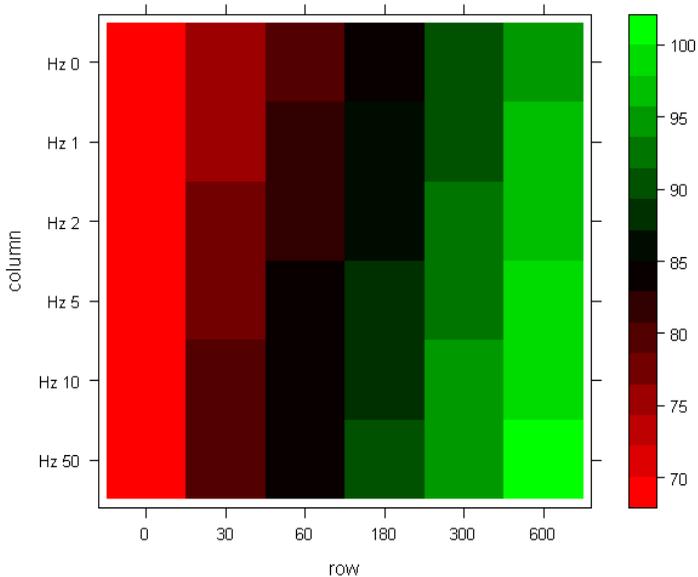


Рис. 1.80. Варнограмма (`col.regions = c(rgb(50:0/50, 0, 0),rgb(0.0:50/50,0,0))`)

```

data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions=rainbow(75))
(рис. 1.81).

```

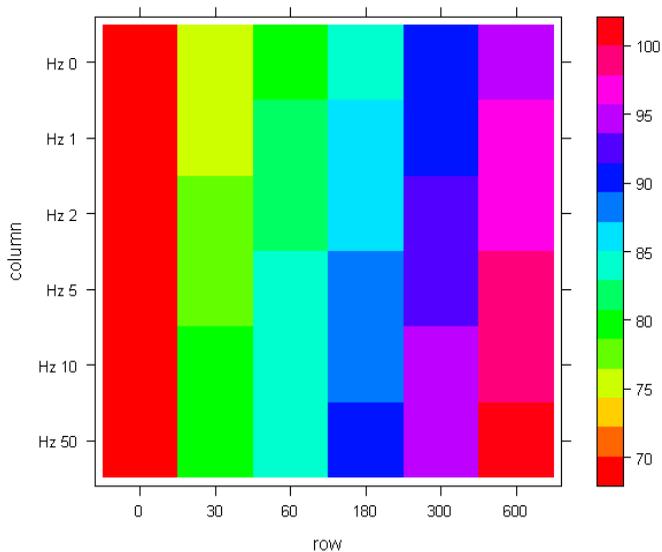


Рис. 1.81. Вариограмма (col.regions=rainbow(75))

```

data(Dataset)
library("lattice")
data=matrix(runif(36, 0, 0) , 6 , 6)
colnames(Dataset)=paste(c(0,30,60,180,300,600) , sep=" ")
rownames(Dataset)=paste( rep("Hz",6) , c(0,1,2,5,10,50) , sep=" ")
par(mar=c(3,4,2,2))
levelplot(t(Dataset[c(nrow(Dataset):1) , ]), col.regions=heat.colors(75))
(рис. 1.82).

```

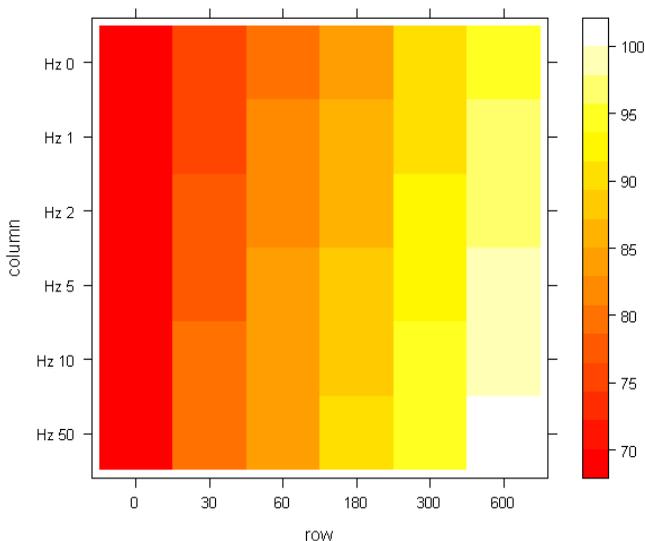


Рис. 1.82. Вариограмма (col.regions=heat.colors(75))

### ***Выявление коллинеарности.***

Когда цель анализа заключается в нахождении переменных (предикторов), связанных со значениями зависимой переменной, важным этапом разведочного анализа данных является обнаружение коллинеарности. Под коллинеарностью (англ. collinearity) понимают наличие линейной зависимости между двумя предикторами. В задачах с несколькими предикторами (например, при выполнении множественного регрессионного анализа) говорят также о мультиколлинеарности (англ. multicollinearity), т. е. о наличии линейной зависимости сразу между несколькими переменными.

Традиционным способом оценки мультиколлинеарности является анализ корреляционной матрицы. Хорошим способом повысить наглядность отображения уровня взаимных зависимостей переменных является использование раскрашенных матриц, создаваемых функцией `corrplot()` из одноименного пакета.

Рассмотрим пример с изучением оценки классности производителей карпа по комплексу признаков и по продуктивности и качеству потомства (age (возраст), weight (масса, г), length (длина, см), O (максимальный обхват), H (максимальная высота), type (соответ-

ствии желательному типу), survival larva (выживаемость личинок), survival fingerlings (выживаемость сеголеток), survival yearlings (выживаемость годовиков)) (табл. 1.12).

Таблица 1.12. Оценка классности производителей карпа по комплексу признаков и по продуктивности и качеству потомства

Age	Weight	Length	O	H	Type	Survival larva	Survival fingerlings	Survival yearlings
5.00	4500.00	55.00	43.00	22.00	1.00	400.00	80.00	83.00
5.00	3500.00	50.00	39.00	20.00	2.00	350.00	73.00	80.00
5.00	3700.00	52.00	40.00	20.00	2.00	250.00	68.00	78.00
5.00	3100.00	49.00	38.00	19.00	3.00	150.00	65.00	75.00
5.00	3000.00	45.00	35.00	17.00	3.00	450.00	81.00	84.00
5.00	2700.00	40.00	34.00	17.00	3.00	350.00	74.00	81.00
5.00	4550.00	57.00	43.00	22.00	1.00	250.00	69.00	79.00
5.00	3550.00	51.00	38.00	19.00	2.00	150.00	66.00	76.00
5.00	3750.00	51.00	41.00	21.00	2.00	400.00	79.00	85.00
5.00	3150.00	49.00	38.00	19.00	3.00	330.00	72.00	84.00
6.00	5500.00	63.00	46.00	23.00	1.00	270.00	67.00	79.00
6.00	4300.00	55.00	42.00	21.00	2.00	180.00	64.00	76.00
6.00	4600.00	57.00	44.00	22.00	2.00	433.00	88.00	80.00
6.00	3900.00	52.00	42.00	21.00	3.00	366.00	77.00	80.00
6.00	3700.00	51.00	41.00	21.00	3.00	255.00	66.00	77.00
6.00	3300.00	49.00	39.00	19.00	3.00	160.00	66.00	75.00
6.00	6350.00	63.00	46.00	23.00	1.00	440.00	81.00	89.00
6.00	5150.00	59.00	44.00	22.00	2.00	378.00	75.00	87.00
6.00	5350.00	58.00	45.00	23.00	2.00	222.00	65.00	84.00
6.00	4550.00	55.00	42.00	21.00	2.00	111.00	56.00	85.00

Для построения раскрашенной матрицы необходимо установить и запустить пакет `corrplot` по аналогии с установкой и запуском пакета `R Commander`, описанными выше.

После загрузки данных введем в строку основного рабочего окна консоли `R` следующий скрипт:

```
M <- cor(Dataset)
library(corrplot)
col4 <- colorRampPalette (c("#7F0000","red", "#FF7F00","yellow",
"#7FFF7F", "cyan", "#007FFF", "blue", "#00007F"))
corrplot(M, method="color", col=col4(20), cl.length=21, order = "AOE",
addCoef.col="green").
```

Получим раскрашенную матрицу, представленную на рис. 1.83.

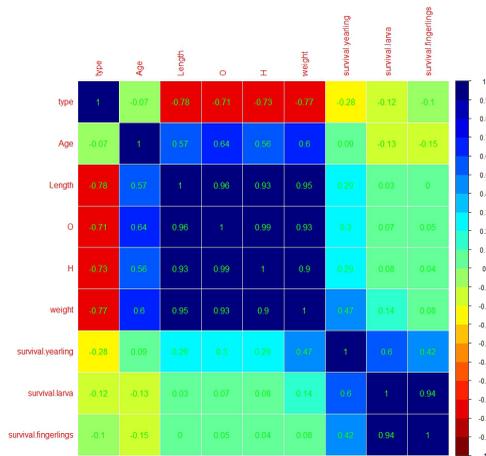


Рис. 1.83. Цветная корреляционная матрица

Осуществить построение корреляционной матрицы возможно также через пакет R Commander, нажимая последовательно **Графики** → **Матрица точечных графиков** (рис. 1.84).

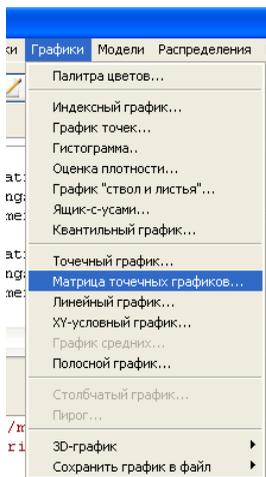


Рис. 1.84. Диалоговое окно, показывающее принцип выбора матрицы точечных графиков

Построение корреляционной матрицы представлено на рис. 1.85.

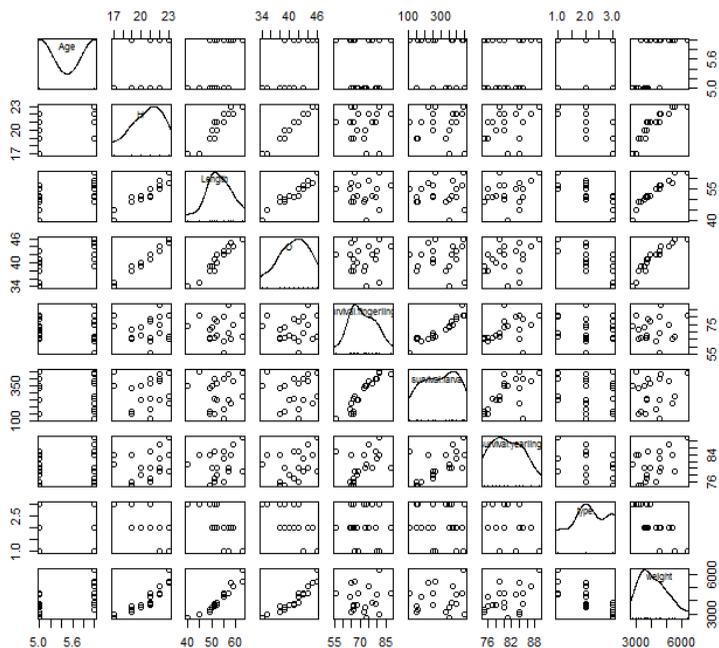


Рис. 1.85. Построение корреляционной матрицы

### *Прочие графики и рисунки.*

Графические возможности R представлены множеством вариантов и сочетаний и ежемесячно совершенствуются. Как было указано выше, читатель, интересующийся всем спектром графических возможностей R, может посетить сайт [R Graph Gallery](#), а также другие подобные сайты, на которых представлены не только примеры всевозможных графиков, но и исходный R-код, использованный для их построения. Так, например, хорошие рисунки можно строить в пакете R Commander в функции **Графики**. Для активации доступности некоторых графиков необходимо изменить вид ваших данных, например две колонки – на несколько колонок с данными каждой группы в отдельной колонке или наоборот и т. д. На рис. 1.86–1.90 представлены простейшие примеры графиков.

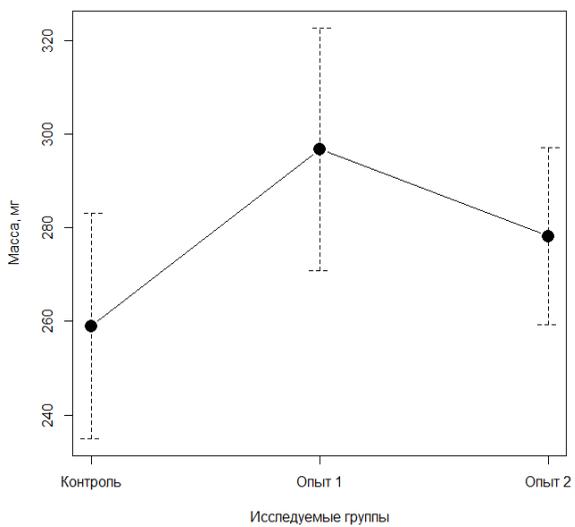


Рис. 1.86. График средних («усы» показывают размах стандартной ошибки средней)

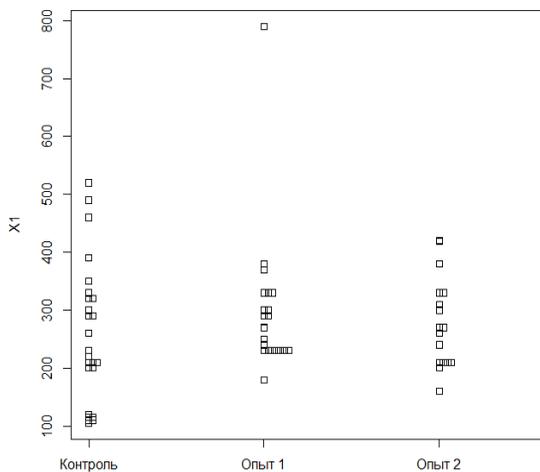


Рис. 1.87. Полостной график

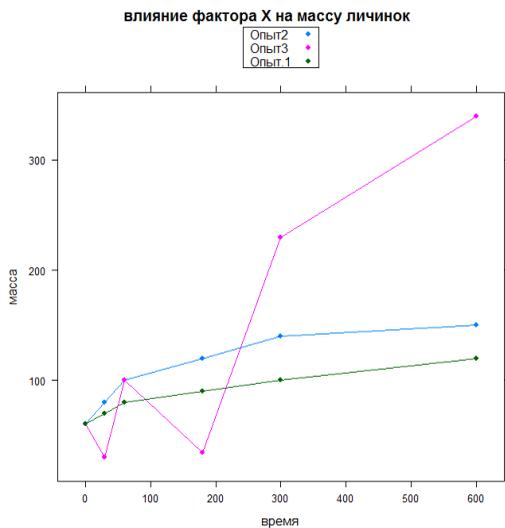


Рис. 1.88. XY – условный график

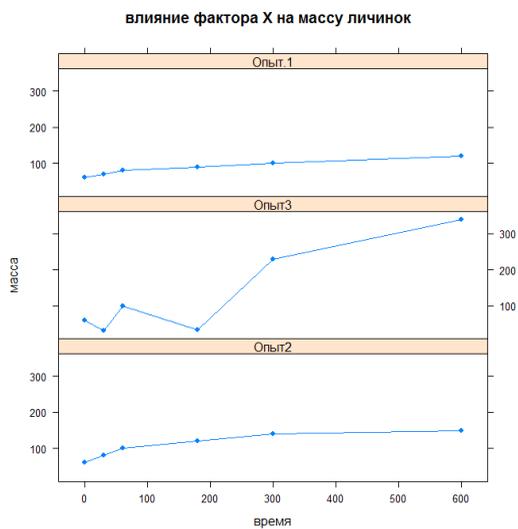


Рис. 1.89. XY – условный график, разделенный на различные панели

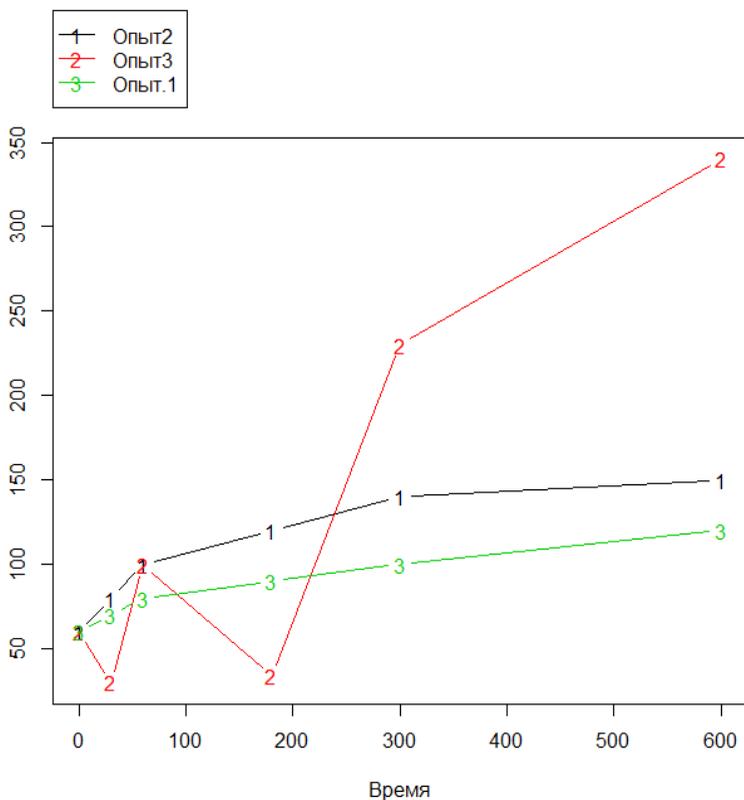


Рис. 1.90. Линейный график

Также интересное графическое отображение можно получить, используя пакет `beeswarm()`, который необходимо предварительно установить и запустить.

Использование этого пакета рассмотрим на примере данных, полученных при исследовании концентрации фермента АСТ в сыворотке крови у положительно (А) и отрицательно (В) ответивших самок сибирского осетра на гормональное инъецирование в различных группах, под влиянием фактора X (табл. 1.13).

Таблица 1.13. Концентрация фермента АСТ в сыворотке крови у сибирского осетра

	group	AST	response		group	AST	response
1	control	40	A	31	opyt1	140	B
2	control	50	A	32	opyt1	120	B
3	control	60	A	33	opyt1	123	B
4	control	30	A	34	opyt1	100	B
5	control	40	A	35	opyt1	130	B
6	control	34	A	36	opyt1	120	B
7	control	40	A	37	opyt1	150	B
8	control	50	A	38	opyt1	190	B
9	control	60	A	39	opyt1	190	B
10	control	70	A	40	opyt1	200	B
11	control	120	B	41	opyt2	300	B
12	control	130	B	42	opyt2	250	B
13	control	120	B	43	opyt2	230	B
14	control	110	B	44	opyt2	230	B
15	control	130	B	45	opyt2	200	B
16	control	120	B	46	opyt2	250	B
17	control	130	B	47	opyt2	230	B
18	control	120	B	48	opyt2	200	B
19	control	100	B	49	opyt2	190	B
20	control	130	B	50	opyt2	189	B
21	opyt1	20	A	51	opyt2	189	B
22	opyt1	30	A	52	opyt2	203	B
23	opyt1	20	A	53	opyt2	210	B
24	opyt1	20	A	54	opyt2	250	B
25	opyt1	30	A	55	opyt2	340	B
26	opyt1	40	A	56	opyt2	200	B
27	opyt1	100	B	57	opyt2	199	B
28	opyt1	120	B	58	opyt2	23	A
29	opyt1	120	B	59	opyt2	24	A
30	opyt1	120	B	60	opyt2	20	A

Введем следующий скрипт:

```

bimodal <- c(rnorm(250, -2, 0.6), rnorm(250, 2, 0.6))
uniform <- runif(500, -4, 4)
normal <- rnorm(500, 0, 1.5)
dataf <- data.frame (group = rep(c("bimodal","uniform", "normal"),
each = 500), xv = c(bimodal, uniform, normal), cg = rep( c("A","B"), 750))
require(beeswarm)
beeswarm(AST ~ group, data = Dataset,method = 'swarm', pch = 16,
pwcol = as.numeric(response), xlab = "", ylab = 'AST', labels = c('контроль',
'опыт1', 'опыт2')).

```

В результате получим рис. 1.91, на котором черные точки – это концентрация АСТ у положительно ответивших на гормональное стимулирование самок, а красные точки – концентрация АСТ у отрицательно ответивших на данное инъектирование самок. Из рисунка видно, что в контрольной группе доля положительно и отрицательно ответивших самок составляет около 50 %. В первой опытной группе под воздействием фактора X в одной дозировке количество отрицательно ответивших самок увеличилось, во второй опытной группе под воздействием фактора X в другой дозировке это количество увеличилось еще больше.

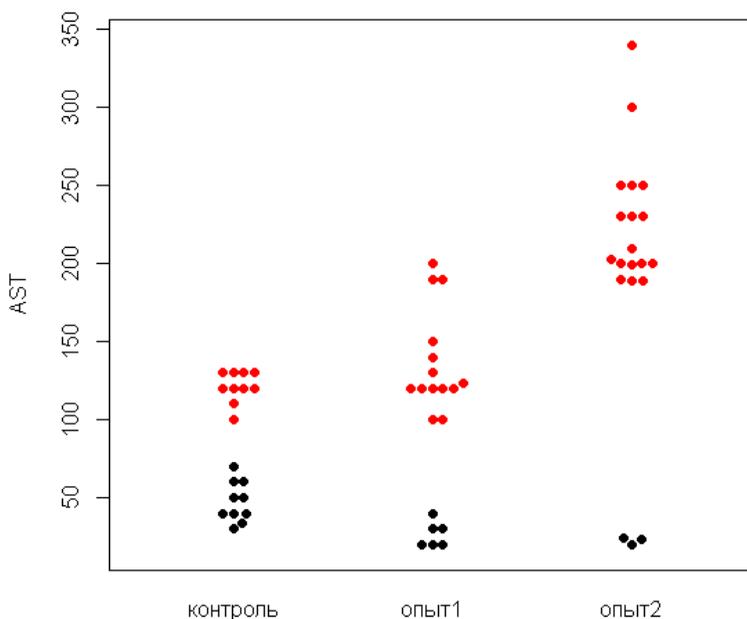


Рис. 1.91. График с использованием пакета beeswarm

График выглядит еще более информативным, если исследуемых данных больше (рис. 1.92).

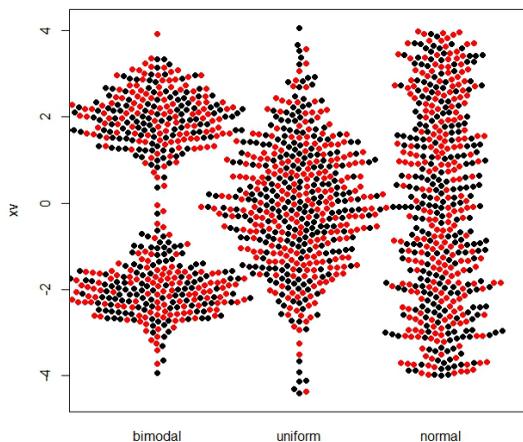


Рис. 1.92. График с использованием пакета beeswarm

Также более подробно с методами визуализации данных можно ознакомиться в электронной книге С. Э. Мостицкого, В. К. Шитикова «Статистический анализ и визуализация данных с помощью R» (2014) (адрес доступа: <http://r-analytics.blogspot.com>) и в книге Г. Джеймса, Д. Уиттона, Т. Хасты, Р. Тибширани «Введение в статистическое обучение с примерами на языке R» (2016) (адрес доступа: <http://dmkpress.com/catalog/computer/statistics/978-5-97060-293-5/>).

## 2. АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В СТАТИСТИЧЕСКОМ АНАЛИЗЕ

Использование алгоритмов машинного обучения в статистическом анализе рассмотрим на примере идентификации пола стерляди по строению спинных костных пластин.

В результате исследования морфологического строения спинных костных пластинок стерляди был собран массив данных, включающий длину, ширину каждой костной пластинки и другие морфологические изменения. На основании полученных морфологических изменений были рассчитаны морфологические индексы.

Далее были отобраны наиболее значимые морфологические параметры и индексы для идентификации пола при использовании нейронных сетей, метода Random Forrest и алгоритма Boruta.

При использовании алгоритма Boruta для отбора наиболее значимых полоспецифических морфологических индексов было установлено, что при анализе первых десяти спинных костных пластин наиболее высокие значения важности (mean importance или meanImp) имел индекс Дз/Шз – 40,77, а также число зубцов – 27,94 (рис. 2.1). Результаты при использовании нейронных сетей, метода Random Forrest (рис. 2.2–2.4), а также алгоритма Feature Importance были аналогичными.

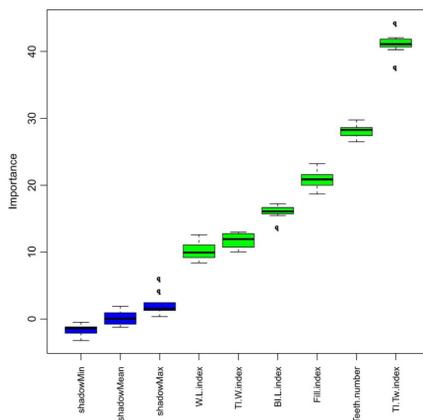


Рис. 2.1. Визуализация селекции значимых морфологических индексов с использованием алгоритма Boruta для создания модели определения пола взрослой стерляди

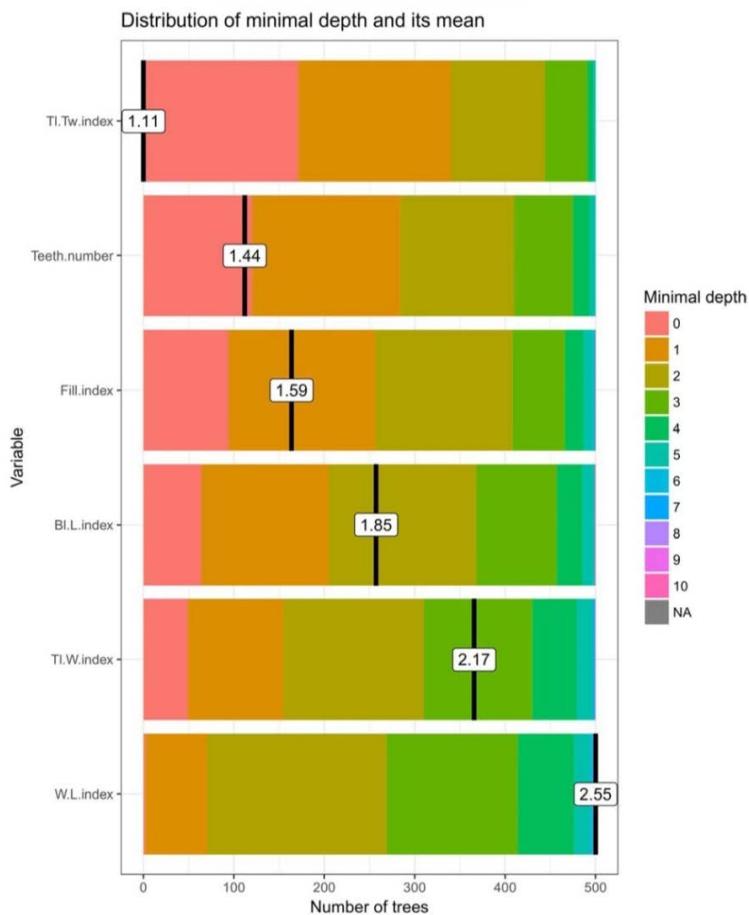
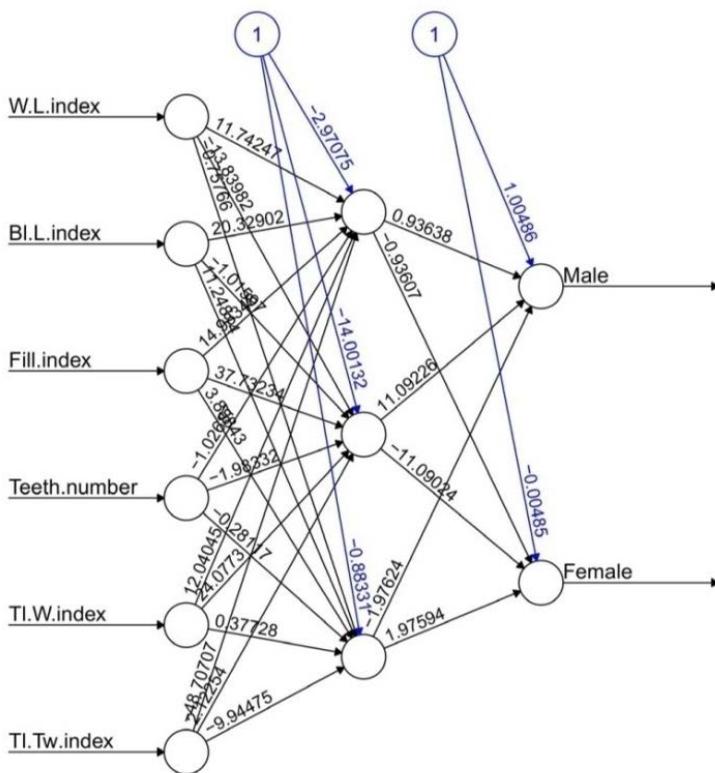


Рис. 2.2. Визуализация селекции значимых морфологических индексов с использованием метода Random Forrest для создания модели определения пола взрослой стерляди



Error: 14.631941 Steps: 11020

Рис. 2.3. Пример схемы нейронных сетей для селекции значимых морфологических индексов

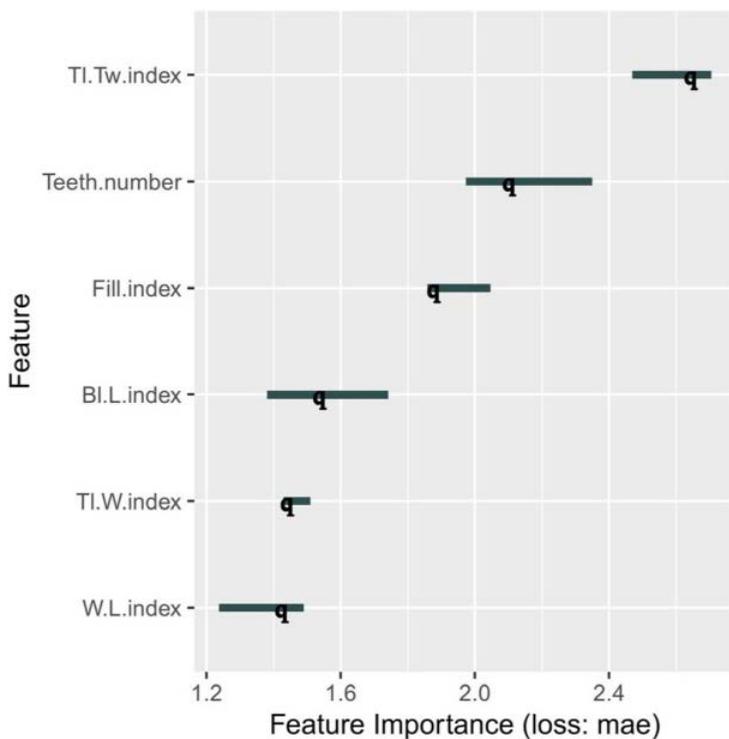
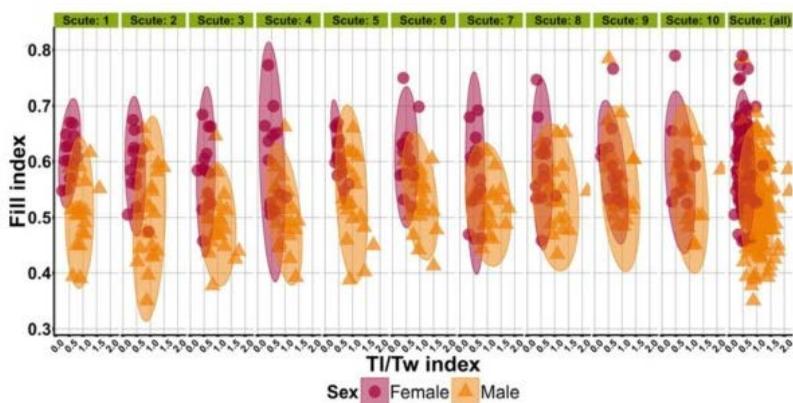
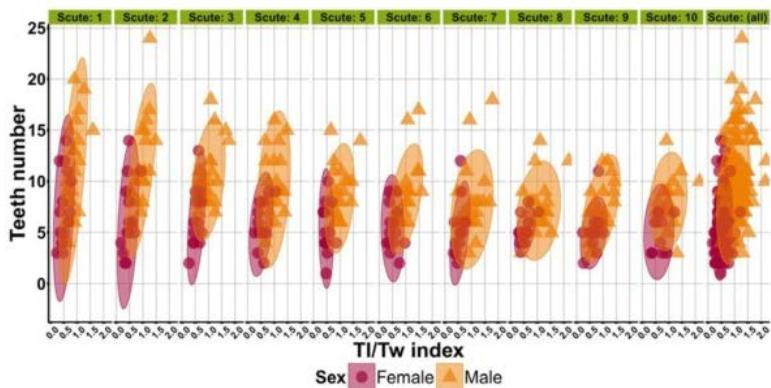


Рис. 2.4. Визуализация селекции значимых морфологических индексов с использованием алгоритма Feature Importance для создания модели определения пола взрослой стерляди

На основании осуществленного отбора можно выделить наиболее значимые морфологические индексы при анализе первых десяти спинных костных пластин: индекс Дз/Шз, число зубцов и индекс заполнения. Однако построение графика распределения точек и ординационных диаграмм не позволило выявить выраженное образование полоспецифических классов (кластеров) между отобранными показателями морфологических индексов спинных жучек взрослой стерляди (рис. 2.5, 2.6).

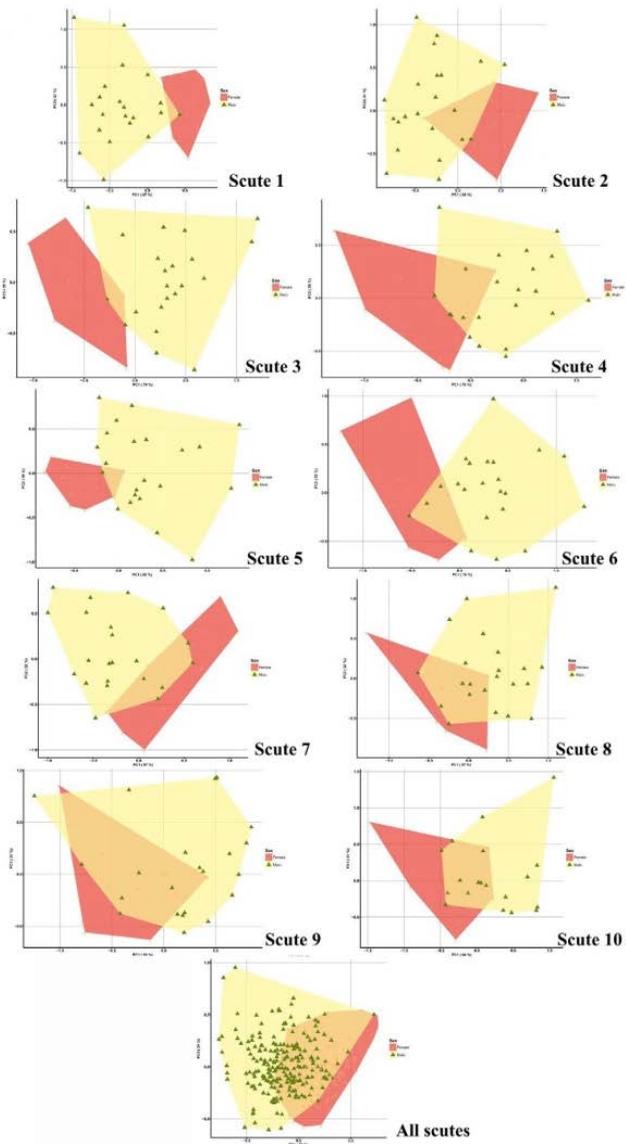


*a*

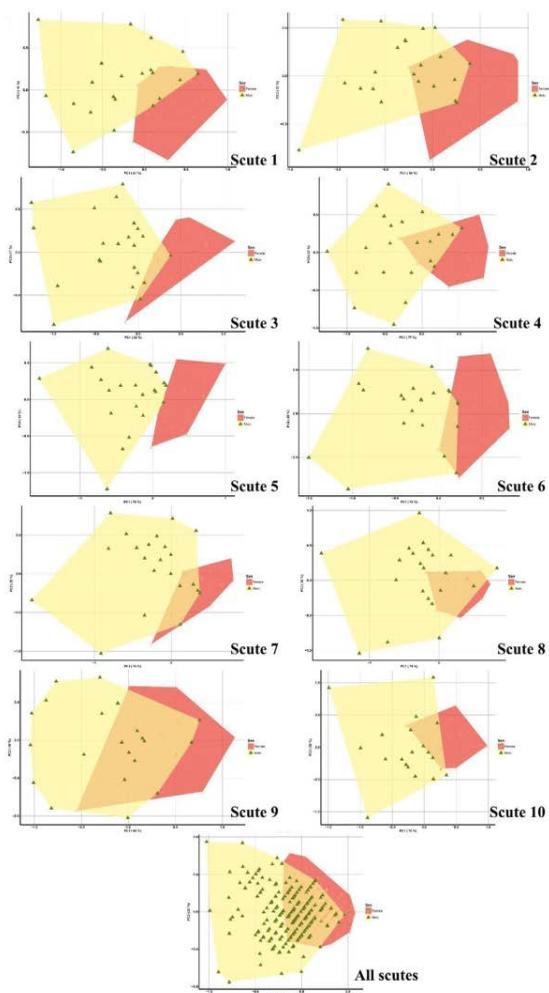


*b*

Рис. 2.5. Результаты распределения точек, отражающие образование полоспецифических классов (кластеров) между отобранными показателями морфологических индексов спинных жучек взрослой стерляди: *a* – между индексом заполнения и индексом Дз/Шз; *b* – между индексом Дз/Шз и числом зубцов (количество просмотренных рыб – 37 шт.)



*a*



*б*

Рис. 2.6. Ординационная диаграмма, отражающая образование популяционно-географических классов (кластеров) между отобранными показателями морфологических индексов спинных жучек взрослой стерляди: *а* – между индексом заполнения и индексом Дз/Шз; *б* – между индексом Дз/Шз и числом зубов (количество просмотренных рыб – 37 шт.)

Для получения наиболее точных моделей определения пола нами использовался метод построения деревьев решений (рис. 2.7) на основе рекурсивного разбиения. Мы построили шесть моделей, применяя указанный метод: первая модель, использующая только индекс Дз/Шз, вторая – только индекс заполнения, третья – только число зубов, четвертая (комбинированная модель № 1) – индекс Дз/Шз и число зубов, пятая (комбинированная модель № 2) – индекс Дз/Шз и индекс заполнения, шестая (комбинированная модель № 3) – индекс заполнения и число зубов.

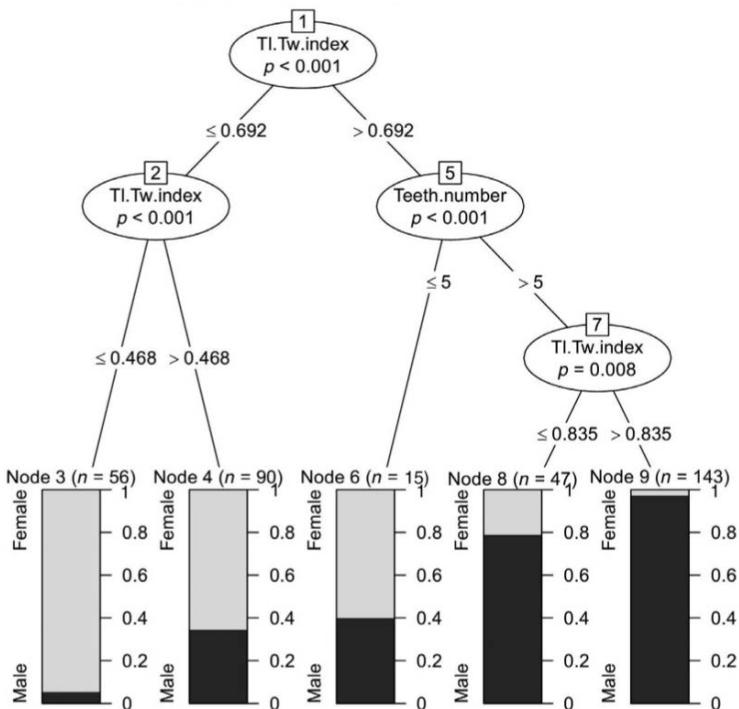


Рис. 2.7. Лучшая модель дерева решений для идентификации пола молодой стерляди по первым десяти спинным костным пластинкам, в которых использованы индекс Дз/Шз и число зубов

На основании построенных деревьев решений нами была составлена балльная шкала для определения пола по первым десяти спинным жучкам для каждой из вышеперечисленных моделей (табл. 2.1).

**Таблица 2.1. Балльная шкала для идентификации пола стерляди по первым десяти спинным костным пластинкам на основании результатов построения деревьев решений**

№ модели/индекс	Формула начисления баллов	Точность определения, %
Модель № 1/индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз > 0,692	93,55
Модель № 2/индекс заполнения	1 балл для каждой жучки, если индекс заполнения ≤ 0,523	73,33
Модель № 3/число зубцов	1 балл для каждой жучки, если число зубцов > 5 шт.	86,67
Модель № 4 (комбинированная модель № 1)/индекс Дз/Шз + число зубцов	1 балл для каждой жучки, если индекс Дз/Шз > 0,692 и число зубцов > 5 шт.	96,67
Модель № 5 (комбинированная модель № 2)/индекс Дз/Шз + индекс заполнения	1 балл для каждой жучки, если индекс Дз/Шз > 0,857 или индекс Дз/Шз > 0,692 и индекс заполнения ≤ 0,526	93,33
Модель № 6 (комбинированная модель № 3)/индекс заполнения + число зубцов	1 балл для каждой жучки, если индекс заполнения ≤ 0,595 и число зубцов > 5 шт.	93,33

При тестировании вышеописанных шести моделей для определения пола стерляди по первым десяти спинным жучкам в производственных условиях количество набираемых баллов достоверно выше ( $p < 0,001$ , критерий Манна – Уитни) было у самцов, а именно (8,26 ± 0,50) балла для модели № 1, (5,15 ± 0,55) балла для модели № 2, (9,36 ± 0,27) балла для модели № 3, (8,00 ± 0,51) балла для модели № 4, (7,26 ± 0,64) балла для модели № 5, (8,10 ± 0,38) балла для модели № 6. Тестируемые самки набрали следующие баллы: (1,75 ± 0,53) балла при использовании модели № 1, (1,09 ± 0,53) балла при использовании модели № 2, (5,00 ± 0,94) балла при использовании модели № 3, (1,27 ± 0,54) балла при использовании модели № 4, (0,63 ± 0,36) балла при использовании модели № 5, (2,27 ± 0,64) балла при использовании модели № 6 (рис. 2.8).

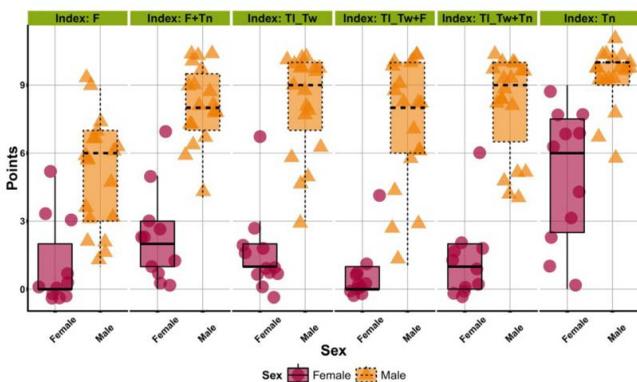


Рис. 2.8. Совмещенная диаграмма одномерного рассеяния и размахов распределения баллов при производственном тестировании моделей № 1–6 для определения пола половозрелой стерляди по первым десяти спинным жучкам

Для оценки точности полученных моделей определения пола стерляди по первым десяти спинным жучкам использовали бинарную матрицу, а также метод *binary discriminant analysis* с использованием алгоритма binDA и метод бинарного дерева решений с использованием алгоритма ID3 (Interactive Dichotomizer). Критерием перехода из 1 в 0 в бинарной матрице являлись полученные результаты при использовании метода построения деревьев решений на основе рекурсивного разбиения, а также полученные результаты при производственном тестировании данных моделей. Если общее число набираемых баллов было  $\geq 5$  (для модели № 1),  $\geq 6$  (для модели № 2),  $\geq 9$  (для модели № 3),  $\geq 4$  (для модели № 4),  $\geq 3$  (для модели № 5),  $\geq 6$  (для модели № 6), то исследуемый экземпляр рыбы относился к самцам (1 в бинарной матрице). Если исследуемый экземпляр рыбы набирал меньшее количество баллов по каждой модели, то он относился к самкам (0 в бинарной матрице).

Согласно оценке точности полученных моделей (см. табл. 2.1), лучшей моделью оказалась модель № 4, или комбинированная модель № 1, которая показала 96,67 % точности.

Таким образом, для определения пола стерляди при использовании первых десяти спинных жучек нами рекомендуется применение комбинированной модели с двумя параметрами: индекс Дз/Шз + число зубцов.

Однако определение пола стерляди при использовании первых десяти спинных жучек будет представлять достаточно длительный процесс в практике аквакультуры, поэтому нами была оценена возможность использования для определения пола стерляди первых пяти жучек, первых трех жучек и отдельные жучки.

При анализе первых пяти спинных костных пластин с использованием алгоритма Boruta морфологические индексы разместились по убыванию значения важности (meanImp) в следующем порядке: индекс Дз/Шз (34,59), индекс заполнения (20,36), число зубцов (16,27), индекс Дз/Ш (13,28), индекс Дл/Д (12,45), индекс Ш/Д (10,55). Результаты при использовании нейронных сетей, метода Random Forrest, а также алгоритма Feature Importance были аналогичными.

На основании осуществленного анализа можно выделить наиболее значимые морфологические индексы при анализе первых пяти спинных костных пластин: индекс Дз/Шз и индекс заполнения.

Нами было построено три модели с применением метода построения деревьев решений на основе рекурсивного разбиения: первая модель, использующая только индекс Дз/Шз, вторая – только индекс заполнения, третья (комбинированная) – индекс Дз/Шз и индекс заполнения (рис. 2.9).

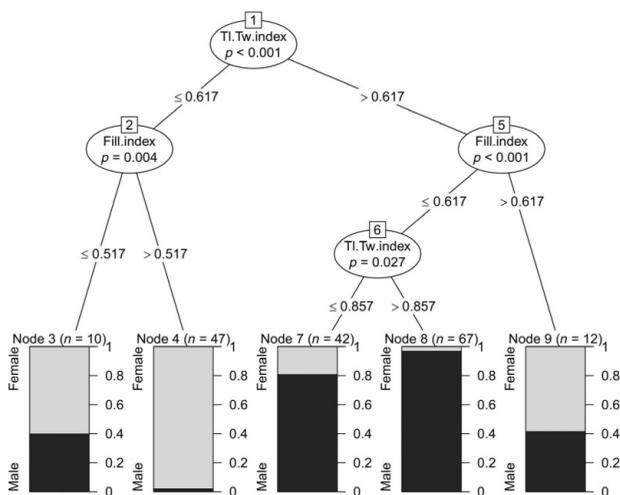


Рис. 2.9. Лучшая модель дерева решений для идентификации пола молоди стерляди по первым пяти спинным костным пластинкам, в которой использованы индекс Дз/Шз и индекс заполнения

На основании построенных деревьев решений нами была составлена балльная шкала для определения пола по первым пяти спинным жучкам для каждой из вышеперечисленных моделей (табл. 2.2).

Таблица 2.2. Балльная шкала для идентификации пола стерляди по первым пяти спинным костным пластинкам на основании результатов построения деревьев решений

№ модели/индекс	Формула начисления баллов	Точность определения, %
Модель № 1/индекс заполнения	1 балл для каждой жучки, если индекс заполнения $\leq 0,597$	80,56
Модель № 2/индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,617$	94,44
Модель № 3 (комбинированная)/индекс Дз/Шз + индекс заполнения	1 балл для каждой жучки, если индекс Дз/Шз $> 0,617$ и индекс заполнения $\leq 0,617$	97,06

При тестировании вышеописанных трех моделей для определения пола стерляди по первым пяти спинным жучкам в производственных условиях количество набираемых баллов достоверно выше ( $p < 0,001$ , критерий Манна – Уитни) было у самцов, а именно ( $4,54 \pm 0,15$ ) балла для модели № 1, ( $4,77 \pm 0,09$ ) балла для модели № 2, ( $4,95 \pm 0,04$ ) балла для модели № 3. Тестируемые самки набрали следующие баллы: ( $2,21 \pm 0,45$ ) балла при использовании модели № 1, ( $1,28 \pm 0,39$ ) балла при использовании модели № 2, ( $0,71 \pm 0,36$ ) балла при использовании модели № 3 (рис. 2.10).

Для оценки точности полученных моделей определения пола стерляди по первым пяти спинным жучкам использовали бинарную матрицу, а также метод *binary discriminant analysis* с использованием алгоритма binDA и метод бинарного дерева решений с использованием алгоритма ID3 (Interactive Dichotomizer). Если общее число набираемых баллов было  $\geq 3$  (для всех моделей), то исследуемый экземпляр рыбы относился к самцам (1 в бинарной матрице). Если исследуемый экземпляр рыбы набирал меньшее количество баллов по каждой модели, то он относился к самкам (0 в бинарной матрице).

Согласно оценке точности полученных моделей (см. табл. 2.2), лучшей моделью оказалась модель № 3, которая показала 97,06 % точности.

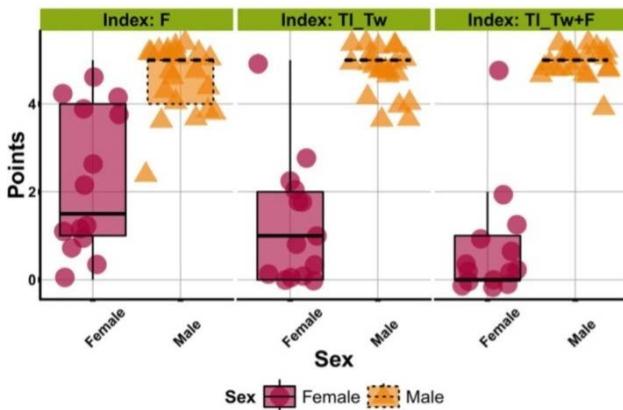


Рис. 2.10. Совмещенная диаграмма одномерного рассеяния и размахов распределения баллов при производственном тестировании моделей № 1–3 для определения пола половозрелой стерляди по первым пяти спинным жучкам

Таким образом, для определения пола стерляди при использовании первых пяти спинных жучек нами рекомендуется применение комбинированной модели с двумя параметрами: индекс Дз/Шз + индекс заполнения.

При анализе первых трех спинных костных пластин с использованием алгоритма Voruta морфологические индексы разместились по убыванию значения важности (meanImp) в следующем порядке: индекс Дз/Шз (27,42), индекс заполнения (16,85), индекс Дл/Д (14,15), число зубцов (12,25), индекс Ш/Д (10,75), индекс Дз/Ш (7,43). Результаты при использовании нейронных сетей, метода Random Forrest, а также алгоритма Feature Importance были аналогичными.

На основании осуществленного анализа можно выделить наиболее значимые морфологические индексы при анализе первых трех спинных костных пластин: индекс Дз/Шз и индекс заполнения.

Нами было построено три модели с применением метода построения деревьев решений на основе рекурсивного разбиения: первая модель, использующая только индекс Дз/Шз, вторая – только индекс заполнения, третья (комбинированная) – индекс Дз/Шз и индекс заполнения (рис. 2.11).

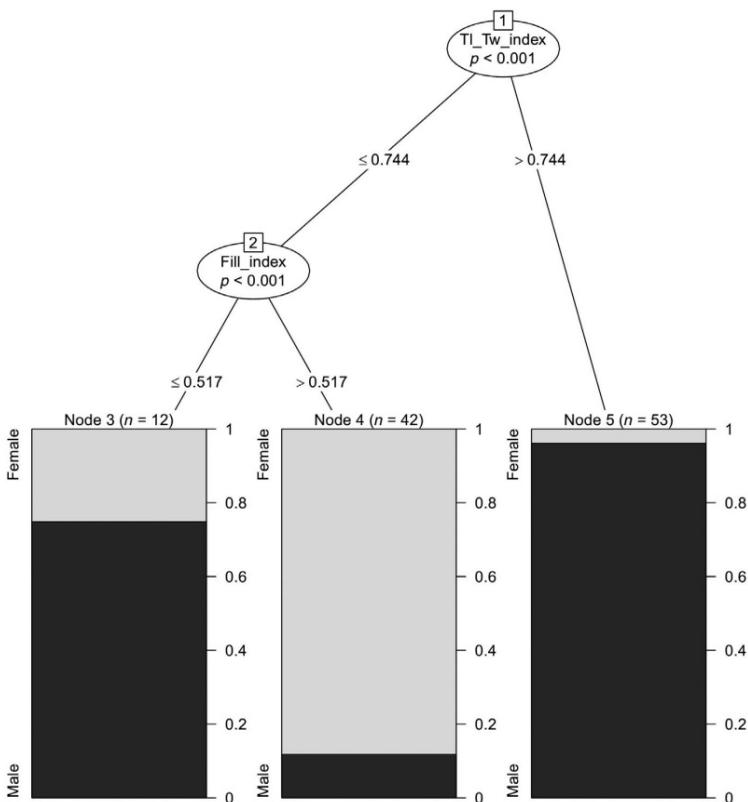


Рис. 2.11. Лучшая модель дерева решений для идентификации пола молоди стерляди по первым трем спинным костным пластинкам, в которой использованы индекс Дз/Шз и индекс заполнения

На основании построенных деревьев решений нами была составлена балльная шкала для определения пола по первым трем спинным жучкам для каждой из вышеперечисленных моделей (табл. 2.3).

При тестировании вышеописанных трех моделей для определения пола стерляди по первым трем спинным жучкам в производственных условиях количество набираемых баллов достоверно выше ( $p < 0,001$ , критерий Манна – Уитни) было у самцов, а именно ( $4,54 \pm 0,15$ ) балла для модели № 1, ( $4,77 \pm 0,09$ ) балла для модели № 2, ( $4,95 \pm 0,04$ ) балла для модели № 3. Тестируемые самки набрали следующие баллы:

( $2,21 \pm 0,45$ ) балла при использовании модели № 1, ( $1,28 \pm 0,39$ ) балла при использовании модели № 2, ( $0,71 \pm 0,36$ ) балла при использовании модели № 3 (рис. 2.12).

Таблица 2.3. Балльная шкала для идентификации пола стерляди по первым трем спинным костным пластинкам на основании результатов построения деревьев решений

№ модели/индекс	Формула начисления баллов	Точность определения, %
Модель № 1/индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,744$	82,86
Модель № 2/индекс заполнения	1 балл для каждой жучки, если индекс заполнения $\leq 0,52$	88,24
Модель № 3 (комбинированная)/индекс Дз/Шз + индекс заполнения	1 балл для каждой жучки, если индекс Дз/Шз $> 0,744$ или индекс Дз/Шз $\leq 0,744$ и индекс заполнения $\leq 0,517$	94,28

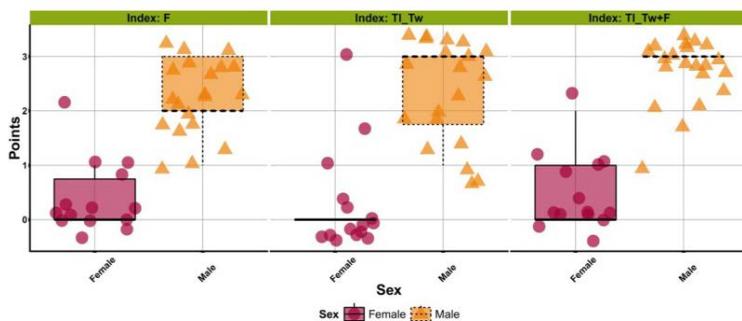


Рис. 2.12. Совмещенная диаграмма одномерного рассеяния и размахов распределения баллов при производственном тестировании моделей № 1–3 для определения пола половозрелой стерляди по первым трем спинным жучкам

Для оценки точности полученных моделей определения пола стерляди по первым трем спинным жучкам использовали бинарную матрицу, а также метод *binary discriminant analysis* с использованием алгоритма binDA и метод бинарного дерева решений с использованием алгоритма ID3 (Interactive Dichotomizer). Если общее число набираемых баллов было  $\geq 2$  (для всех моделей), то исследуемый экземпляр рыбы относился к самцам (1 в бинарной матрице). Если исследуемый

экземпляр рыбы набирал меньшее количество баллов по каждой модели, то он относился к самкам (0 в бинарной матрице).

Согласно оценке точности полученных моделей (см. табл. 2.3), лучшей моделью оказалась модель № 3, которая показала 94,28 % точности.

Таким образом, для определения пола стерляди при использовании первых трех спинных жучек нами рекомендуется применение комбинированной модели с двумя параметрами: индекс Дз/Шз + индекс заполнения.

При отдельном анализе каждой спинной костной пластинки (жучки) с использованием алгоритма Voruta наиболее высокие значения важности (meanImp) получили следующие морфологические индексы: индекс заполнения (14,79) и индекс Дз/Шз (11,82) для первой жучки; индекс Дз/Шз (17,35) для второй жучки; индекс Дз/Шз (14,56) и индекс Дл/Д (13,61) для третьей жучки; индекс Дз/Шз (16,17) для четвертой жучки; индекс Дз/Шз (21,28) для пятой жучки; индекс Дз/Шз (13,30) и число зубцов (11,75) для шестой жучки; индекс Дз/Шз (16,85) для седьмой жучки; индекс Дз/Шз (16,30) для восьмой жучки; число зубцов (13,92) для девятой жучки; число зубцов (10,59) и индекс Дз/Шз (8,71) для десятой жучки (рис. 2.13).

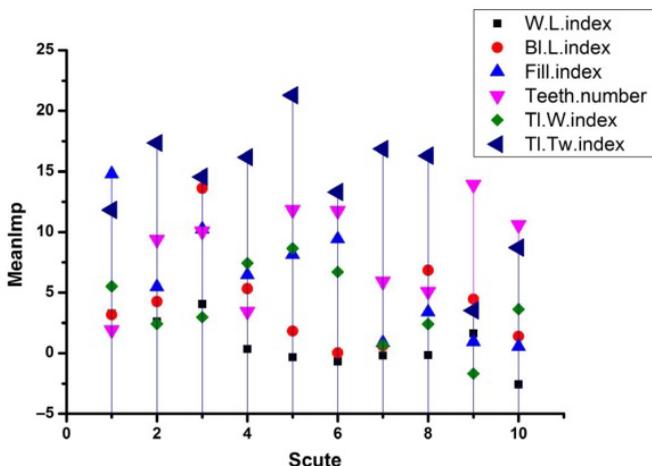


Рис. 2.13. Изменение среднего значения важности морфологических индексов спинных костных пластинок стерляди в зависимости от их номера

На основании осуществленного анализа можно выделить наиболее значимые морфологические индексы для каждой спинной костной пластинки: индекс Дз/Шз для жучек № 1–8, индекс заполнения для жучки № 1, число зубцов для жучки № 9, индекс Дл/Д для жучки № 3. Далее нами было осуществлено построение 14 моделей с применением метода построения деревьев решений на основе рекурсивного разбиения: первая модель, использующая только индекс Дз/Шз, вторая – только индекс заполнения, третья (комбинированная) – индекс Дз/Шз и индекс заполнения. На основании построенных деревьев решений нами была составлена балльная шкала для определения пола по каждой спинной жучке (табл. 2.4).

Таблица 2.4. Балльная шкала для идентификации пола стерляди для каждой спинной костной пластинки на основании результатов построения деревьев решений

№ спинной костной пластинки	Главный индекс	Формула начисления баллов	Точность определения, %
1	2	3	4
1	Индекс заполнения	1 балл для каждой жучки, если индекс заполнения $\leq 0,591$	88,57
	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,75$	82,86
2	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,51$	91,43
3	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,75$	85,71
	Индекс Дл/Д	1 балл для каждой жучки, если индекс Дл/Д $\leq 0,744$	82,86
4	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,63$	88,87
5	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,63$	91,67
6	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз $> 0,7$	83,33
	Число зубцов	1 балл для каждой жучки, если число зубцов $> 5$ шт.	80,56

1	2	3	4
7	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз > 0,79	83,33
8	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз > 0,71	86,11
9	Число зубцов	1 балл для каждой жучки, если число зубцов > 5 шт.	82,86
10	Число зубцов	1 балл для каждой жучки, если число зубцов > 7 шт.	70,00
	Индекс Дз/Шз	1 балл для каждой жучки, если индекс Дз/Шз > 0,76	76,67

Лучшие результаты показала модель для спинной жучки № 5 (91,67 %), при тестировании которой самцы набирали  $(1,00 \pm 0,00)$  балла, а самки –  $(0,20 \pm 0,10)$  балла, а также модель для спинной жучки № 2 (91,43 %), при тестировании которой самцы набирали  $(1,00 \pm 0,00)$  балла, а самки –  $(0,21 \pm 0,11)$  балла (рис. 2.14).

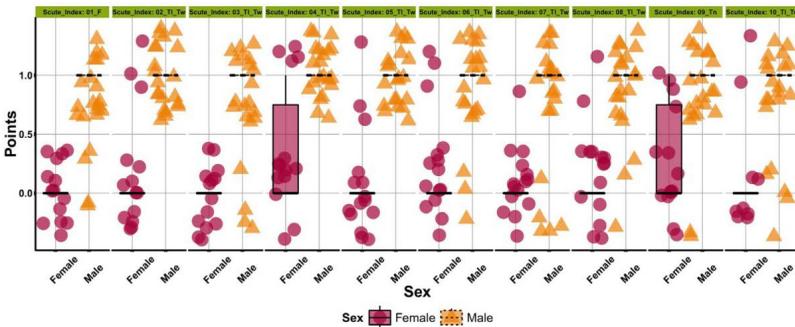


Рис. 2.14. Совмещенная диаграмма одномерного рассеяния и размахов распределения баллов при производственном тестировании моделей № 1–3 для каждой спинной костной пластинки

Примечание. F – индекс заполнения; T1\_Tw – индекс Дз/Шз; Tn – число зубцов; female – самка; male – самец; points – баллы; sex – пол

Таким образом, в результате проведенных исследований было установлено, что для диагностики пола по спинным костным пластинкам с использованием компьютерных алгоритмов анализа рекомендуется использовать комбинированные модели, осуществляющие определение пола по двум основным морфологическим индексам: индекс Дз/Шз и индекс заполнения. При этом точность определения пола таких моделей составляет: 96,67 % для первых десяти спинных жучек, 97,22 % для первых пяти спинных жучек и 94,28 % для первых трех спинных жучек. При необходимости определения пола по одной жучке рекомендуется использовать жучку № 2 или жучку № 5 и анализировать ее по индексу Дз/Шз (точность определения пола – 91,43 и 91,67 % соответственно).

Нами разработан протокол по определению пола у личинок, молодежи и взрослой стерляди на основании полученных результатов. В наших исследованиях изучались только спинные жучки стерляди. Исследование других рядов жучек стерляди нуждается в дополнительных исследованиях. Предварительные результаты показывают, что боковые и брюшные жучки стерляди также имеют различия в строении в зависимости от пола. Нами была получена модель, позволяющая с высокой вероятностью осуществлять определение пола у личинок, молодежи и взрослой стерляди, применяя методы нейронных сетей, Random Forrest и алгоритм Boruta, критерий  $\chi^2$ , binary discriminant analysis с использованием алгоритма binDA, бинарного дерева решений с использованием алгоритма ID3 (Interactive Dichotomizer).

Найденные зависимости с использованием алгоритмов машинного обучения открывают перспективы для создания оборудования для прижизненного определения пола при использовании искусственного интеллекта для распознавания изображений.

## СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	3
1. ИСПОЛЬЗОВАНИЕ ПРОГРАММНОЙ СРЕДЫ R ПРИ СТАТИСТИЧЕСКОМ АНАЛИЗЕ .....	4
1.1. Проверка на нормальность распределения.....	14
1.2. Проверка на однородность групповых дисперсий.....	20
1.3. Параметрические критерии .....	23
1.4. Непараметрические критерии.....	26
1.5. Оценка полученных результатов на соответствие нормативным значениям .....	30
1.6. Базовые графические возможности R.....	33
2. АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В СТАТИСТИЧЕСКОМ АНАЛИЗЕ .....	83

Учебное издание

**Барулин** Николай Валерьевич  
**Шумский** Константин Леонардович

ФУНДАМЕНТАЛЬНЫЕ И ПРИКЛАДНЫЕ  
НАУЧНЫЕ ИССЛЕДОВАНИЯ В АКВАКУЛЬТУРЕ

В трех частях

Часть 1

ИСПОЛЬЗОВАНИЕ ПРОГРАММНОЙ СРЕДЫ R  
ПРИ СТАТИСТИЧЕСКОМ АНАЛИЗЕ

Учебно-методическое пособие

Редактор *Н. Н. Пьянусова*  
Технический редактор *Н. Л. Якубовская*  
Корректор *Е. В. Ширалиева*

Подписано в печать 08.12.2022. Формат 60×84 <sup>1</sup>/<sub>16</sub>. Бумага офсетная.  
Ризография. Гарнитура «Таймс». Усл. печ. л. 6,04. Уч.-изд. л. 5,42.  
Тираж 25 экз. Заказ .

УО «Белорусская государственная сельскохозяйственная академия».  
Свидетельство о ГРИИРПИ № 1/52 от 09.10.2013.  
Ул. Мичурина, 13, 213407, г. Горки.

Отпечатано в УО «Белорусская государственная сельскохозяйственная академия».  
Ул. Мичурина, 5, 213407, г. Горки.