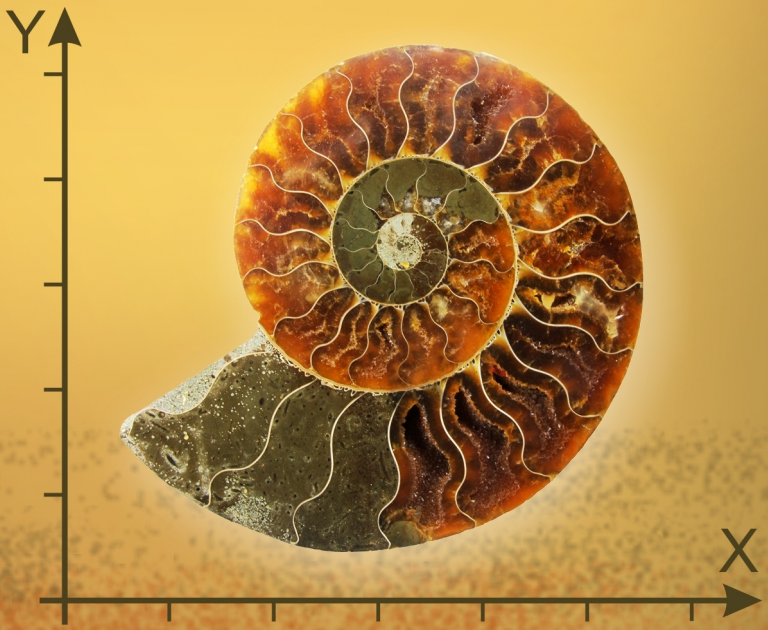
КЛАССИЧЕСКОЕ
УНИВЕРСИТЕТСКОЕ
ОБРАЗОВАНИЕ

Челябинский государственный университет

Д. Ю. Нохрин

ЛАБОРАТОРНЫЙ ПРАКТИКУМ ПО БИОСТАТИСТИКЕ



Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Челябинский государственный университет»

КЛАССИЧЕСКОЕ УНИВЕРСИТЕТСКОЕ ОБРАЗОВАНИЕ

Д. Ю. Нохрин

ЛАБОРАТОРНЫЙ ПРАКТИКУМ ПО БИОСТАТИСТИКЕ

Челябинск
Издательство Челябинского государственного университета
2018

УДК 57.08
ББК Е.я7
Н858

Серия основана в 2008 году

Печатается по решению редакционно-издательского совета
Челябинского государственного университета

Р е ц е н з е н т ы:
В. Е. Лазарев, доктор технических наук, доцент,
заведующий кафедрой двигателей внутреннего сгорания
и электронных систем автомобилей Южно-Уральского государственного университета
(национального исследовательского университета);
Л. А. Рязанова, кандидат биологических наук,
доцент кафедры общей биологии и физиологии
Южно-Уральского государственного гуманитарно-педагогического университета

Нохрин, Д. Ю.

Н858 Лабораторный практикум по биостатистике / Д. Ю. Нохрин. Челябинск : Изд-во Челяб. гос. ун-та, 2018. 289 с. (Классическое университетское образование).

ISBN 978-5-7271-1487-2

Представлены конспекты 18 лабораторных занятий по программе курса «Основы биометрического анализа и планирования эксперимента», содержащие необходимые сведения теоретического характера. На примерах из области биологии и медицины рассмотрен алгоритм расчётов с использованием статистического пакета, а также даются рекомендации по оформлению результатов исследования в квалификационной работе или научной статье.

Предназначено для бакалавров направления 06.03.01 «Биология», магистрантов направления 06.04.01 «Биология» (направленности — микробиология и вирусология, медико-биологические науки, лабораторная диагностика в клинической практике для биологов, прикладные и фундаментальные вопросы биотехнологии) и аспирантов направления 30.06.01 «Фундаментальная медицина (направленность — клиническая иммунология и аллергология)». Издание будет полезно школьникам старших классов с углублённым изучением биологии и аспирантам медицинских и биологических направлений обучения.

УДК 57.087.1(075.8)
ББК Е.с13я73-5

ISBN 978-5-7271-1487-2

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Лабораторный практикум разработан с целью повышения эффективности обучения студентов основам биостатистики, а также помощи в написании и оформлении студенческих квалификационных работ и научных статей. В основу практикума положен курс «Основы биометрического анализа и планирования эксперимента», который автор длительное время читает студентам биологического факультета Челябинского государственного университета, поэтому объём и структура издания находятся в полном соответствии с рабочей программой указанной дисциплины. Вместе с тем форма подачи материала в виде развёрнутого конспекта занятия позволяет использовать его для самостоятельного освоения широкого спектра статистических методов школьниками старших классов, студентами и аспирантами, проводящими научные исследования в области естественных наук и медицины.

Структура практикума и самостоятельная работа с ним

Практикум состоит из 18 лабораторных занятий, на которые отводится 36 аудиторных академических часов. В его основе лежит принцип деления практических задач на 7 категорий, а исследуемых признаков (типов данных) — на 3 категории. Таким образом, значительная часть наиболее востребованных в исследовательской практике статистических методов может быть структурно представлена в виде таблицы 7×3, которую мы предлагаем заполнить читателю самостоятельно по мере прохождения курса.

Перечислим и кратко охарактеризуем ***типичные задачи*** в исследовании.

1. Описание данных. Данная задача решается в подавляющем большинстве работ, поскольку исследователю необходимо представить результаты в компактном виде, должным образом их обобщив. Например, для количественных показателей это может быть среднее значение или медиана, для качественных —

частота в процентах; также приводят меры рассеяния или точности оценки показателя по выборке. Поскольку грамотное описание данных подразумевает знание характера распределения признака, этапу описания данных часто предшествует анализ распределения: графический и/или статистический. Эта задача рассматривается на лабораторных занятиях № 2–4 (табл. 1).

2. Сравнение двух выборок. Очень распространённая задача. Обычно одна выборка является экспериментальной (в медицине — «основная группа»), а вторая — контрольной (в медицине — «группа сравнения»). Также это могут быть группы разного пола, возраста, разных видов и т. д. Сравнение проводят по мерам положения, рассеяния и форме распределения. На лабораторных занятиях № 5–7 данная задача рассматривается только применительно к мерам положения.

3. Сравнение нескольких выборок. Также распространённая задача. Это могут быть несколько экспериментальных групп и контрольная группа, разные виды, условия и т. д. Если исследуется не один действующий фактор, а несколько — требуется одновременная их оценка для выявления возможных неаддитивных эффектов взаимодействия факторов. Методы для решения этой задачи рассматриваются применительно только к мерам положения на занятиях № 8–10.

4. Поиск связей. Важная задача, которую часто называют поиском корреляций или ассоциаций между признаками. В этом случае все признаки рассматриваются как равноценные, то есть их отношения по типу «причина — следствие» если и возможны, то не предполагаются строго. Этой задаче посвящено занятие № 11.

5. Поиск зависимостей. Данная задача подразумевает наличие двух разнокачественных показателей: одни являются независимыми (регрессорами), а другие — зависимыми (откликами). Установление факта наличия зависимости откликов от регрессоров, а также её формы проводится в ходе регрессионного анализа, для которого разработано большое число статистических техник. Также цель может заключаться в прогнозе величины количественного показателя или прогнозе наступления качественного состояния. Ряд техник регрессионного анализа рассматривается на лабораторных занятиях № 12–14. В экологических работах нередко требуется изобразить значения показателей с привязкой

к местности и построить карту (загрязнения, распространения видов и т. п.). Задаче интерполяции пространственных данных посвящено занятие № 15.

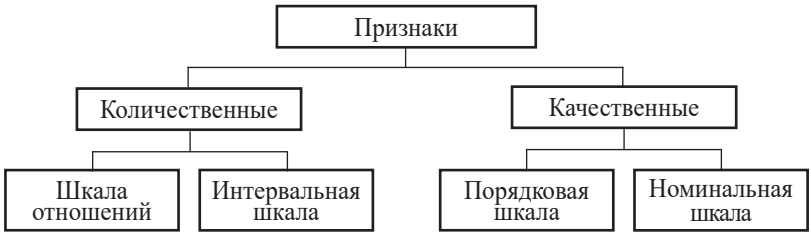
6. Многомерный разведочный (эксплораторный) анализ. Зачастую показателей в работе оказывается так много, а многочисленные связи между ними представляются столь непростыми, что исследователю сложно приступить к анализу: неясно, с чего нужно начинать в первую очередь. Методы эксплораторного анализа помогают обобщить сложный материал, увидеть в нём разнонаправленные тенденции, естественные группировки объектов и сходным образом изменяющиеся показатели. При этом ценность в работе может представлять как достигнутый с применением многомерных техник уровень обобщения, так и напротив — сужение спектра гипотез до 1–2 простых, которые удобнее проверять менее сложными классическими статистическими техниками. Типичные методы разведочного анализа рассматриваются на занятиях № 16–17.

7. Специфические задачи. В целом таких задач много, хотя в конкретных областях биологии и медицины это единичные специфические случаи. Объём практического курса не позволяет рассмотреть даже по 1–2 таких задач для каждого из направлений обучения, поэтому было решено включить три специальные задачи, выбранные по другим принципам. Анализ выживаемости (занятие № 14) знакомит с таким особым типом данных, как неполные, или цензурированные наблюдения. Анализ чувствительности и специфичности диагностических методов и тест-систем (занятие № 14) крайне важен в экспериментальной биологии и медицине. На последнем, 18-м занятии рассматриваются необходимые для любого исследования вопросы планирования: выбор экспериментального плана и расчёты объёмов выборок.

Перечислим и кратко охарактеризуем ***типы признаков*** и ***шкалы данных***. Все биологические признаки можно разделить на количественные и качественные, представленные четырьмя шкалами (см. теоретический материал и рис. на с. 8).

С точки зрения анализа данных их удобнее разбить на 3 группы.

I. Количественные признаки с нормальным распределением (шкала отношений и интервальная шкала). Методы анализа таких признаков разработаны очень хорошо и составляют



направление классической ***параметрической статистики***, задействующей в расчётах параметры известных распределений, главным образом — параметры нормального распределения (математическое ожидаемое μ «мю» и стандартное отклонение σ «сигма»). Отметим, что вопреки распространённому заблуждению нормальным распределение должно быть не в выборке, а в генеральной совокупности, откуда данная выборка извлечена.

II. Количественные признаки с ненормальным распределением и порядковые признаки (шкала отношений, интервальная шкала и порядковая шкала). Большинство биологических показателей относится к этому типу. Если мы не уверены в нормальности распределения признака или точно знаем, что он распределён ненормально, анализ данных можно провести тремя способами:

1) нормализовать данные с помощью специальных преобразований шкалы (логарифмирование, преобразование арксинуса, Бокса — Кокса и др.) и использовать далее параметрические методы;

2) преобразовать шкалы отношений или интервалов в порядковую шкалу и работать далее ***непараметрическими методами*** порядковой статистики. Такой способ традиционен и популярен (медианы и квартили, корреляция Спирмена, критерии Уилкоксона — Манна — Уитни, Краскела — Уоллиса, Фридмана и др.). Однако важно отдавать себе отчёт в том, что понижение шкалы до порядковой сопровождается определённой потерей информации;

3) работать с исходными непреобразованными данными методами, устойчивыми к отклонениям от нормальности. Это могут быть либо методы робастной статистики (усечённые средние или отличные от среднего М-оценки, средние абсолютные отклонения вместо среднеквадратичных и т. д.), либо современные

ресэмплинг-техники, основанные на вычислительных возможностях компьютеров (складной нож, бутстреп, рандомизационные методы Монте-Карло).

При решении разных задач мы так или иначе познакомимся со всеми тремя способами. Исключение составят методы робастной статистики, которые в практике биостатистики не являются традиционными и применяются редко.

III. Качественные номинальные признаки (номинальная шкала). Такие признаки представляют собой определённые состояния (вид, пол, цвет, есть или нет, жив или мёртв и т. п.). Обычно их описывают частотами, выраженными в процентах от общего числа (а также в промилле, единицах на 10 тыс. и т. п.).

* * *

При самостоятельном прохождении данного практического курса рекомендуем придерживаться имеющегося порядка выполнения лабораторных занятий. Тем не менее быстрый поиск необходимой группы методов может быть осуществлён с помощью карты лабораторных занятий (табл. 1). Для этого нужно выбрать задачу для статистического анализа в строке, тип данных — в столбце, а на пересечении найти номер занятия. Если на пересечении окажется пустой блок ($-^{n}$), значит, в практикуме методы данной группы не рассматривается; самостоятельный их поиск можно начать с методов, указанных в примечании к таблице.

Таблица 1

Карта лабораторных занятий
в зависимости от задачи исследования и типа данных

Задача	Признак		
	I Количественные с нормальным распределением	II Количественные с ненормальным распределением и порядковые	III Качественные номинальные
1. Описание данных			
– меры	№ 2		
– графики	№ 3		
– распределения	№ 4		$-^{1}$

Окончание табл. 1

Задача	Признак		
	I Количественные с нормальным распределением	II Количественные с ненормальным распределением и порядковые	III Качественные номинальные
2. Сравнение двух выборок по мерам положения			
– независимые выборки	№ 5		№ 6
– зависимые выборки	№ 7		
3. Сравнение трёх и более выборок по мерам положения			
– однофакторное	№ 8		№ 9
– многофакторное	№ 10		–[2]
4. Поиск связей	№ 11		
5. Поиск зависимостей			
– линейные	№ 12	–[3]	–[4]
– нелинейные	№ 13		
– в пространстве (карты)	№ 15		
6. Многомерный эксплораторный анализ			
– кластеризация	№ 16		
– ординация (проекция)	№ 16		№ 17
7. Специфические задачи			
– анализ выживаемости	№ 14		
– диагностические тесты	№ 14		
– планирование исследования	№ 18		

Примечание. [1] Анализ распределения для качественных признаков: биномиальное, отрицательное биномиальное, пуассоновское. [2] Многофакторное сравнение трёх и более выборок для качественных признаков: иерархический логлинейный анализ (Loglinear analysis). [3] Линейные зависимости для ненормально распределённых признаков: робастная регрессия Кендалла — Тейла (Kendall-Teil regression). [4] Линейные зависимости для качественных номинальных признаков: анализ таблицы сопряжённости на тренд методом Кохрана — Армитаж (Cochran-Armitage test for trend) или на линейную ассоциацию методом Мантеля — Хензеля (Mantel-Haenszel linear-by-linear association).

Обратите внимание на правильное написание названий статистических методов, указанных в примечании. В предстоящем практическом курсе будет много статистических техник, названных в честь их разработчиков. Если разработчиков было два и более, то согласно правилам русского языка такие сложные эпонимы пишутся через знак тире: критерий Уилкоксона — Манна — Уитни, преобразование Бокса — Кокса, регрес-

сия Кендалла — Тейла и т. п. В англоязычном написании таких терминов используется знак дефис: Wilcoxon-Mann-Whitney test, Box-Cox transformation, Kendall-Teil regression. Также математические символы принято писать курсивом: объём выборки *n*, число степеней свободы *df*, *t*-критерий Стьюдента, *F*-критерий Снедекора — Фишера и т. д. Помните, что ваша печатная работа — это ваше лицо в науке, поэтому грамотно оформляйте свои работы и следите за тем, чтобы ваши старшие наставники или редакции журналов не исправили верное написание на неверное (к сожалению, иногда такое случается).

В конце практикума представлен указатель статистических терминов и названий методов на русском и английском языках с указанием номера страницы, где они упоминаются. Поэтому быстрый поиск информации по интересующему методу можно также начать с указателя.

Структура лабораторного занятия

Каждое лабораторное занятие состоит из четырёх основных частей.

1. Введение. В нём приводится минимально необходимый для выполнения заданий набор сведений теоретического и/или практического характера, включая ***термины*** (выделены полужирным курсивом) и определения (обозначены знаком ►). Для облегчения поиска в пакетах и на интернет-ресурсах статистические термины даны также на английском языке и приведены светлым курсивом в скобках: (*term*). Вводный раздел не заменяет теоретический курс, а призван дополнить его и адаптировать к практике. Указания на необходимость обращения к теоретическому курсу приводятся в соответствующих местах пособия в скобках: (см. теоретический материал). Они подразумевают самостоятельную работу с литературой или чтение лекционного материала. Важные сведения и рекомендации предваряются словом «**ВАЖНО!**». Поскольку лабораторный курс сильно опережает лекционный, для ряда занятий введение получилось весьма объёмным. Такие лабораторные занятия рационально частично выносить на самостоятельное изучение. Небольшие

занятия, оставляющие много свободного времени, можно дополнить опросом и контрольными работами.

2. П р и м е р. Содержит данные для освоения метода и формулировку задания. Небольшие наборы данных представлены в тексте практикума и должны вноситься в статистический пакет самостоятельно. Объёмные данные представлены в виде готовых файлов; их можно найти в папке «Данные» по ссылке: https://yadi.sk/d/g50i73pt3J6pAa

3. Алгоритм расчётов. Данный раздел содержит пронумерованную последовательность действий с необходимыми скриншотами и комментариями. Для облегчения структуры практикума скриншоты не нумеровались, а ряд небольших рисунков не содержит подрисуночных подписей.

Шаги, требующие работы в программе, начинаются с соответствующей иконки:

Пакет PAST — основной пакет практикума. Бесплатный; не требует установки. URL: https://folk.uio.no/ohammer/past/

Пакет TpX — векторный графический редактор для правки рисунков. Бесплатный (лицензия GNU GPL), не требует установки.
URL: https://sourceforge.net/projects/tpx/ (исходный код)
URL: https://ctan.org/tex-archive/graphics/tpx (zip-архив с программой, скомпиллированной под ОС Windows)

Электронная таблица Excel из пакета Microsoft Office (коммерческая). Может быть заменена на Calc из пакетов Apache OpenOffice (бесплатный, лицензия Appache; https://www.openoffice.org/ru) или LibreOffice (ru.libreoffice.org, лицензия GNU GPL)

Браузер — для расчётов в онлайновых калькуляторах. Можно использовать любой современный браузер.

Пакет 3DField — для интерполяции пространственных данных и построения карт-схем (функционально ограниченная бесплатная версия). URL: http://3dfmaps.com

Пакет GPower — для расчётов объёмов выборок, мощности и др. при планировании исследования (бесплатный).
URL: http://www.gpower.hhu.de/en.html

Часть занятий содержит также расчёты по формулам, которые можно проводить на ручном или виртуальном калькуляторе.

В алгоритмической части имеются вопросы (**Вопрос:** …), над которыми нужно подумать и дать ответ *обязательно до дальнейшего прочтения текста*, поскольку далее ответ, как правило, приводится. Также имеются задания для самостоятельного выполнения (**Задание:** …).

4. Вариант оформления в квалификационной работе. Обычно приведены краткие формулировки и рекомендации для трёх разделов квалификационной работы или научной статьи: 1) «Материалы и методы»; 2) «Результаты и обсуждение»; 3) «Выводы». Представленные шаблоны не являются строгими и должны дорабатываться авторами под конкретные методы и результаты работы.

Некоторые занятия содержат «**Комментарий**» и пометку «**К сведению**». Для ряда лабораторных занятий дано обязательное к выполнению  «**Домашнее задание**».

Как подготовить файл данных для статистического анализа в рамках статьи, дипломного проекта или диссертации

► **Данные** (*data*) — сведения, полученные путём наблюдения, счёта, измерения, а также логических или арифметических операций, представленные в форме, пригодной для хранения, передачи и обработки.

Как показывает практика, несмотря на успешное прохождение курса биостатистики большинством студентов, при работе над дипломным проектом у многих из них возникают сложности, связанные с отсутствием навыков работы с большими массивами разнородных данных. В данном разделе представлен ряд рекомендаций по структурированию собираемых данных для последующего статистического анализа.

1. Сбор данных должен осуществляться строго в той шкале, к которой эти данные относятся (см. выше «типы признаков и шкалы данных»). **Недопустимо на стадии сбора материала проводить его классификацию**, тем более с понижением шкалы.

Например, если данные получаются в виде какой-то цифры (шкала отношений или интервалов), нельзя записать их как «ниже нормы», «норма» или «выше нормы» (порядковая шкала). На стадии сбора данных должно быть записано конкретное число, тогда как классификация может быть проведена значительно позже — уже при интерпретации результатов. Придерживайтесь алгоритма: 1) всё, что можно записать числом, записывается числом; 2) что нельзя записать числом, но можно упорядочить — записывается как ранг или упорядоченная категория (1–2–3, мало — средне — много и т. п.); 3) только то, что нельзя записать ни числом, ни упорядочить, фиксируется как номинальная категория (вид, цвет, форма, диагноз и т. п.).

2. Хранение данных. Данные удобнее всего создать и хранить в формате электронных книг Microsoft Office Excel, OpenOffice Calc или LibreOffice Calc. Они позволяют работать с форматами, которые поддерживают все статистические пакеты: *.xls, *.xlsx, *.csv, *.txt. Вместе с тем они привычны для большинства пользователей и предоставляют весь необходимый инструментарий для подготовки данных к анализу (сортировка, фильтрация и т. п.).

3. Количество файлов. Файлов с данными не должно быть много, идеально — один файл на один проект. Для студенческих и магистерских работ **все собранные данные должны быть размещены на одном листе электронной таблицы**. То есть контрольные и экспериментальные группы, мужчины и женщины, здоровые и больные и т. д. — все должны быть размещены на одном листе. Нет ничего страшного в том, если не все ячейки такой большой таблицы будут заполнены, главное — чтобы таблица была одна. Удобно переименовать «Лист 1» созданной таблицы в «Данные» или «Исходные данные».

4. Структура файла с данными.

4.1. Первый ряд таблицы — шапка таблицы — содержит названия колонок. **Все названия необходимо разместить в ячейках только первой строки**, при этом объединять ячейки нельзя. Такая шапка может смотреться не очень красиво, названия колонок могут частично дублироваться (например: Показатель 1–24 ч, Показатель 1–48 ч, Показатель 1–3 сут …). Но мы создаём файл не для распечатки в работу, а для хранения данных таким образом, чтобы их можно было легко передать статистическому пакету. Желательно не использовать для шапки каких-

то особенных шрифтов, но, чтобы она отличалась от остальных данных, строку шапки допустимо залить цветом.

4.2. В каждом ряду таблицы, начиная со второго ряда, располагается информация по одному **объекту** исследования (животное, пациент, образец и т. п.). Таких строк будет столько, сколько было уникальных объектов в исследовании. В колонку 1 помещаются уникальные метки, которые позволяют однозначно идентифицировать объект. Если файл данных создаётся на основе фрагмента базы данных, картотеки и т. п., поместите в первую колонку именно эти коды: в случае необходимости данные по этому объекту можно будет уточнить. В медико-биологических работах часто вместо кода помещают фамилию, имя и отчество пациента, но следует помнить, что такая информация попадает в разряд врачебной тайны и требует должного обращения. Так или иначе, но все объекты исследования должны иметь подобную уникальную метку, то есть **в первой колонке не должно быть пропусков**.

4.3. В столбцах таблицы располагается информация об **атрибутах** объектов, то есть каких-либо их характеристиках, признаках и т. п. Обычно колонку 2 называют «Группа», в неё помещается метка принадлежности объекта к какой-то группе исследования, например, экспериментальной или контрольной. Некоторые пакеты не понимают текстовый и особенно кириллический формат меток, поэтому можно сделать метки цифровыми. Например, в ячейках контрольной группы поместить цифру 1, в ячейках экспериментальной — 2. Чтобы не забыть, какую группу какой цифрой мы обозначили, на втором листе электронной таблицы, который следует переименовать в «Коды», параллельно размещается информация с расшифровкой цифровых кодов первого листа, например: «Группа 1 — контроль, 2 — опыт». Для удобства работы с файлом данных можно строки объектов разных групп выделить разным цветом.

Далее в столбцы 3 и 4 в медико-биологических работах обычно помещают информацию о возрасте и половой принадлежности объекта. **Возраст должен быть представлен одной цифрой в одинаковых для всех объектов единицах**. Если возраст учитывается с точностью до месяца, то в ячейке нельзя написать, например: «2 года 4 месяца» или «2; 4» или «2,4» (в году 12 месяцев, то есть 4 месяца это не 0,4, а 4/12 = 0,333(3)). В этом

случае под возраст нужно создать 3 колонки: «Возраст полные годы», «Возраст месяцы к годам» и итоговую колонку «Возраст», куда поместить формулу для расчёта возраста в годах по колонкам с годами и месяцами. В случае «2 года 4 месяца» по колонкам со значениями 2 и 4 здесь будет вычислено значение 2,3 (2 + 4 : 12 = 2,333). Все три колонки полезно оставить даже после таких расчётов, поскольку в официальной статистике возраст обычно считается в годах, и колонка «Возраст полные годы» далее может также понадобиться. Пол можно закодировать цифрами 1 и 2, которые нужно расшифровать на втором листе.

Заполняйте колонку «пол» непосредственно в процессе внесения данных, так как восстановить пол позже по фамилиям не всегда возможно (например, в случае украинского, французского (и т. д.) и иного происхождения фамилии — Ковальчук, Порте, Шмидт, Русских и пр.).

В последующих столбцах располагается информация о других атрибутах объектов. В версиях Excel до Office 2007 максимальное число столбцов в листе было 256, и их иногда не хватало; начиная с Office 2007 количество столбцов было увеличено до 16 385.

5. Данные в ячейках могут быть цифровыми или текстовыми. **Категорически нельзя смешивать эти типы данных**. Например, в ячейках колонки «Размер опухоли» нельзя написать «2×4×1,5» — ни одна программа не обработает такую ячейку как цифровую. В этом конкретном случае следует создать 3 столбца под каждый из линейных размеров опухоли, а в четвёртом поместить ту интегральную характеристику, которую планируется использовать (максимальный из трёх размеров или рассчитанный по формуле объём опухоли). Аналогично поступают с данными по артериальному давлению (систолическое и диастолическое), шкале Апгар (на 1-й и 5-й минутах с момента рождения ребёнка) и другим показателям, совмещающим сразу несколько цифровых значений.

5.1. В некоторых случаях кодировать цифрами текст на листе 1 и расшифровывать всё на листе 2 слишком трудоёмко, поэтому какую-то информацию можно оставить текстовой, особенно если пока нет уверенности в том, что её потребуется статистически обрабатывать. Тем не менее для текстовых переменных следите за тем, чтобы одинаковые названия были везде пропи-

саны одинаково. То есть недопустимо, чтобы, например, хронический холецистит в разных ячейках был записан как «Хр. Холецистит», «Хр. холецистит», «Хронич. холецистит» — **нужно выбрать строго одну форму записи и скопировать её во все нужные ячейки для унификации**. Это позволит при необходимости фильтровать и сортировать текстовую информацию, а также легко заменять её цифровыми кодами.

5.2. Альтернативные признаки (обычно — наличие (+) или отсутствие (–) какой-либо характеристики) кодируются как 1 и 0. Имеет смысл расписать таким образом все сложные характеристики. Например, видовой состав сообщества следует представить столбцами с названиями видов, а в соответствующих ячейках поместить цифры 0 — если данный вид не был обнаружен, 1 — если был обнаружен. Аналогично в медицинских исследованиях как 0 или 1 кодируются различные сопутствующие заболевания и осложнения.

Если проводился количественный учёт видов, то вместо цифры 1 следует указать конкретные численности, например 28, 100, 528. В микробиологических работах численности организмов обычно велики: 10^5, 10^6, 10^7 и т. п. Для таких случаев можно указывать в ячейках только степени: 5, 6, 7 (то есть десятичные логарифмы численностей), однако на листе «Коды» об этом следует оставить напоминание.

5.3. **Пустые ячейки — это не ноли!** Пустые ячейки означают, что данных для этой ячейки нет и она не будет участвовать в расчётах. Соответственно, везде, где регистрация признака проводилась, но интересующей характеристики обнаружено не было, следует написать 0, а пустыми оставить только те ячейки, для которых данные по какой-то причине не были собраны. По окончании заполнения таблицы станет видно, для каких объектов и атрибутов есть пропуски и какова их доля. Решение по таким случаям следует принимать вместе с научным руководителем или специалистом по анализу данных, поскольку в разных случаях и для разных задач возможны все три варианта действий: 1) удалить строку или столбец с большим числом пропусков целиком; 2) оставить с пропусками и далее так и обсчитывать; 3) заменить малочисленные пропуски медианой или специально подобранными с помощью *техник* ***множественной импутации*** (*multiple imputation*) значениями.

5.4. ***Цензурированные данные*** *(censored data)* — особый и неудобный тип данных, сочетающий количественные и качественные характеристики. Цензурированные данные типа «более чем» типичны для исследований в области медицины и некоторых областях экспериментальной биологии, где они образуют особое направление — ***анализ выживаемости*** (см. лабораторное занятие № 14). Цензурированные данные типа «менее чем» появляются вследствие ограниченной разрешающей способности аналитических методов, когда часть значений измеряемого показателя оказывается ниже границы чувствительности метода. На стадии заполнения таблицы их можно вносить как текстовые, например «<0,02». Но для статистической обработки их потребуется заменить цифрой. Обычно это 0 (ноль), однако в ходе работы специфическими для таких данных методами эти значения не будут учтены как простые нули. Иногда, такие значения вносят как половину чувствительности метода, то есть не «<0,02», а количественное значение «0,01», однако это не вполне корректно. Способы анализа таких данных в пособии не рассматриваются, поэтому здесь также рекомендуем консультироваться с научным руководителем или биостатистиком.

По окончании внесения данных мы должны иметь один файл в формате *.xls или *.xlsx из двух листов: «Данные» и «Коды».

6. Передача файла статистическому пакету и верификация данных. Созданный файл нужно попытаться открыть из статистического пакета, в котором предполагается провести бóльшую часть обработки данных. Любой пакет ищет по умолчанию именно свой тип данных, поэтому нужно выбрать соответствующее файлу данных расширение (*.xls, *.xlsx), в качестве листа указать Лист 1 — «Данные» и отметить опцию, чтобы названия переменных были взяты из первой строки. Если статпакет не поддерживает открытие файлов электронных таблиц, можно вставить в него предварительно скопированный в буфер обмена Лист 1. Часто сразу после этого статистический пакет начинает выдавать какие-либо предупреждающие сообщения. **Не закрывайте их сразу, а внимательно читайте**, **английский текст — переводите**: помните, что наша цель — не побыстрее открыть данные в статпакете, наша цель — верифицировать данные, и этап их портирования в другой пакет является первой стадией такой проверки. Типичными предупреждающими сообщениями

являются сообщения об отсутствии данных в каких-то ячейках, о наличии текстовой информации в ячейках, о каких-либо несоответствиях, например, конфликте десятичных разделителей (некоторые пакеты «не понимают» русского стиля с запятой в качестве десятичного разделителя, а не точки). Убедитесь, что всё, что «не понравилось» статпакету, не содержит ошибок. Если же ошибки были обнаружены — устраните их **обязательно сразу же, не откладывая**, также в исходном файле с данными.

Открытый в статпакете файл следует сохранить в формате этого пакета и провести 1–2 анализа: получить описательную статистику (обязательно с минимумом и максимумом) и/или посмотреть гистограммы распределения. В данном практикуме эти задачи рассматриваются, соответственно, на лабораторных занятиях № 2 и 4. Это необходимо сделать для того, чтобы исключить грубые ошибки набора данных. Например, максимальный возраст 115, это, скорее всего, лишняя цифра при наборе 11 или 15 лет, масса 875 кг — пропущенная запятая в 87,5 и т. п. Гистограммы распределения также помогают обнаруживать выбросы и ещё дают представление об однородности выборки и форме распределения.

7. Данные нельзя потерять! Созданный и верифицированный файл данных сохраните в нескольких местах (в компьютере, на флеш-карте, отправьте себе на электронную почту и т. д.). **Держите исходный файл неизменным**. Для работы сохраните файл под другим именем, например, «Диплом для анализа» или «Диплом_26.03.2018». Это позволит не потерять исходные данные в непредвиденных ситуациях.

Как получить помощь по статистическому анализу данных

Как уже было сказано, в настоящем пособии не представлен ряд направлений анализа данных, а решение подавляющего большинства задач продемонстрировано в одном из множества пакетов — пакете PAST. Другие направления, методы и пакеты потребуют самостоятельного освоения. Кроме того, бывают случаи, к статистическому анализу которых можно подойти

с использованием разных техник, а требуется выбрать: 1) наиболее эффективную в плане достижения цели; 2) наиболее мощную в статистическом смысле этого термина; 3) достаточно традиционную в конкретной области науки. Поэтому рано или поздно любой исследователь оказывается в ситуации, когда осознаёт ограниченность своих знаний и необходимость применения новых методов, которыми не владеет он сам и коллеги в его ближайшем окружении. Приведём несколько практических советов, как рационально действовать в такой ситуации.

1. «Я — биолог (я — врач), а не статистик». Такая не всегда верная установка тиражируется в некоторых медицинских, педагогических и даже научных коллективах, а потому встречается не так уж редко. Она справедлива до тех пор, пока человек не приступает к выполнению научной квалификационной работы. **Статистический анализ данных является неотъемлемой частью современной научной методологии**. Поэтому если человек работает над школьным научным проектом, дипломной работой бакалавра, магистерской, кандидатской или докторской диссертацией, он должен предъявить соответствующие данному квалификационному уровню умения грамотно получать данные и выделять из них наиболее существенные закономерности с использованием статистических методов. Следовательно, пока вы занимаетесь научной работой, вы — статистик.

2. Начните поиск информации самостоятельно. К настоящему времени русскоязычный сегмент Интернета существенно наполнился биостатистической информацией, поэтому формулируйте запросы в поисковике и просматривайте первые 3–5 страниц. Ищите такие материалы, которые изложены понятным для вашего уровня языком, и такие статьи, где всё описано логично, хорошо прописан анализ данных и указаны программы для статистических расчётов.

3. Если вы попадаете на форум — сначала ознакомьтесь с имеющимися на нём материалами, воспользовавшись поиском по форуму. **Категорически не рекомендуется начинать сразу задавать вопросы**: примитивные и одинаковые вопросы новичков у опытных участников форумов вызывают только раздражение, а те, кто поспешит вам ответить, скорее всего знают немногим больше вашего. Хорошие биостатистические форумы Рунета:

– http://forum.disser.ru, раздел «Медицинская статистика»;
– http://molbiol.ru/forums, раздел «Биофизика и матметоды в биологии».

Пользуясь случаем, автор выражает свою благодарность А. Г. Виноградову, И. П. Гайдышеву, Е. И. Драгомирецкой, С. В. Петрову, С. Л. Плавинскому, Ю. А. Тукачёву, А. Б. Шипунову и другим создателям и активным консультантам статистических форумов за возможность развиваться в русскоязычной (био)статистической среде.

4. Ознакомившись с новой информацией, попробуйте выполнить расчёты самостоятельно. Если при этом что-то не получается, убедитесь, что вы ознакомились с «помощью» к пакету (разделы «Помощь» или «Help», описания, инструкции, manual). Многие просто забывают про этот источник информации, а именно он обычно и содержит ответы технического характера. Для самых популярных статистических пакетов также имеется специализированная литература, которую можно найти в сети.

5. Если, проделав шаги 2–4, вы всё ещё не справились с проблемой, сомневаетесь в полученном результате или его интерпретации — обращайтесь за помощью к специалистам в вашем окружении или на форумах. К этому моменту вы уже будете достаточно осведомлены об интересующем предмете, поэтому сможете грамотно и лаконично изложить суть вопроса. При этом нужно кратко описать шаги, которые вы уже предприняли, — это покажет собеседникам, что они имеют дело не с лентяем, который хочет, чтобы его задачу решили за него, а с младшим коллегой, которому действительно нужна помощь.

6. Если на каком-то этапе вам потребуется платная помощь биостатистиков, старайтесь находить такие формы взаимодействия со специалистами, когда вас научат чему-то новому, а не просто сделают вашу работу.

Желаем успехов в вашем дальнейшем профессиональном росте!

ЛАБОРАТОРНАЯ РАБОТА № 1

Знакомство со статистическим пакетом для ПК на примере PAST

Тема 5. Описательная статистика.
Количество часов: 2.
Цель: познакомиться с интерфейсом и основными приёмами работы с данными в бесплатном статистическом пакете для ПК. Работа на ПК.

Статистический анализ данных можно проводить в пакетах трёх типов: электронных таблицах, математических пакетах и статистических пакетах.

I. Процессоры электронных таблиц (электронные таблицы)

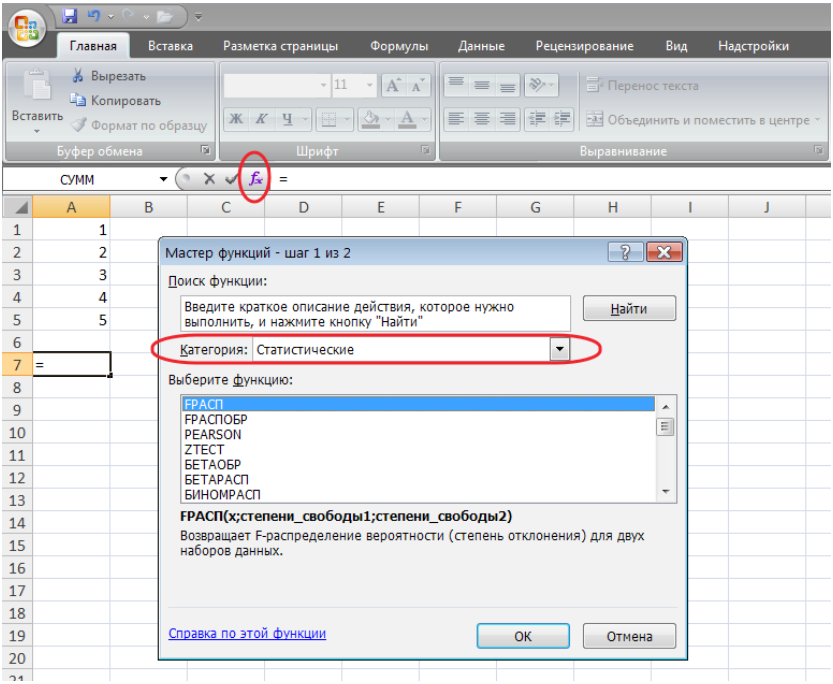
► **Электронные таблицы** — это программы, позволяющие проводить операции и вычисления с данными, представленными в виде двумерных массивов, имитирующих бумажные таблицы. Исторически первой такой программой была VisiCalc Дэниела Бриклина, написанная в 1978 г. на ассемблере для Apple-2. В настоящее время наиболее распространены и известны следующие пакеты:

1) для ОС Windows: Microsoft Office Excel;
2) для ОС Linux: GNOME Office Gnumeric, KOffice KSpread;
3) кроссплатформенные: Apache OpenOffice Calc и LibreOffice Calc.

Рабочее пространство таких программ представляет собой развёрнутый ***разлинованный лист*** (*spreadsheet*), в ячейки которого можно вводить различные данные и формулы для организации автоматических вычислений. Пакеты имеют обширный раздел статистических формул и логических операций для программирования расчётов.

В пакете Excel

① Запустите MS Excel, вбейте в первую колонку пять произвольных значений (например, 1, 2, 3, 4, 5), а в произвольной ячейке — знак равенства и нажмите на значок функции f_x. Посмотрите, какие есть категории функций и выберите «Статистические».



Первые две функции задействуют *F*-распределение Снедекора — Фишера, далее следуют: показатель корреляции Пирсона, *z*-критерий (площадь под кривой стандартного нормального распределения), функции с бета-распределением, биномиальным распределением и т. д.

② Пролистайте список далее, найдите и прочитайте описание для функций МАКС, МЕДИАНА, МИН, СРЗНАЧ.

③ Выберите СРЗНАЧ, оттащите мышью форму правее, чтобы она не загораживала данные, выделите введённую область значений и нажмите ОК. В выбранной ячейке появится среднее значение (для нашего примера — число 3).

Вы увидели в списке большое количество функций и распределений, с помощью комбинаций которых можно задать подавляющее число известных статистических методов для анализа данных. У такого подхода к анализу есть свои плюсы и минусы.

Плюсы: понимание принципа того, как считает программируемый статистический метод.

Минусы:

1) отсутствие умения грамотно перенести формулы из статистической литературы в расчётный блок электронной таблицы;

2) затраты времени на устранение ошибок и отладку расчётного блока. Даже известные коммерческие пакеты содержат ошибки и недоработки, что уж говорить о качестве программы обычного пользователя ПК! Поэтому, если в научной статье в разделе «Материалы и методы» вы встречаете фразу «Статистический анализ данных выполнен стандартными методами вариационной статистики в пакете MS Excel», вы имеете полное право не доверять автору. Скорее всего, он не обладает требуемой квалификацией и либо использовал для расчётов нелицензионный «пиратский» софт, либо вообще не знает, как были проанализированы кем-то его данные (к сожалению, такое встречается);

3) статистические библиотеки даже самых известных процессоров электронных таблиц вроде Excel считают неточно. Неточность проистекает от недостаточной точности вычислений и несовершенства расчётных алгоритмов. Это известно давно, и попытки исправить ситуацию предпринимаются; однако по мере устранения старых ошибок в новых версиях пакетов появляются новые ошибки. Специалистами в области статистического программирования разработаны специальные тесты на правильность статистических вычислений. Наиболее известен тестовый набор Statistics Quiz, называемый также ***тестом Вилкинсона*** (*Wilkinson's test*). С ним в разной степени не справляются почти все статистические пакеты, включая известные коммерческие разработки, а тем более процессоры электронных таблиц. Поэтому даже если статистические расчёты в пакетах типа Excel или Calc запрограммированы правильно, их использование в качестве статистических пакетов указывает на некий непрофессионализм, поскольку не гарантирует правильности сложных расчётов и построенных на их основе выводов.

Исключение составляют ***программы-надстройки*** (*add-on*), большинство из которых написано для MS Excel. Такие пакеты представляют собой самостоятельные программы, использующие лишь интерфейс электронных таблиц, поэтому их качество должно оцениваться самостоятельно. Наиболее известны: XLSTAT, PopTool, AtteStat.

В целом процессоры электронных таблиц представляют собой удобную среду для ввода, хранения и операций с данными, а также для автоматизации вычислений, создания деловой и несложной научной графики и др. Однако они не могут быть использованы в качестве полноценного самостоятельного средства для статистического анализа данных. По ходу нашего курса мы будем использовать электронные таблицы Excel, но лишь в качестве дополнения к статистическому пакету.

II. Математические пакеты

Наиболее мощными математическими пакетами являются ***системы компьютерной алгебры*** (СКА). ►**СКА** — программные комплексы для символьных вычислений. Они служат для работы с математическими выражениями в аналитической (символьной) форме. Первой успешной СКА была разработка голландского физика, нобелевского лауреата по физике 1999 г. Мартинуса Велтмана, который в 1963 г. создал программу Schoonschip для символьных вычислений в области физики высоких энергий. Современные СКА позволяют проводить весь цикл разработки математической модели: от поиска и просмотра необходимой литературы до численного или аналитического решения задачи и подготовки отчёта, публикации. Наиболее известные и распространённые СКА: Maple, Mathematica, MATLAB, Mathcad.

Такие программные комплексы могут: упрощать сложные математические выражения, разлагать их на множители, дифференцировать и интегрировать функции, проводить операции с матрицами и многое другое, включая даже автоматическое доказательство теорем. Естественно, они позволяют программировать любые статистические функции и проводить по ним расчёты с высокой точностью. До появления ***программно-статистической среды R*** ряд передовых (*advanced*) статистических техник был доступен пользователям исключительно в виде программного кода к СКА.

Работа в СКА требует высокого уровня математической подготовки, а также навыков работы в подобных системах, включая владение языками программирования. Естественно, что они не подходят для подавляющего большинства биологов и врачей, вследствие больших затрат времени на сложные непрофильные работы в области математики и программирования, которые к тому же не застрахованы от ошибок.

III. Статистические пакеты

► **Статистический пакет** — программный продукт, предназначенный для статистической обработки данных. Рассмотрим классификацию таких пакетов.

1. По назначению: универсальные или специализированные. ***Универсальные пакеты*** позиционируются разработчиками как средства для анализа данных в самом широком диапазоне научных исследований.

Плюсы:

а) обычно — многоплатформенность (под Windows, Linux, Mac);

б) относительно высокое качество алгоритмов;

в) в целом стандартный интерфейс: пользователь, знакомый с одним таким пакетом, без большого труда найдёт нужные опции и методы в другом универсальном пакете.

Минусы:

а) большой размер;

б) сложность интерфейса. Попытка угодить исследователям разных направлений приводит к увеличению размеров пакета, а также усложнению интерфейса: в них действительно содержится много методов, но нужно уметь найти их в пакете и выбрать из перечня оптимальный;

в) как правило, универсальные пакеты — коммерческие.

Специализированные пакеты нацелены на решение узкого диапазона вычислительных задач и/или небольшую область науки. Как правило, они имеют небольшой размер, зачастую — нестандартный интерфейс, но содержат методы, отсутствующие в больших универсальных пакетах.

2. По типу интерфейса: с графическим интерфейсом или с текстовым интерфейсом. Большинство статистических пакетов имеют кнопочный ***графический интерфейс***. Это удобно для рядовых пользователей, но одновременно является ограничени-

ем для профессионалов, поскольку здесь имеется возможность лишь следовать алгоритмам разработчиков в ущерб скорости и гибкости. Профессионалы используют пакеты с ***консольным текстовым интерфейсом***, что позволяет быстро оперировать данными и проводить расчёты, вводя определённые команды с клавиатуры. Однако для этого необходимо знать соответствующий программный язык.

Наиболее популярные пакеты сочетают графический и текстовый интерфейс. Такие пакеты либо представляют собой среду для вычислений на определённом программном языке, поверх которого надстраивается кнопочный графический интерфейс, либо изначально имеют графический интерфейс, но с возможностью программировать команды на специальном ***сценарном языке*** (*scripting language*) — с помощью ***скриптов***. Это даёт возможность работать в пакете как начинающим пользователям — через систему меню, так и продвинутым пользователям — путём непосредственного набора команд в специальном окне программы.

3. По цене для пользователя: платные или бесплатные. Независимо от рассмотренных выше классификаций пакеты могут быть платными или бесплатными.

Как правило, платные коммерческие пакеты до покупки лицензии работают в режиме демо-версии, демонстрирующей возможности пакета. Такая демо-версия является полноценным продуктом, но содержит обратимые ограничения: либо функциональные — по спектру доступных методов, либо временны́е — по времени использования (обычно 30 дней), которые после покупки лицензии снимаются. Платный пакет может иметь модульную структуру, когда наряду с базовым пакетом продаются отдельные специализированные модули.

Бесплатные пакеты очень разнообразны. Это могут быть «младшие» версии коммерческих разработок — академические версии для некоммерческого использования или старые версии флагманского пакета. Такие пакеты имеют закрытый программный код, а лицензия позволяет только ограниченно использовать пакет. Также это могут быть написанные энтузиастами программы или расчётные блоки онлайновых калькуляторов без каких-либо ограничений или лицензий. Однако чаще всего бесплатные пакеты поставляются с так называемой GPL-лицензией.

► GNU **G**eneral **P**ublic **L**icense — ***универсальная общедоступная***

лицензия на свободное программное обеспечение с сохранением авторских прав. Она позволяет использовать программу в любых целях, модифицировать её, свободно распространять копии и модификации, но запрещает включать программу в частные коммерческие разработки.

Перечислим самые популярные в России статистические пакеты:

STATISTICA — универсальный пакет, имеющий разный набор модулей в зависимости от комплекта поставки. Интерфейс — графический, начиная с версии 6.1 — русскоязычный; установлен поверх языка STATISTICA Visual Basic. Пакет — коммерческий, с сильно функционально урезанной демо-версией. Популярность пакету обеспечили широкодоступные пиратские копии во время массового перехода пользователей с операционной системы DOS на Windows в середине 1990-х гг.: какое-то время это был самый доступный в России статистический пакет под Windows с большим набором методов и отличной графикой.


SPSS (**S**tatistical **P**ackage for the **S**ocial **S**ciences) — универсальный пакет с долгой историей (с 1968 г.) и модульной структурой. Интерфейс — графический, начиная с версии 18 — русскоязычный. Написан на Java. Встроенный язык позволяет гибко работать с данными и писать самостоятельные программы-макросы. Пакет — коммерческий, с полноценной демо-версией.

SAS/STAT (**S**tatistical **A**nalysis **S**ystem) — универсальный пакет с долгой историей (с 1976 г.). Мощный коммерческий продукт с текстовым интерфейсом, написанный на SAS programming language. До появления R (см. далее) был почти единственным инструментом для отечественных статистиков-профессионалов (популярные за рубежом коммерческие пакеты STATA и Origin не распространены в России).


MedCalc — универсальный пакет, с акцентом на анализ данных в области медицины. Интерфейс — графический, начиная с версии 15.2 — русскоязычный. Пакет — коммерческий с полноценной демо-версией.

 R — программно-статистическая среда с интерфейсом командной строки и открытым исходным кодом (лицензия GNU GPL). В доступном онлайн-депозитарии CRAN (**C**omprehensive **R** **A**rchive **N**etwork) находятся дистрибутивы пакета, а также многочисленные (тысячи) пакеты, создаваемые энтузиастами и профессионалами со всего мира. Есть ряд весьма успешных реализаций графического интерфейса к R, что облегчает использование среды новичками, но лишает гибкости. В настоящее время среда R становится стандартом в области математической статистики и анализа данных.

* * *

При выборе статистического пакета для данного лабораторного курса мы руководствовались его бесплатностью и лёгкостью в освоении для практического использования при написании курсовых и дипломных работ. Среда R помимо знаний в области статистики требует неплохого знания английского языка и навыков программирования. Поэтому для короткого начального курса был выбран бесплатный пакет с графическим интерфейсом — PAST. Полагаем, что его освоение послужит хорошей базой для дальнейшего профессионального роста.

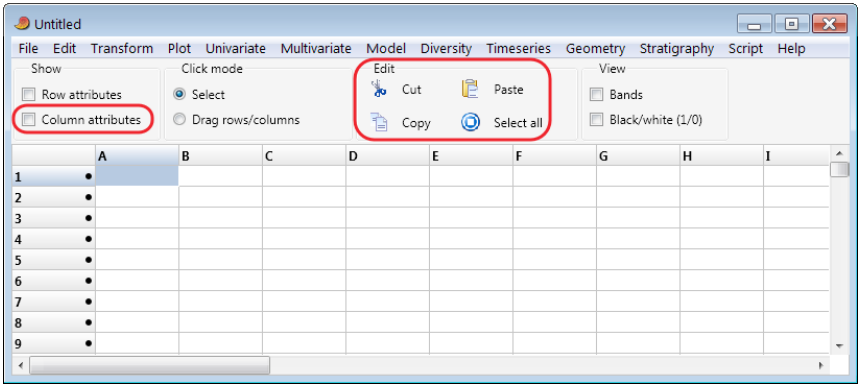
 PAST (**Pa**leontological **St**atistics) — универсальный пакет с акцентом на анализ данных в палеонтологии. Разработчиками пакета являются учёные из трёх европейских университетов: Эйвинд Хаммер из Палеонтологического музея университета Осло (Норвегия), Дэвид Харпер из Геологического музея университета Копенгагена (Дания) и Пол Райан из Геологического департамента Национального университета Ирландии (Ирландия). Помимо основных статистических методов в пакете представлены методы для морфометрического анализа размеров и формы, а также для анализа сообществ организмов, несложных генетических расчётов и др. Это делает его полезным инструментом для биологов (особенно экологов) и медиков. Особенностью пакета является внедрение в большинство модулей современных ресэмплинг-техник (рандомизационная техника Монте-Карло, бутстреп, точные перестановочные критерии), отсутствующих даже в популярных коммерческих пакетах. К недостаткам пакета можно отнести англоязычный интерфейс и довольно бедные графические возмож-

ности; однако по ходу курса мы переведём основные термины и научимся дорабатывать графики до совершенства во внешнем графическом редакторе.

В пакете PAST

① Запустите программу. Вы видите, что интерфейс пакета представляет собой разлинованный лист, похожий на лист Excel.

② Кратко познакомимся с меню и отметим наиболее важные для предстоящего практического курса разделы.



File — файл. Стандартное меню. В данном модуле собраны основные операции с файлами. Также здесь можно узнать данные о программе, включая номер версии, адрес сайта, имена разработчиков и правила оформления ссылки на программу в публикациях.

Edit — правка. Стандартное меню. Полезные элементы:

Undo — откатиться назад, Redo — вернуться вперёд.

Cut — вырезать, Copy — копировать, Paste — вставить, Select all — выделить всё (эти команды вынесены также в отдельное графическое подменю Edit).

Find — найти, Replace — заменить.

Rearrange (изменение порядка) ▶

– Transpose — транспонировать (поменять строки и колонки местами);

– Observations to contingency table — наблюдения в таблицу сопряжённости;

– Value pairs to matrix — пары значений в матрицу.

Transform — преобразовать. Преобразования данных. Полезно:
Log — логарифм десятичный;
Box-Cox — преобразование Бокса — Кокса.

Plot — график. Будем активно использовать этот модуль для построения графиков.

Univariate — одномерные методы. В модуле собраны наиболее распространённые статистические методы для описания данных, выборочных сравнений и поиска связей. Будем активно использовать этот модуль.

Multivariate — многомерные методы, включая проекционные техники (главные компоненты, анализ соответствий и др.), кластерный анализ, индексы сходства и расстояний.

Model — модель. В модуле представлен широкий выбор регрессионных техник для поиска зависимостей и сглаживания рядов данных.

Diversity — разнообразие. Методы для анализа видового богатства и биоразнообразия.

Timeseries — временны́е ряды. Специфические методы работы с рядами динамики.

Geometry — геометрия. Модуль анализа формы и размеров с использованием техник геометрической морфометрии.

Stratigraphy — стратиграфия. Раздел со специальными палеонтологическими методами ***биостратиграфии*** — анализа распределения ископаемых объектов по геологическим пластам.

Script — скрипт. Окно написания скриптов.

Help — помощь. Загружает с сайта проекта руководство к текущей версии пакета в формате *.pdf.

③ Создадим небольшой файл данных. Для этого в подменю Show (Показать) поставим галочку в ☐ Column attributes (Атрибуты колонки) и вместо буквы A в строке Name (Имя) введём название колонки — Длина стопы. Это будут данные по длине стопы 49 восточноевропейских полёвок *Microtus rossiaemeridionalis* (Ognev, 1926) из первого поколения лабораторной колонии. Далее снимаем галочку в ☑ Column attributes. Если название не уместилось в ширину колонки — раздвигаем колонки, как в Excel: подводим указатель мыши к границе колонок «Длина стопы» и «B» (указатель меняет вид на двустороннюю стрелку) и, удерживая правую кнопку мыши, раздвигаем колонку.

④ Вводим данные. Привычка работать на ноутбуке замедляет процесс ввода данных. Имея полноразмерную клавиатуру, полезно научиться вводу данных в её правом цифровом блоке (рис. 1.1).

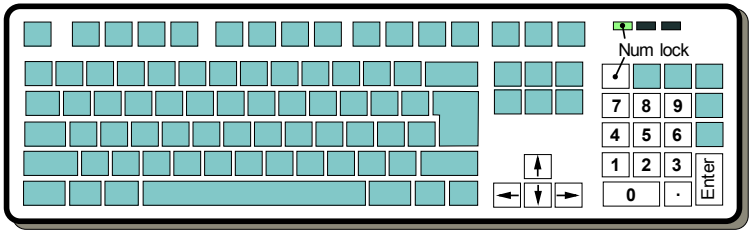


Рис. 1.1. Полноразмерная клавиатура с блоком ввода данных

На клавиатуре должен светиться индикатор Num Lock. Пальцами правой руки набиваем данные. Кнопка десятичного разделителя [.] находится здесь же. В Excel вносить данные удобнее, поскольку ввод значения осуществляется расположенной с краю клавишей [Enter], на которую можно нажимать также пальцем правой руки. В PAST ввод значения осуществляется кнопкой [↓], нажимать на которую удобно левой рукой.

Введите в строчках 1 и 2 число 15.

Далее со строки 3 по 9 — число 15,5.

С 10 по 26 — число 16.

С 27 по 31 — число 16,5.

С 32 по 43 — число 17.

С 44 по 48 — число 17,5.

В последнюю строчку 49 вносим число 18.

⑤ Сохраняем файл. Как и во всех пакетах, при первом сохранении файла ему нужно дать название, а далее можно сохранять изменения, используя комбинацию клавиш [Ctrl + S].

Путь: File — Save as… — находим нужную папку — называем файл «Длина стопы» — [OK].

В верхнем левом углу рядом с пиктограммой пакета название Untitled (Неназванный) изменится на **Длина стопы.dat**.

ЛАБОРАТОРНАЯ РАБОТА № 2

Описательная статистика

Тема 5. Описательная статистика.
Количество часов: 2.
Цель: освоить расчёт показателей описательной статистики в статпакете. Научиться грамотно округлять данные и представлять их в табличном виде. Работа на ПК.

В некоторых областях науки, при работе с крайне редкими или уникальными случаями, считается допустимым приводить все собранные данные целиком, без обобщения. В качестве примера можно привести такое редкое генетическое заболевание, как детская прогерия (синдром ускоренного старения). В каждый момент времени на планете одновременно проживает лишь несколько индивидов с таким заболеванием, и поэтому вполне уместно исследование только одного — единственного случая, а также публикация его результатов. В эту же категорию попадают уникальные или редкие виды животных и растений. Но эти примеры — исключение из общего правила. В подавляющем большинстве исследований размер ***генеральной совокупности*** (*population*) очень велик, ***выборки*** (*samples*) из неё также не слишком малы, а потому нет никакой возможности публикации в работах развёрнутых индивидуальных данных. Более того, компактное описание данных позволяет выделить в них ряд закономерностей, которые не видны в наборе цифр ***исходных данных*** (*raw data*). Для компактного описания и обобщения данных используются методы ***описательной статистики*** (*descriptive statistics, summary statistics*).

Данные можно кратко охарактеризовать с использованием трёх групп мер: мер положения, мер рассеяния и мер формы распределения (см. теоретический материал). В публикациях и квалификационных научных работах обычно используют первые две группы мер:

1. ***Меры положения***, или ***меры оценки центральной тенденции*** показывают положение центра, вокруг которого группируются данные. В качестве таковых для количественных признаков используют среднее значение и медиану, а для качественных номинальных — частоту.

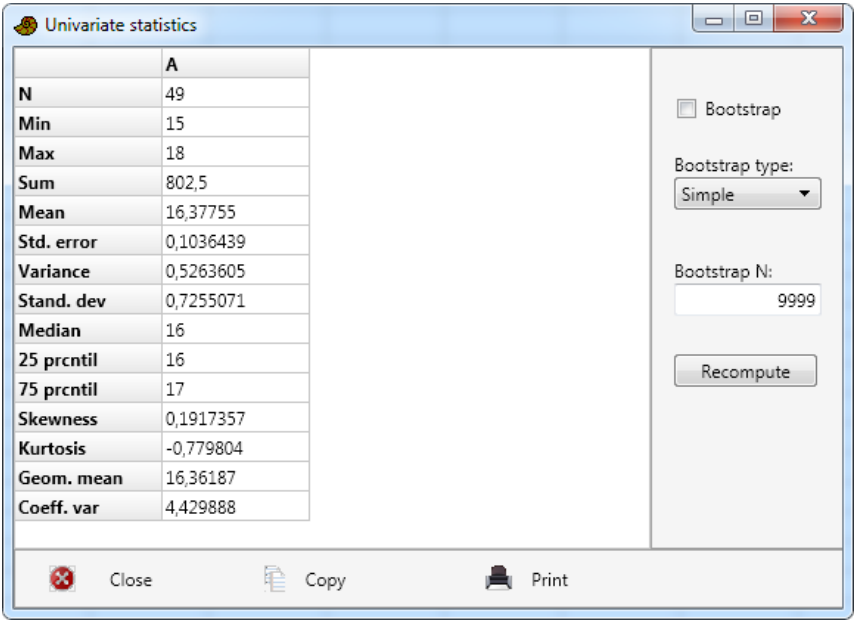
2. ***Меры рассеяния***, или ***меры масштаба*** показывают разброс значений относительно центра. В качестве них используются: стандартное отклонение, размах, межква́ртильный размах. Раньше вместо меры рассеяния часто приводилась стандартная ошибка среднего, а в настоящее время обычно приводится интервальная оценка среднего значения или частоты — ***доверительный интервал*** *(ДИ)*, как правило, 95%-ный ДИ.

Существуют стандарты представления мер описательной статистики. Рассмотрим их для трёх разных типов данных: I — количественных признаков с приблизительно нормальным распределением, II — количественных признаков с ненормальным распределением и порядковых признаков, III — качественных признаков.

В пакете PAST

① Откройте файл «Длина стопы»: File — Open — найти свой файл.

② Выделите область значений: мышью или стрелками на клавиатуре при удерживании нажатой клавиши [Shift]. Также можно кликнуть на название колонки, чтобы выделить её целиком.



③ Путь: Univariate — Summary statistics.

Выпишите в тетрадь в столбик английские названия, а рядом запишите перевод:

N	количество наблюдений (объём выборки)
Min (Minimum)	минимум
Max (Maximum)	максимум
Sum	сумма
Mean	среднее арифметическое
Std. Error (Standard error)	стандартная ошибка
Variance	дисперсия
Stand. dev (Standard deviation)	стандартное отклонение
Median	медиана
25 prctil (Percentile)	25-й проце́нтиль, или нижняя ква́ртиль
75 prctil	75-й процентиль, или верхняя ква́ртиль
Skewness	асимметрия
Kurtosis	эксцесс
Geom. mean (Geometric mean)	среднее геометрическое
Coeff. var (Coefficient of variation)	коэффициент вариации

> **Комментарий**. В статистической литературе советского периода преобладало употребление термина «квартиль» в женском роде, иногда — с ударением на последний слог. В настоящее время термин часто используется в мужском роде (нижний или верхний квартиль). В пособии используется вариант написания согласно словарям А. М. Микиша и В. Б. Орлова (1989), а также «Энциклопедии статистических терминов» (2011).

I. Количественные признаки с нормальным распределением

Сведения о нормальности распределения показателя в генеральной совокупности (популяции) берутся из литературы, предыдущих исследований или непосредственно из результатов проверки данных на нормальность, если это позволяет объём выборки (30 и более наблюдений). Приблизительно нормально распределённые данные можно получить с помощью подходящих ***нормализующих преобразований*** (логарифмирование,

преобразование Бокса — Кокса, угловые преобразования для частот и др. — см. теоретический материал).

СТАНДАРТ 1. Среднее ± стандартная ошибка среднего

Знак «±» читается как «плюс-минус».

В символьной форме: $\bar{x} \pm m$ («икс среднее плюс-минус эм»).

Англ. Mean ± Standard error of mean.

В публикациях встречаются обозначения: M±m, Mean ± SE, Mean ± SEM.

► **Среднее (арифметическое)** — мера положения, или центральная тенденция набора данных. Рассчитывается суммированием всех значений в наборе данных и делением суммы на объём выборки: $\bar{x} = \sum x_i / n$. Выборочное среднее является несмещённой оценкой математического ожидаемого μ генеральной совокупности.

► **Стандартная ошибка среднего значения выборки** — теоретическое стандартное отклонение всех средних значений выборок размера *n*, извлекаемых из генеральной совокупности. Рассчитывается как отношение выборочного стандартного отклонения к квадратному корню из объёма выборки: $m = s / \sqrt{n}$.

Почти весь XX в. этот стандарт доминировал в науке, но в последнее время считается устаревшим. Центральные зарубежные журналы с высоким импакт-фактором не принимают статьи с такими данными, рекомендуя использовать среднее и 95%-ный доверительный интервал (95% ДИ) для него [2; 12]. Причины отказа от стандарта:

1) знак «±» подразумевает симметричное нормальное распределение, а исследователи далеко не всегда могут быть уверены в этом или проверяют данное требование;

2) стандартную ошибку, в отличие от 95% ДИ, неудобно использовать для визуальной оценки степени различий средних между выборками и оценки значимости этих различий;

3) не все учёные принимают подход с оценкой значимости по показателю *P* (*p-value*). Для некоторых из них сопоставление 95% ДИ в выборках является альтернативой расчёту *P*.

Тем не менее в некоторых областях науки данный стандарт до сих пор используется и его можно встретить в современных научных журналах.

В пакете PAST

① Выписываем Mean ± Std. Error: 16,37755 ± 0,1036439.

② Результат нужно правильно округлить. Есть несколько вариантов округления:

1) как принято в данной области науки (см. публикации в центральных журналах — там за этим следят рецензенты);

2) среднее привести на знак точнее точности измерений, ошибку — ещё на знак точнее. Например, если рост человека измеряется с точностью до сантиметра, среднее можем привести с точностью до десятых;

3) учесть вариабельность признака и объём выборки: результат для жёстких признаков и/или для больших выборок приводить точнее, а для пластичных признаков и/или для малых выборок — менее точно. В качестве меры вариабельности и объёма выборки используется непосредственно значение стандартной ошибки, формула которой включает как стандартное отклонение, так и объём выборки (см. выше).

Алгоритм из учебника Сокала и Рольфа [21]:

а) разделить стандартную ошибку на 3;

б) определить место после запятой первой цифры, не равной нулю;

в) с такой точностью округлить среднее, а ошибку привести на знак точнее;

г) если при делении стандартной ошибки на 3 получается число больше 1, то среднее нужно округлить до целых, стандартную ошибку — до десятых. Этого пункта в оригинале не было, но на практике такие ситуации встречаются.

В нашем случае:

а) 0,1036439/3=0,0**3**454;

б) первая цифра, не равная нулю, — 3; подчеркните её;

в) она стоит на месте сотых, поэтому округляем среднее до сотых, стандартную ошибку — до тысячных. Обведите рамкой окончательный результат:

16,38 ± 0,104

> **К сведению.** Чтобы правильно округлить число, нужно оставить требуемое количество цифр после запятой и смотреть только на следующую: если она ≥5, то предыдущая цифра округляется в большую сторону, если <5 — отбрасывается.

Данный алгоритм не следует использовать фанатично. Если у вас много выборок, достаточно проверить 1–2, но в таблице результатов значения для всех выборок обязательно приводить с одинаковой точностью. **ВАЖНО!** Если для данного показателя мы решили округлять значения, например, до десятых, то с такой точностью должны быть представлены все однотипные данные, например: не 12,4; 13,22; 18, а 12,4; 13,2; 18,0. Внимательно следите за этим, иначе у грамотного читателя вашей работы сложится впечатление, что вы получили данные по одному показателю с разной точностью для разных групп (что вызывает лишние вопросы) или неаккуратны в оформлении (а может быть, и всей своей работе вообще).

СТАНДАРТ 2. Среднее; стандартное отклонение

В символьной форме: $\overline{x}$; s.

Англ. Mean; Standard deviation.

В публикациях может обозначаться немного иначе: *Mean*; *SD, M*; *s. d.*, приводиться со скобками $\overline{x}$ (s) или в соседних столбцах таблицы описательной статистики для набора показателей. Часто даже в зарубежных статьях приводится через знак «±»: $\overline{x} \pm s$, но лучшими учебниками по биостатистике такая форма записи не рекомендуется.

► **Стандартное**, или **среднеквадратическое отклонение** — мера изменчивости, или дисперсии набора данных, представляющая собой среднее расстояние отдельных наблюдений от среднего значения выборки. Рассчитывается как положительное значение квадратного корня из дисперсии выборки:

$$s = \sqrt{s^2}\ .$$

Является смещённой оценкой стандартного отклонения σ генеральной совокупности.

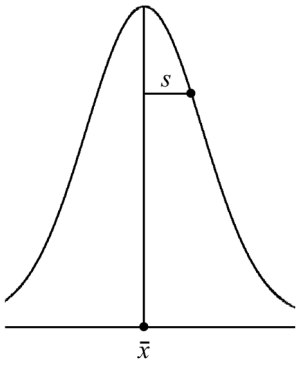


Рис. 2.1. Кривая нормального распределения с двумя его параметрами

В случае нормального распределения стандартное отклонение имеет геометрическую интерпретацию: это расстояние от перпендикуляра, опущенного с вер-

шины на среднее значение, до точки перегиба гауссовой кривой (рис. 2.1).

Стандартное отклонение имеет тот же масштаб, что и среднее, поэтому округляется с той же точностью.

В пакете PAST

Выписываем Mean; Stand. Dev; округляем. Окончательный результат:

16,38; 0,73

СТАНДАРТ 3. Среднее [Доверительный интервал]

В символьной форме: $\bar{x}$ [*ДИ*; *ДИ*], $\bar{x}$ (95% ДИ: от *ДИ* до *ДИ*), где *ДИ* — нижняя граница доверительного интервала; *ДИ* — верхняя граница интервала. Такая форма записи ДИ рекомендуется рядом организаций по унификации представления статистики в публикациях (American Medical Assosiation, American Psychological Assosiation). В журнальных публикациях также встречается форма $\bar{x}$ (*ДИ* – *ДИ*) и другие.

Англ. Mean (Confidence Interval, CI), нижняя граница — Lower Confidence Limit (LCL, LL), верхняя граница — Upper Confidence Limit (UCL, UL). В публикациях встречаются обозначения: *M* [*CI*, *CI*], *Mean* (95%*CI*: *CI to CI*), *Mean* (*CI* – *CI*), *M* (*LL* – *UL*) и т. п.

Доверительный интервал (ДИ) не является собственно мерой рассеяния, но часто её заменяет. ► **ДИ** — интервал, который покрывает неизвестный параметр с заданной надёжностью. Для среднего значения обычно рассчитывают 95%-ный ДИ, который будет содержать 95 % средних значений выборок, извлекаемых из бесконечной генеральной совокупности. Таким образом, ДИ — интервальная оценка среднего, дополняющая точечную оценку. **ВАЖНО**: чем выше надёжность, тем ДИ шире, т. е. 95% ДИ шире чем 90% ДИ (см. теоретический материал). Вариантов расчёта несколько.

3.1. Строго для нормального распределения

Если распределение признаков в популяции нормальное, то при расчёте ДИ можно задействовать ***t-распределение Стьюдента*** (*Student's t-distribution*), которое стремится к нормальному распределению с ростом объёма выборки.

 **В пакете PAST** такой 95% ДИ можно получить в модуле сравнения выборочного среднего с параметром генеральной совокупности.

① Файл «Длина стопы» открыт, область значений выделена.

② Путь: Univariate — One-sample tests — Закладка по умолчанию t-test.

③ В графе параметра генеральной совокупности Given mean (задаваемое среднее) оставить значение по умолчанию — ноль и нажать кнопку [Compute].

④ Из результатов выписать 95 % conf. interval, округлить границы ДИ с той же точностью, что среднее:

16,38 [16,17; 16,59]

или

16,38 (95% ДИ: от 16,17 до 16,59)

3.2. Для любого распределения, включая нормальное

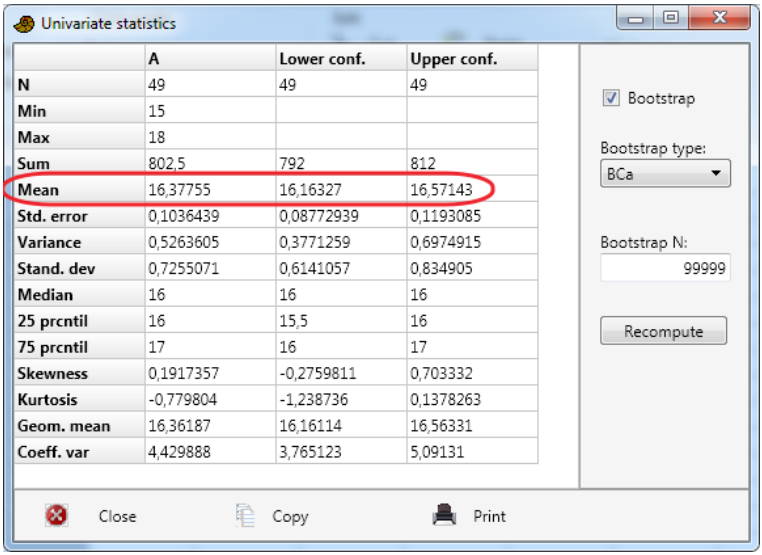
ДИ можно построить с использованием ***ресэмплинг-техник*** (*resampling*): методом складного ножа (см. теоретический материал) или более современным методом бутстрепа.

► **Бутстреп** (*bootstrap, bootstrapping*) — это современная ресэмплинг-техника, то есть техника, основанная на взятии повторных (*re…*) выборок (*…sample*). Представьте, что мы выписываем все 49 значений длины стопы полёвок на отдельные карточки, перемешиваем их и достаём одну. Выписываем значение (например, 18), возвращаем карточку в колоду и снова перемешиваем. Затем достаём новую случайную карточку и выписываем второе значение. **Вопрос:** может оказаться, что это тоже значение 18? Затем третью и т. д. столько раз, сколько наблюдений в выборке, то есть в нашем случае 49 раз. **Вопрос:** может ли оказаться, что вся такая выборка будет состоять только из значений 18? Теоретически — может, хотя это и маловероятно. Таким образом, из исходных данных мы *сгенерировали выборку с возвратом* значений. Далее в сгенерированной выборке рассчитывается интересующая статистика, в нашем случае — среднее значение. Так делается много раз: 10–500 тысяч. Следовательно, имея изначально только 49 значений в одной выборке и одно среднее

значение, мы получаем тысячи средних значений. Далее строится распределение этих средних значений (это мы научимся делать в ходе лабораторной работы № 4) и с его концов отрезается по 2,5 % площади распределения, т. е. по ***2,5 процентиля***. В результате остаётся непараметрический 95% ДИ для среднего, вычисленный с помощью процедуры бутстрепа ***процентильным методом***. Кроме него есть и другие методы бутстрепа, один из лучших — ***метод BCa*** (*Bias Corrected accelerated* — ускоренный бутстреп с поправкой на смещение).

В пакете PAST

① Файл «Длина стопы.dat» открыт, область значений выделена.
② Путь: Univariate — Summary statistics.
③ В меню Bootstrap type выбрать метод [BCa].
④ Число бутстреп-выборок N можно увеличить: добавьте ещё одну девятку — 99 999. **Вопрос**: почему лучше брать нечётное число?
⑤ Поставьте галочку в ☑ Bootstrap и программа проведёт расчёт.
⑥ Пересчитайте несколько раз, нажимая на кнопку [Recompute]. Некоторые значения в таблице результатов изменяются. **Вопрос:** почему?



	A	Lower conf.	Upper conf.
N	49	49	49
Min	15		
Max	18		
Sum	802,5	792	812
Mean	16,37755	16,16327	16,57143
Std. error	0,1036439	0,08772939	0,1193085
Variance	0,5263605	0,3771259	0,6974915
Stand. dev	0,7255071	0,6141057	0,834905
Median	16	16	16
25 prcntil	16	15,5	16
75 prcntil	17	16	17
Skewness	0,1917357	-0,2759811	0,703332
Kurtosis	-0,779804	-1,238736	0,1378263
Geom. mean	16,36187	16,16114	16,56331
Coeff. var	4,429888	3,765123	5,09131

⑦ Смотрим строчку Mean и выписываем нижнюю границу ДИ — Lower conf. и верхнюю границу ДИ — Upper conf. Округляем их с такой же точностью, как среднее, в нашем случае — до сотых.

Результат:

16,38 [16,16; 16,57]

или

16,38 (95% ДИ: от 16,16 до 16,57)

ВАЖНО: техника бутстрепа достаточно универсальная, рекомендуется считать ДИ именно бутстрепом.

> **К сведению**. Во введении мы рекомендовали хранить свои данные в электронных таблицах Excel с двумя листами: «Данные» и «Коды». Проведя в пакете PAST расчёт описательной статистики, можно нажать на кнопку [Copy] под таблицей результатов, а в Excel создать третий лист «Описательная статистика» (или «Статистика» — если предполагается продолжение наполнения листа результатами статобработки) и вставить в него результаты из буфера обмена. Таким образом, результаты статистического анализа окажутся сохранёнными вместе с данными и будут легко доступны в случае необходимости.

II. Количественные признаки с ненормальным распределением и порядковые признаки

1. Если распределение ненормальное и известно, какое именно, то данные можно нормализовать с помощью подходящего преобразования (логарифм, квадратный корень, преобразование Бокса — Кокса, угловое фи-преобразование и т. д.). На рис. 2.2 принцип нормализующего преобразования показан на примере преобразования степенной функцией.

В результате преобразования данные становятся приблизительно нормально распределены. По ним рассчитывают среднее и границы ДИ, которые затем ретрансформируют в исходную шкалу с помощью обратного преобразования.

Так, например, работая с признаком «площадь», следует извлечь из площадей квадратный корень, провести расчёт среднего и границ ДИ, а полученные числа возвести в квадрат. Работая с приблизительно *логнормально* распределёнными данными (например, численности организмов, скорости реакций, концентра-

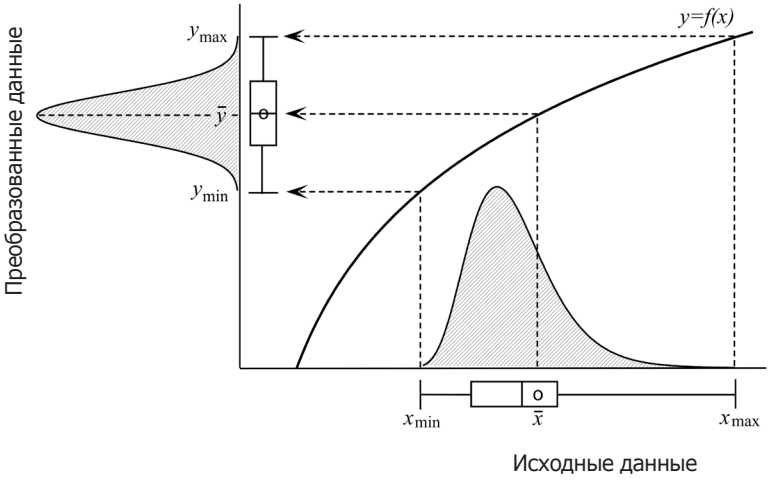


Рис. 2.2. Степенное преобразование, переводящее положительно асимметричное распределение в нормальное

ции в широком диапазоне и т. п. — см. теоретический материал), следует взять их десятичный логарифм, провести расчёт среднего и границ ДИ и полученные числа пересчитать в исходную шкалу путём возведения числа 10 в эти степени. При таком подходе ДИ станут немного асимметричными — как и само распределение.

> **К сведению**. В специальной литературе описанное выше обратное преобразование называется ***наивной ретрансформацией*** (*naïve retransformation*). Оно точно́ только алгебраически, но не статистически, поскольку не учитывает изменение ошибки разброса данных (дисперсии) при пересчёте в исходную шкалу. В результате среднее и ДИ оказываются немного смещёнными. Учёт величины такого смещения может быть важен в особых случаях (например, установление референтных значений показателя в популяции), хотя практики обычно им пренебрегают.

2. Если распределение ненормальное или неизвестное, можно использовать ресэмплинг-техники: складной нож или бутстреп. **ВАЖНО!** Но для сильно асимметричных распределений получаемые в результате ДИ оказываются не слишком точны, то есть не обеспечивают заданного покрытия истинного среднего значения генеральной совокупности. Поэтому данный подход можно сочетать с предыдущим, то есть работать по алгоритму: 1) преобразование; 2) расчёт среднего и ДИ бутстрепом; 3) обратное преобразование.

3. Если распределение ненормальное, неизвестное или признаки порядковые, часто переходят к *порядковым статистикам*, то есть рассчитывают медиану, квартили, процентили.

Медиана (нижняя квартиль — верхняя квартиль)

В символьной форме: *Me* ($Q_1 - Q_3$).

Англ. *Median* ($Q_1 - Q_3$), *Median* ($IQR = Q_1 - Q_3$), *Median* (q_1 to q_3) и др.

► **Медиана** (50-й процентиль) — это значение в центре упорядоченного по возрастанию или убыванию ряда. Так, например, если группу студентов построить в ряд по увеличению роста, то рост оказавшегося в центре человека и будет медианой. *Нижняя квартиль* Q_1, или 25-й *процентиль*, отсекает 25 % наблюдений начиная от минимума, а *верхняя квартиль* Q_3, или 75-й процентиль, отсекает 25 % в конце ряда, включая максимум (рис. 2.3). Таким образом, в ***межквартильном размахе*** (*interquartile range*, *IQR*) между Q_1 и Q_3 находится 50 % наблюдений.

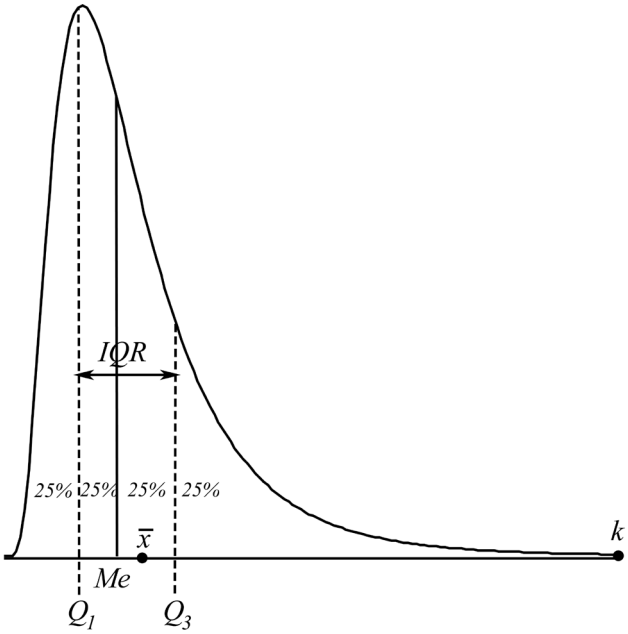


Рис. 2.3. Показатели центральной тенденции и размаха для асимметричного распределения

В отличие от среднего значения порядковые статистики очень устойчивы к отклонениям распределения от нормального — ***робастны*** *(robust)*. Так, если в ряду студентов последнего человека с ростом 185 см заменить на гиганта *k* с ростом 230 см, то среднее значение выборки сразу увеличится. **Вопрос**: а как изменится медиана? Медиана в данном случае не изменится, поскольку в центре ряда будет находиться тот же самый студент. В силу таких свойств, медиана и квартили популярны в некоторых областях биологии и медицины, где распределения показателей сильно асимметричны.

 **В пакете PAST** выписываем: Median, 25 prcntil, 75 prcntil. Медиана и квартили имеют такую же точность, как и сами наблюдения, поэтому естественно приводить их с тем же числом знаков, что и исходные наблюдения, хотя пакеты обычно выдают их с бóльшим числом знаков. Окончательный результат:

16 (16–17)

К сведению 1. Не существует единой точки зрения на то, как рассчитывать порядковые статистики в случае, когда в центре ряда оказывается не одно значение (в случае нечётного объёма выборки), а два (в случае чётного). Есть сторонники подхода с использованием интерполяции между соседними значениями в центре, то есть, например, если в центре окажутся цифры 5 и 6, то следует брать их среднее — 5,5 (так считают и авторы PAST). Другие считают это неправильным: поскольку порядковая статистика дискретна, то следует брать конкретное значение, например, большее из двух центральных, то есть не 5,5, а 6. Разные программы могут использовать разные подходы, поэтому в части медиан, квартилей и процентилей результаты расчётов в разных пакетах могут отличаться.

К сведению 2. Представленный стандарт написания является менее обоснованным по сравнению с ДИ. В отечественной технической литературе для обозначения подобных диапазонов и интервалов используется форма (*a*…*b*). Межквартильный размах также представляет собой интервал (между нижней и верхней квартилями), причём сами значения квартилей в него не входят. Поэтому, вероятно, правильнее будет представлять медиану и квартили в форме *Me* (Q_1…Q_3) или *Me* (Q_1; Q_3).

* * *

Разные стандарты можно сочетать: например, в ячейке таблицы привести над чертой среднее и 95% ДИ, а под чертой медиану и квартили:

16,38 [16,16; 16,57]
16 (16–17)

Указание на это следует привести в конце названия таблицы или в примечании под таблицей.

III. Качественные номинальные признаки

Такие признаки нельзя упорядочить естественным образом. Примеры: вид организма, цвет венчика цветка, сорт сельскохозяйственной культуры, диагноз пациента. Эти данные представляют собой частоты:

1) абсолютные (в штуках) или

2) относительные (в долях единицы, в процентах, в промилле и др.).

Раньше такие данные обычно выражали в процентах, изредка снабжая стандартной ошибкой процента, вычисленной по ***формуле Вальда*** (для больших выборок). В настоящее время принято приводить и абсолютные, и относительные частоты, а последние снабжать 95% ДИ. 95% ДИ можно вычислить разными способами; лучшие методы: ***метод Джеффриса*** (*Jeffreys' CI for proportion*), ***метод Уилсона*** (*Wilson…*), ***метод Агрести — Коулла*** (*Agresti-Coull…*, откорректированный метод Вальда). Традиционен, но несколько более консервативен ***точный метод Клоппера — Пирсона*** (*Clopper-Pearson…*). ***Метод Вальда*** (*Wald CI for proportion*), который описан во всех учебниках по статистике, в настоящее время не рекомендуется использовать даже для больших выборок (Brown et al., 2001). Все эти методы редки в пакетах, но есть в многочисленных онлайновых калькуляторах.

П р и м е р. Из 100 проанализированных клеток 5 содержали хромосомные аберрации. Задание: найти среднюю частоту аберрантных клеток и 95% ДИ для неё.

Способ 1. В пакете Excel

Воспользуемся расчётным файлом Excel из набора материалов к лабораторному практикуму «Доверительный интервал для долей.xls».

Изменяем значения в жёлтых полях: в поле [N] вносим объём выборки 100, в поле [k] — абсолютную частоту 5. Выписываем частоту P и ДИ методом Клоппера — Пирсона.

Задание. Посмотрите, как изменяется ширина ДИ и его значения для разных методов с изменением объёма выборки. Давайте оставим относительную частоту 5 %, но рассчитаем её как 1 случай из 20, 5 из 100, 50 из 1000, 500 из 10000. Обратите внимание:

– при 1/20 методы Вальда и Агрести — Коулла дают невозможное отрицательное значение нижней границы ДИ;

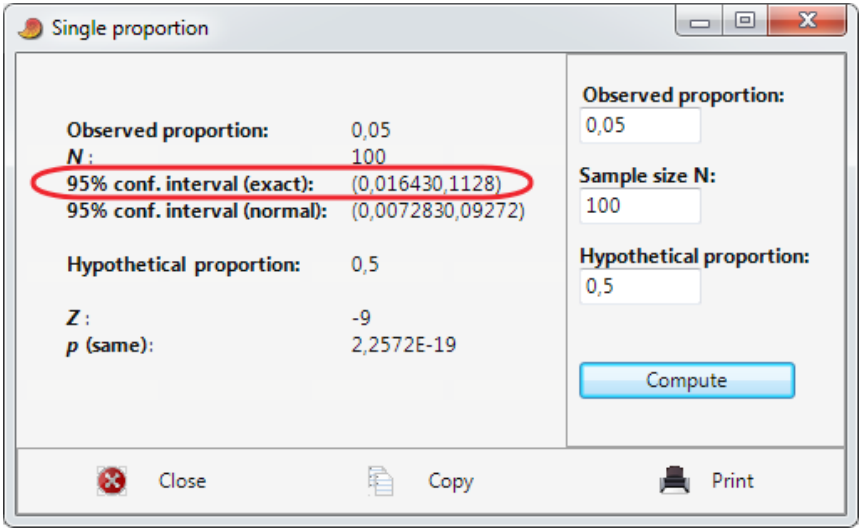
– при 50/1000 (большая выборка) все методы дают очень близкие значения ДИ, а при 500/10 000 — одинаковые.

Способ 2. **В пакете PAST**

① Путь: Univariate — Single proportion test.

② В поле [Observed proportion] (Наблюдаемое отношение) вбить относительную частоту в долях единицы, а в поле [Sample size N] — Размер выборки и нажать [Compute] (Вычислить).

③ Границы 95% ДИ по Клопперу — Пирсону взять из строки 95% conf. interval (exact).



Обратите внимание, что в пакете PAST вводить данные менее удобно, и мы не можем выбрать иной интервал, кроме 95% ДИ. Зато пакет предоставляет возможность оценить статистическую

значимость отклонения эмпирической доли от теоретической. Если таковая известна, её следует ввести в поле [Hypothetical proportion] и после расчёта выписать значение *p* (same).

Проценты обычно округляют до десятых, но в случае очень больших выборок (тысячи) можно округлять до сотых.

Итак, окончательно средняя частота клеток с аберрациями [95% ДИ] составила:

5,0 % [1,6; 11,3]

или

5,0 % (95% ДИ: от 1,6 до 11,3)

④ **Оформление в квалификационной работе (вариант)**.

4.1. Статистическая часть раздела «Материалы и методы».

При описании данных для количественных признаков рассчитывали средние значения с 95% ДИ, вычисленными процедурой бутстрепа (метод BCa, $n = 99\,999$), а также медиану с квартилями. Для качественных признаков находили абсолютные и относительные (в %) частоты; последние снабжали 95% ДИ, вычисленными по Клопперу — Пирсону. Расчёты выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

4.2. Раздел «Результаты и обсуждение».

Даются таблицы с абсолютными и относительными частотами с 95% ДИ.

* * *

Домашнее задание. Найдите и выпишите в тетрадь в продолжение этого практического занятия адреса четырёх онлайновых калькуляторов. Рассчитайте в них среднее и ДИ для одного случая из 100 и одного случая из 1000. Сформулируйте фразы для оформления в квалификационной работе.

В браузере

В строке поиска браузера нужно напечатать «Clopper-Pearson CI calculator», а в калькуляторе — выбрать 95% ДИ или 0,95, объём выборки N и число интересующих событий в этой выборке X или K.

Посмотрите, какими ещё методами возможен расчёт ДИ. Обратите внимание на пакеты, которые позволяют рассчитать наи-

более рекомендуемый в настоящее время байесовский ДИ Джеффриса (*Jeffreys' CI for proportion*). Он всегда находится внутри ДИ Клоппера — Пирсона, а потому менее консервативен.

Выделите в тетради подчёркиванием или цветом адрес самого, на ваш взгляд, удобного и/или информативного калькулятора.

В печатных работах на онлайновые калькуляторы и другие интернет-ресурсы можно ссылаться; ссылки на них должны быть грамотно оформлены в соответствии с ГОСТ или требованиями журнала.

ЛАБОРАТОРНАЯ РАБОТА № 3

Графические возможности статистических пакетов. Описательная статистика на графиках

Тема 5. Описательная статистика.

Количество часов: 2 (примечание: по причине большого объёма теоретической части, разделы I и II выносятся на самостоятельное изучение).

Цель: Познакомиться с графическими возможностями статпакета и освоить построение столбчатых и коробчатых графиков для характеристики центральной тенденции и рассеяния данных. Работа на ПК.

Средний исследователь пишет в год 3–5 научных статей и участвует в одной конференции. Одна статья содержит обычно 1–3 рисунка, презентация — немного больше — 3–5, поскольку включает фотографии, карты, блок-схемы, схемы и т. п. Научные отчёты в расчёт можно не брать, так как там, как правило, будут те же самые иллюстрации. Таким образом, за год получается около $5 \times 3 = 15$ рисунков в статьях и 5 в презентациях, то есть около 20 рисунков. Даже новичок в состоянии по инструкции подготовить за 1–2 часа в день один добротный рисунок, причём с навыком время будет сокращаться. Таким образом, затратив всего около двух рабочих дней в году, исследователь полностью иллюстрирует свою научную работу. Читатели вашей квалификационной работы или статьи могут не знать, насколько вы хороший работник (грамотный, ответственный, аккуратный и т. д.) и будут составлять впечатление о вас, ваших соавторах, а возможно, и всей организации, где вы учитесь или работаете, по косвенным признакам. То есть они будут оценивать: владение темой исследования, актуальность и новизну работы, приборное обеспечение лаборатории, методологическую и статистическую грамотность, логику изложения результатов, а также удобство подачи информации в таблицах и рисунках. Если исследователь не в состоянии одинаково округлить и отформатировать данные в таблице, забывает подписать оси на графиках, использует в оформлении 10 шрифтов и т. п., значит, и работник он такой же: торопится, не следует методикам, что-нибудь забывает и вообще занимается не наукой, а художественным творчеством. **ВАЖНО!** Поэтому не экономьте время на качестве иллюстраций к своей

работе, пусть они выглядят строго, добротно и будут выполнены в едином стиле.

На этом занятии мы познакомимся с принципами создания качественной научной графики, научимся строить графики средствами статистического пакета, а также, при необходимости, доводить их до совершенства в графических редакторах.

I. Основные принципы создания качественной научной графики

График должен помогать восприятию информации, но никак не отвлекать от неё. Сейчас у вас мало опыта для восприятия более «тонких» рекомендаций, поэтому придерживайтесь нескольких главных.

1. Полный отказ от ложного третьего измерения!

Посмотрите на рис. 3.1 — каждый из вас видел такие ***столбчатые диаграммы*** (*Bar chart*) в статьях, в презентациях, а может быть, строил сам. На этом рисунке собрано сразу несколько типичных ошибок, но главной из них является ложное 3D. **Задание**: скажите, какие значения иллюстрируют эти три столбца. Выпишем все варианты: {18, 6, 1}, {19, 7, 2}, {20, 8, 3}.

В действительности здесь отображены значения {20, 8, 3}. А раз существуют разные мнения, значит, неуместное 3D вводит в заблуждение и мешает восприятию информации.

Существует и более глубокая причина, по которой серьёзные исследователи избегают подобных графиков. Дело в том, что

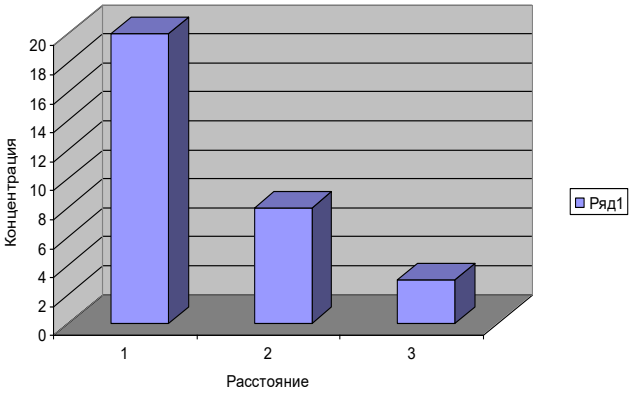


Рис. 3.1. Уменьшение концентрации меди в почве с расстоянием от источника загрязнения

помимо точности наука также «любит» экономность. Если мы хотим отобразить изменение одного показателя в зависимости от изменения другого, значит мы вправе потратить на эти два показателя только два измерения графика. Третье измерение, придающее рисунку объём, в данном случае является лишним, поскольку на него не приходится никакого содержательного третьего показателя. Это относится и к ***круговым диаграммам*** (*Pie chart*), которые часто строят в виде объёмной таблетки или разрезанного на куски торта. Получается, что за кажущейся солидностью такого 3D-графика кроется непонимание одного из основных методологических принципов науки, известного как *бритва Óккама*: **не следует множить сущее без необходимости**. Данный принцип экономии упоминается обычно в связи с научной трактовкой явлений, но может быть применен и к нашей ситуации. Помните, что если вы строите графики с ложным третьим измерением, часть грамотного научного сообщества уже только по этой причине будет оценивать вас не как равных себе, в лучшем случае — снисходительно.

2. Рациональное использование пространства рисунка. График должен, по возможности, занимать всё пространство рисунка. Следите за тем, чтобы слева, справа и сверху не оставалось больших пустот. Избавиться от пустот поможет правильная настройка разметки осей: обычно это настройка минимума и максимума, но может потребоваться также логарифмическая шкала. Если, в силу особенностей отображаемой информации, остаётся большая пустая область, в неё можно поместить легенду. В первую очередь, это относится к рисункам квалификационных работ, отчётов и презентаций, поскольку центральные журналы не любят большие легенды и потребуют доработать график так, чтобы легенда была минимальной, а как можно больше информации из неё было перенесено в подрисуночную подпись.

3. Простота при информационной насыщенности: «максимум информации при минимуме чернил!» Для новичка это более сложное требование по сравнению с первыми двумя. Пока рассмотрим три варианта повышения информационной ёмкости.

3.1. Рекомендация 1: указывайте 95% ДИ и аналогичные меры. Любая статистическая оценка получается нами на основе изучения выборки, а потому подвержена ошибкам. Эти ошибки будут тем больше, чем меньше объём выборки и/или выше вариа-

бельность признака. Поэтому точечные оценки отображаемых параметров желательно снабжать интервалами, отражающими степень нашей уверенности в этой точечной оценке. Обычно это 95%-ные ***доверительные интервалы*** (ДИ) для средних значений, 95%-ные ***доверительные границы*** для линий регрессии, 95%-ные ***доверительные эллипсы*** для корреляций. Оцените, насколько более точным, информационно ёмким, «научным» является график на рис. 3.2 по сравнению с 3.1. Он не только не допускает вариантов интерпретации значений 20, 8 и 3, но также позволяет ориентировочно оценить статистическую значимость различий. **ВАЖНО!** Если 95% ДИ не перекрываются, значит, с вероятностью 95 % средние относятся к разным генеральным совокупностям, то есть различия статистически значимы ($P \leq 0{,}05$). Если 95% ДИ перекрываются, значит, скорее всего, они относятся к одной совокупности, то есть не различаются ($P > 0{,}05$) (почему «скорее всего» — см. в теоретическом материале). На нашем рисунке концентрация меди в первой точке *статистически значимо* превосходит таковую в точках 2 и 3, а различия между точками 2 и 3 при данном способе проверки *статистически незначимы* (о том, почему правильнее говорить «статистическая значимость», а не «достоверность», см. комментарий 2 на с. 82).

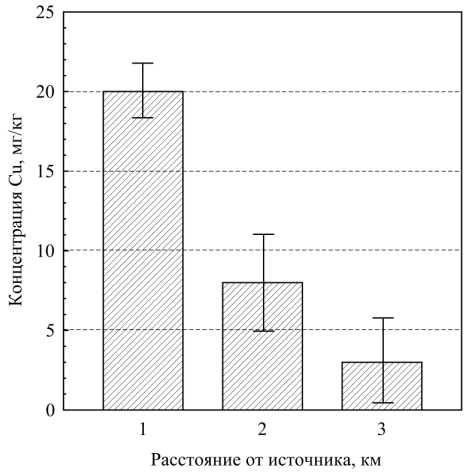


Рис. 3.2. Изменение концентрации меди в почве с расстоянием от источника загрязнения. Усы: 95% ДИ

3.2. Рекомендация 2: совмещайте несколько графиков в одном, когда это уместно. Возможно, в статьях вы встречали графики, где помимо привычной оси слева есть ещё и правая ось, на которой отложен ещё один показатель. В нашем примере с медью таким третьим показателем могло быть, например, видовое разнообразие. Если сделать столбцы рис. 3.2 более узкими и дать разную штриховку для концентрации меди и индекса разнообразия, то получим сдвоенный график. Если мы хотим продемонстрировать, что снижение разнообразия вызвано именно загрязнением медью, такой график будет удачным решением. Также на одном рисунке можно привести гистограмму распределения с кривыми плотности распределения, несколько линий регрессии, рассчитанных по разным моделям и т. п. **ВАЖНО!** Обращайте внимание на сложные графики в научных статьях, подмечайте и перенимайте приёмы повышения информационной насыщенности графиков.

3.3. Рекомендация 3: старайтесь не использовать столбчатые графики, если по оси X находится количественный показатель. Если независимый группирующий показатель измерен в количе-

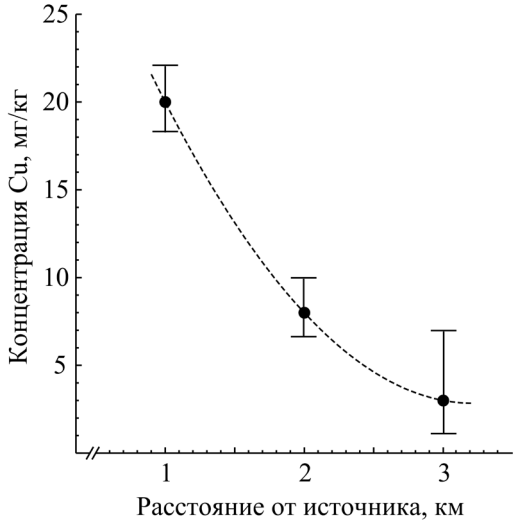


Рис. 3.3. Изменение концентрации меди в почве с увеличением расстояния от источника загрязнения. Усы — 95% ДИ, пунктир — интерполяция полиномом 2-й степени

ственных и порядковых шкалах, то зависимый показатель лучше изобразить точками или аналогичными символами (квадраты, треугольники, ромбы, звёзды и др.). Их можно соединить между собой ломаной *линией профиля* или обобщить зависимость в ходе регрессионного анализа. Наш пример как раз подходит под эту категорию, поскольку по оси X здесь — расстояние, то есть количественный показатель в шкале отношений. Поэтому рис. 3.2 также нельзя считать удачным.

Посмотрите на рис. 3.3, в котором применены все три рекомендации:

1) согласно рекомендации 1 средние значения изображены с 95% ДИ. Обратите внимание, что на этом графике ДИ асимметричны. Это значит, что распределение концентрации меди, вероятно, также имеет положительную асимметрию. Дополнительно это указывает и на квалификацию автора рисунка, который знает, что концентрации часто распределены ненормально, а потому нашёл способ грамотного расчёта ДИ и сумел отобразить результат на графике;

2) согласно рекомендации 2 повышена информативность рисунка: поверх средних с ДИ наложена кривая, интерполирующая значения концентрации между изученными точками. ► **Интерполяция** (*interpolation*) — это нахождение промежуточных значений по имеющемуся дискретному набору известных значений. В отличие от столбчатой диаграммы рис. 3.2 кривая линия лучше визуализирует процесс снижения концентрации меди с удалением от источника. Вообще говоря, со статистической точки зрения, для демонстрации зависимостей более уместной техникой была бы регрессия. Но выбор какой-либо формы нелинейной регрессии при наличии всего трёх точек сложно назвать обоснованным. Возможно, именно поэтому автор рис. 3.3 не стал использовать регрессию, однако нашёл способ помочь нам увидеть форму нелинейной зависимости. В подписи к рисунку указано, что интерполяцию он проводил ***полиномом*** 2-й степени, то есть отрезком ветви параболы. Теоретически это неоправданный, но часто используемый на практике способ приближения неизвестных нелинейных зависимостей с одним изгибом кривой. Поэтому к рис. 3.3 и его автору у нас нет никаких претензий.

ВАЖНО! Соединять точки на графиках плавной линией могут разные пакеты, в том числе часто используемый новичками

MS Excel, и такие графики нередко можно увидеть в публикациях. Тем не менее в помощи к пакету Excel невозможно найти указание на используемый алгоритм сглаживания, а значит, полученными графиками нельзя иллюстрировать научную работу. Всегда указывайте в работе способ сглаживания нелинейных зависимостей! Некоторые способы сглаживания мы рассмотрим на лабораторной работе № 13;

3) согласно рекомендации 3 средние значения приведены не в виде столбцов, а указаны точками. Это позволило гармонично наложить интерполирующую функцию, не перегружая при этом изображениями график.

4. Грамотное оформление осей графика.

4.1. Все оси должны быть подписаны, обязательно (!) с указанием единиц измерения. Если установить единицы измерения сложно или невозможно, следует писать: «у. е.» (условные единицы), «единицы шкалы прибора» и т. п.

4.2. Количество цифровых значений на осях не должно быть слишком большим. Лучше, если эти значения кратны 10^n или 5, например: 10, 20, 30, 40, 50…, или 5, 10, 15, 20…, или 1, 10, 100, 1000 (логарифмический масштаб оси). Значение 0, как правило, отображается статистическим пакетами, но редакции центральных журналов рекомендуют его не отображать (см. рис. 3.3).

4.3. Следите за тем, чтобы число ***малых меток*** (*minor tick marks*) внутри ***большой метки*** (*major tick mark*) было адекватным, поскольку автоматически построенные в статпакетах графики не всегда удачны. Например, если между большими метками 5 и 10 стоит только одна малая метка, то она отображает непонятное число 7,5. Если малых меток в интервале (5, 10) будет четыре, то они будут соответствовать единицам: 6, 7, 8 и 9, что логично и удобно для восприятия (см. рис. 3.3, ось Y).

4.4. Использование десятичной запятой. В английском и других европейских языках в качестве десятичного разделителя принято использовать точку, в то время как в русском языке — запятую. Многие статистические программы строят графики по западным стандартам и могут вместо запятой выдавать точку, например, 2.5 вместо 2,5. Если мы пишем работу на русском языке, то необходимо найти средства, чтобы изображать на осях именно запятую. Тем не менее ряд отечественных центральных журналов также требует десятичную точку вместо запятой на графиках

и даже в тексте. Однако это не говорит о правильности такого формата — редакции просто облегчают себе работу по подготовке переведённых статей для англоязычной версии своего журнала в ущерб грамотности.

5. Шрифты.

5.1. В идеале шрифт на графике должен быть один. Зарубежные научные издания предпочитают строгий шрифт Arial, который относится к бессерифным (от serif — засечка) шрифтам, то есть к шрифтам без декоративных засечек на концах букв. Большинство отечественных центральных журналов по неясным причинам требуют серифный шрифт Times New Roman — помните об этом, готовя рисунки для публикаций.

5.2. Если необходим второй шрифт, то следует выбирать такой же, но только жирный или курсивный. Латинские видовые названия обычно дают курсивом. Жирным шрифтом можно дать названия осей — это хорошо смотрится в презентациях и отчётах, но редакциям центральных журналов не нравится.

5.3. Размер шрифта на рисунке и в тексте должен визуально восприниматься одинаково. Допускается, чтобы шрифт текста рисунка был на 1–2 пункта меньше шрифта основного текста. Например, если текст написан шрифтом с кеглем 14 пунктов, то шрифт на вставленном в текст рисунке может быть 12–14 пунктов. Готовя рисунок к публикации в конкретном журнале, подберите его размер под страницу или колонку, распечатайте и приложите к странице журнала — это поможет увидеть, нуждается ли размер шрифта в изменении.

6. Соотношение сторон рисунка.

Раньше графики строили исходя из воспринимаемой человеком гармоничной пропорции золотого сечения $\left(\frac{\sqrt{5}-1}{2}:\left(1-\frac{\sqrt{5}-1}{2}\right)=0,618:0,382=1,62\right)$ или из соотношения сторон листа международного стандарта ISO 216 (формат A4 и производные), в котором сложенный вдвое лист сохраняет пропорции сторон ($1 : \sqrt{1/2} = 1{,}41$). То есть ширина графика была примерно в 1,5 раза больше высоты.

В настоящее время научные графики чаще строят квадратными: это удобно и для презентаций, и для научных журналов с двумя колонками на странице. Самый популярный в мире

статистический пакет R по умолчанию строит именно квадратные графики.

6.1. Презентации. Если подпись к рисунку расположить под прямоугольным рисунком, то рисунок нужно будет уменьшить; изображения на нём будут приплюснуты и мелковаты (рис. 3.4, слева). Если текст расположить справа от квадратного рисунка, то пространство будет использовано более полно (рис. 3.4, справа). Текст на таком рисунке будет лучше читаться, а сам рисунок будет крупнее.

Рис. 3.4. Расположение рисунка в окне презентации

6.2. Публикации. Вариант 1 (рис. 3.5) редакция журнала постарается не допустить, поскольку это слишком неэкономно (дорого). Прямоугольный рисунок будет скорее всего уменьшен до ширины колонки и будет мелким и приплюснутым (вариант 2). Если мы представим в редакцию квадратный рисунок, то в печать пойдёт относительно крупный, хорошо читаемый рисунок (вариант 3).

7. Форматы для сохранения научной графики.

Для цифровых изображений разработаны растровые и векторные графические форматы.

Растровые форматы сохраняют изображения попиксельно с использованием различных алгоритмов сжатия. Примеры форматов: *.bmp, *.gif, *.png, *.tif, *.jpeg. Если мы будем увеличивать такой рисунок, то качество его фрагментов будет всё более снижаться, и, в конце концов, мы увидим мозаику из квадратных

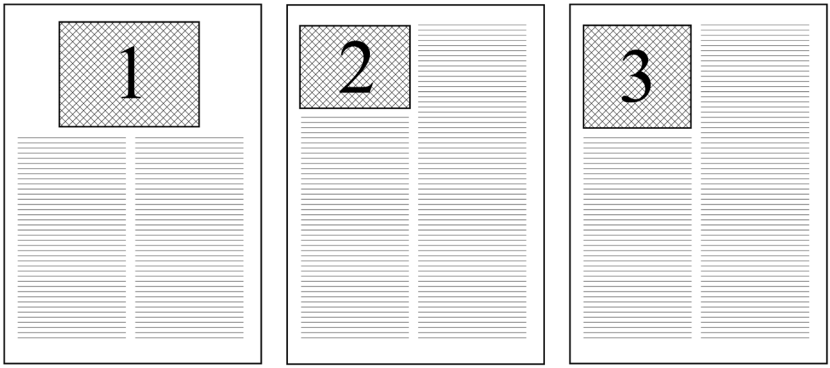


Рис. 3.5. Расположение рисунка в тексте статьи

областей — увеличенных пикселей. Поэтому для растровых форматов важно разрешение, с которым сохраняется рисунок: от этого зависит чёткость форм и детализация его элементов. Для качественного отображения рисунка в распечатанной на принтере работе достаточно разрешения 300 dpi (dots per inch — точек на дюйм, то есть на 2,54 см). Редакции журналов требуют обычно большего качества рисунков — не менее 600 dpi. Минусом растровой графики является их плохая редактируемость или отсутствие таковой. Например, если мы захотим увеличить шрифт, то в графическом редакторе придётся стирать весь (!) текст, заново его набивать (предварительно подобрав размер и тип шрифта) и разносить по областям рисунка. Плюсом растровых форматов является их неизменность: во всех программах и операционных системах они будут отображаться идентично.

Векторные форматы сохраняют изображения в виде набора формул для геометрического описания объектов. Это такие форматы, как *.svg, *.wmf, *.emf, *.eps, *.cdr. Если мы будем увеличивать данные рисунки, то по мере увеличения фрагментов программа будет заново пересчитывать форму объектов по формулам. В результате качество изображения не будет снижаться. К минусам векторной графики относится недостаточно полная совместимость между программами. Например, рисунок, сохра-

нённый в одной программе в формате *.svg, в другой программе может выглядеть несколько иначе: линии могут стать тоньше или толще, пунктирная линия может стать набором отдельных чёрточек, буквы текста могут отображаться не как текст, а как сложные графические объекты и т. п. Тем не менее чаще такие различия легко устранимы, а потому плюсом векторных форматов является их относительно хорошая редактируемость.

7.1. Многие статистические пакеты помимо основных растровых и векторных форматов позволяют сохранять рисунки также и в собственных форматах, что полностью снимает проблему редактируемости при необходимости внесения изменений. Вне зависимости от того, в какую работу готовится рисунок, если пакет позволяет — обязательно сохраняйте рисунок дополнительно в формате этого пакета. Это облегчит его редактирование в будущем.

7.2. Для печатных работ и публикаций сохраняйте рисунки с разрешением 600 dpi. Из растровых форматов редакции центральных журналов предпочитают формат *.tif. Для квалификационных работ и презентаций по соотношению «размер файла / качество изображения» хорош формат *.png.

Из векторных форматов в последнее время редакции отечественных журналов предпочитают формат *.cdr коммерческого пакета Corel Draw. Из бесплатных редакторов, позволяющих работать с форматом *.cdr, следует отметить пакет Inkscape (https://inkscape.org/ru), имеющий собственный открытый формат *.svg (со сжатием *.svgz).

7.3. Для фотографий используйте формат *.jpeg, но никогда (!) не используйте его для научной графики. Данный формат разрабатывался именно для компактного хранения фотографий. В силу специфики алгоритма в данном формате невозможно качественно сохранить чёрную линию или буквы на белом фоне: они будут несколько размыты, строгость и качество графики пострадают.

7.4. Далеко не все статистические пакеты имеют развитый модуль настройки графики либо эти настройки слишком сложны. **ВАЖНО:** это не является оправданием примитивной научной

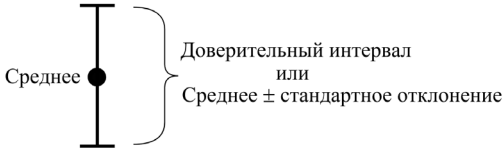
графике в работе! Старайтесь доводить графики до совершенства, используя другие пакеты. Например, уже знакомый нам пакет PAST строит весьма передовые (*advanced*) с точки зрения статистики, но примитивные с точки зрения качества графики. Поэтому мы научимся сохранять их в векторном формате *.svg и дорабатывать в другой программе — векторном редакторе TpX.

Домашнее задание. Найдите в Интернете 10 научных статей разных авторов из разных журналов по своей специальности с графиками. Проанализируйте их на предмет соответствия научной графики критериям, с которыми мы познакомились. На следующем занятии мы объединим результаты и рассчитаем долю исследователей, качественно иллюстрирующих свою работу.

II. Описательная статистика на графиках

Для графической характеристики выборок используются преимущественно те же меры, что и при их табличном описании, то есть либо меры оценки центральной тенденции и рассеяния, либо меры точечной и интервальной оценки центральной тенденции.

1. Количественные показатели (шкала интервалов и шкала отношений) чаще изображают точками с ***усами***. Точка соответствует положению среднего значения, а усы по обе стороны от точки отображают либо стандартное отклонение (мера рассеяния), либо доверительный интервал (интервальная оценка среднего, обычно 95% ДИ):



Вместо точек могут использоваться и другие символы:

○ ● □ ■ ▲ ◊ * и т. д.

Часто, в случае нескольких выборок, значки средних значений соединяют отрезками ломаной линии. Такие выборки могут представлять собой определённую последовательность, например, динамику изменения показателя. Однако отрезками могут быть соединены не только последовательности, но и разнород-

ные группы или даже разные показатели — в этом случае полученная ломаная называется ***профилем*** (рис. 3.6).

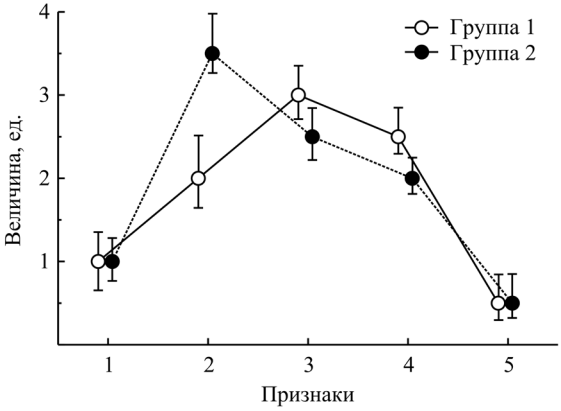


Рис. 3.6. Пример графика с изображением профиля

При анализе литературных данных обязательно находите, что именно автор обозначал усами. Помните, что, в отличие от ДИ, стандартное отклонение не столь удобно для констатации статистической значимости различий. Более того, стандартное отклонение имеет геометрический смысл только в случае нормального распределения. С особым подозрением следует относиться к графикам с одним усом: отсутствие второго уса предполагает зеркальное отображение первого, что справедливо только для симметричных распределений (рис. 3.7). Поэтому даже в случае симметричных усов отображайте оба: это укажет на то, что вы по крайней мере знаете о возможной асимметрии распределения.

Также должны вызывать подозрения слишком узкие усы: возможно, на них изображено среднее ± стандартная ошибка. С точки зрения визуальной оценки значимости различий они намного хуже даже стандартного отклонения, поскольку вообще не позволяют её провести.

ВАЖНО: в своих научных работах обязательно указывайте в легенде графиков или в подписях под ними, что́ именно вы обозначаете усами; лучше, если это будет 95% ДИ.

2. Порядковые показатели (порядковая шкала), а также количественные показатели, которые описывают порядковыми ста-

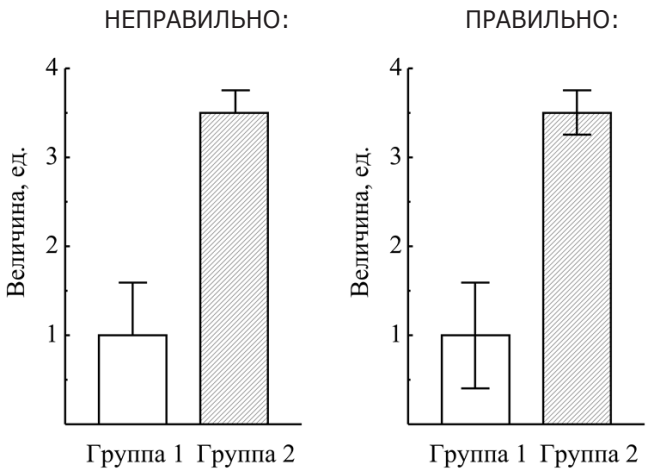


Рис. 3.7. Изображение усов доверительных интервалов на столбчатой диаграмме

тистиками, представляют в виде ***коробчатых,*** или ***ящичковых диаграмм*** (*Box-and-whisker plot*) (рис. 3.8).

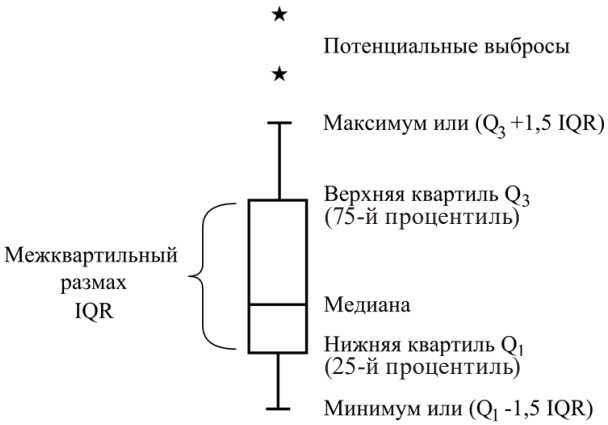


Рис. 3.8. Коробчатые, или ящичковые диаграммы

Такие графики дают отличное представление о форме распределения показателя в выборке, однако с ними связаны три проблемы, которые потребуют от вас внимания:

1) в отличие от 95% ДИ они не позволяют обнаруживать стати-

стически значимые межгрупповые различия. Поэтому для обозначения значимых различий на коробчатых графиках можно встретить скобки, с указанием значимости различий для интересующих пар выборок (см. рис. 8.2 на с. 132);

2) существует неопределённость в показателях, обозначаемых усами. Американский статистик Джон Тьюки, предложивший этот тип графика, обозначал усами 1,5 межквартильных размаха, которые вычитаются из значения нижней квартили или прибавляются к значению верхней. Однако в настоящее время ими часто обозначают минимальное и максимальное значения, а иногда можно встретить также 5-й и 95-й процентили. Если программа позволяет выбрать показатели для усов — рекомендуем привести минимум и максимум: эти показатели понятны и дают хорошее представление о размахе варьирования признака в выборке;

3) большинство пакетов по умолчанию расценивают экстремальные значения как ***выбросы*** (*outliers*), которые обозначают отдельными значками (точки, круги, звёздочки). При этом часто такие значения исключаются из набора данных и показатели описательной статистики для коробчатой диаграммы рассчитываются уже без них, что приводит к противоречиям с описательной статистикой в таблицах. Для беглого знакомства с данными и их проверкой на наличие ошибок графики с потенциальными выбросами являются удобным инструментом. **Однако категорически не рекомендуется давать такие графики в работу!** В настройках программ следует искать опции отказа от детекции потенциальных выбросов. Если в научной статье вы видите графики с потенциальными выбросами, значит, скорее всего, автор просто не знает, что многие биологические показатели распределены резко асимметрично, и позволяет программе считать за выбросы далеко отстоящие значения в хвостах распределений. Также такого автора не смущают различия между медианой и квартилями в таблицах и на графиках. Возможно, он просто не разобрался с настройками статистического пакета и строит графики «по умолчанию». В любом случае это характеризует его не с лучшей стороны и только вызывает вопросы к работе.

3. Качественные номинальные показатели (номинальная шкала) приводятся в основном на столбчатых и круговых диаграммах. Чаще всего такие данные представлены относительными частотами, выраженными в процентах.

3.1. ***Столбчатые диаграммы*** (*Bar chart*) строятся для демонстрации различий выборок. Так, на рис. 3.9 изображена частота некоего показателя в трёх группах, для которых она составляет 10, 50 и 90 %. Для демонстрации эффекта объёма выборки на рис. 3.9 эти частоты были рассчитаны для 25, 50 и 100 наблюдений.

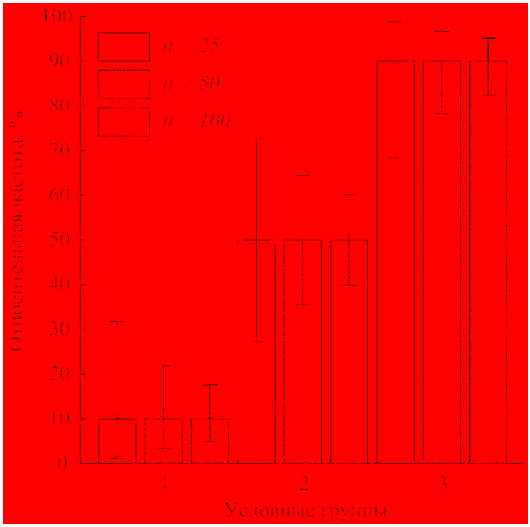


Рис. 3.9. Изменение ширины и симметрии доверительных интервалов для частот

Обратите внимание на усы, которыми обозначены 95% ДИ, вычисленные здесь методом Клоппера — Пирсона.

Во-первых, чем меньше объём выборки, тем шире ДИ — в этом ДИ для частот ничем не отличаются от ДИ для количественных показателей. Для выборок в 25 наблюдений 95% ДИ соседних групп перекрываются, а значит, различия между группами 1–2 и 2–3 сомнительны. На выборках в 50 и 100 наблюдений различия между всеми тремя группами не вызывают сомнений: они статистически значимы.

Во-вторых, ДИ симметричны для 50 %, но резко асимметричны как для 10 % (положительная асимметрия), так и для 90 % (отрицательная асимметрия). Это всегда свойственно частотам, поскольку они «зажаты» между границами 0 и 1 (в долях единицы) или 0 и 100 %, а варьирование в сторону границы ограничено математически. С ростом объёмов выборок асимметрия

становится менее заметной, но она визуально присутствует и для объёма выборки в 100 наблюдений.

Поэтому при анализе литературных данных обращайте внимание на симметрию/асимметрию ДИ для частот. Асимметрия может быть практически незаметна в интервале от 30 до 70 % для любых объёмов выборок или во всём диапазоне от 0 до 100 %, но при больших выборках (сотни и тысячи наблюдений). Если автор приводит симметричные ДИ в области 0–30 % и 70–100 % для небольших и средних объёмов выборок, то, возможно, он не разобрался в программе, а значит, вместо ДИ может быть приведено что угодно. Либо он использовал для построения ДИ метод Вальда, который основан на нормальной аппроксимации и применим только для больших выборок*. В любом случае к таким результатам следует относиться с недоверием.

> ***К сведению**: как уже указывалось выше, в настоящее время профессионалы вообще отказались от использования метода Вальда для расчётов ДИ.

3.2. ***Круговые диаграммы*** (*Pie chart*) очень распространены при описании качественных ***композиционных данных*** *(compositional data)*, то есть таких, композиция которых в сумме составляет 100 %. Например, из 150 изученных объектов у 90 (60 %) отмечалось отсутствие признака, у 45 (30 %) — слабое развитие признака, у оставшихся 15 (10 %) — нормальное развитие признака. Круговая диаграмма для этих данных может выглядеть так, как представлено на рис. 3.10.

Ещё раз напомним: никакого ложного третьего измерения! Лучше снабдить относительные частоты в процентах 95% ДИ и привести их в скобках. Это позволит повысить истинную информативность рисунка, а возможно — также и цитируемость вашей работы, поскольку такой график позволит вашим коллегам провести статистическое сравнение собственных данных с вашими, не прибегая к расчётам, а только сопоставляя ДИ: будут они перекрываться или нет.

П р и м е р. У рыб озера Чеба́куль (Челябинская область) определено содержание никеля в мышечной ткани, в мг/кг сухого вещества. Данные представлены в таблице:

Плотва	1,41	1,79	0,17	1,10	0,17	0,17	1,07	0,20	0,17	0,17
Окунь	0,84	0,46	0,37	0,45	0,37	0,42	0,40	0,44	0,41	0,36

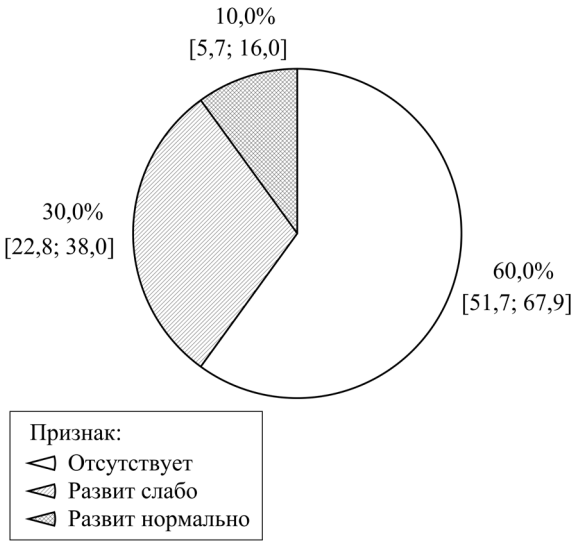


Рис. 3.10. Пример круговой диаграммы

Задание. Построить графики, характеризующие центральную тенденцию и рассеяние показателя в выборках двух видов.

В пакете PAST

① Ввести данные в две колонки, дать колонкам названия (плотва, окунь) и выделить область данных.

② Путь: Plot — Barchart/Boxplot. Plot type: Bar chart — Столбчатая диаграмма. Также можно выбрать тип Mean and Wisker. **Задание**: выберите его, опробуйте все остальные типы и вернитесь на Bar chart.

③ По умолчанию усы на графике означают стандартное отклонение. Мы отмечали, что более полезным является доверительный интервал, поэтому в Whisker length (Длина усов) ставим радиометку в положение ◉ 95% interval.

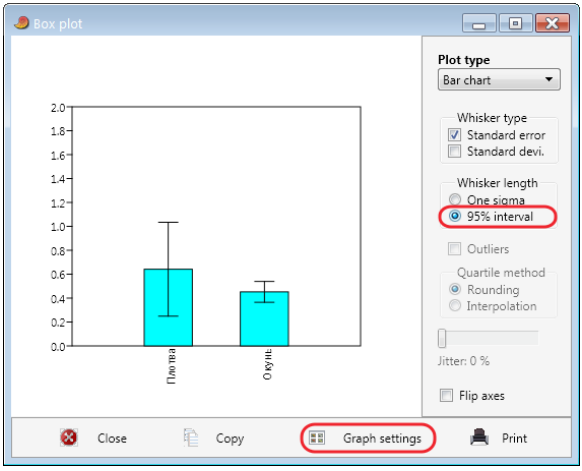
Вопрос 1: Глядя на рисунок, что мы можем сказать: 1) о средних значениях содержания никеля в мышечной ткани двух видов; 2) об изменчивости этого показателя; 3) о статистической значимости различий?

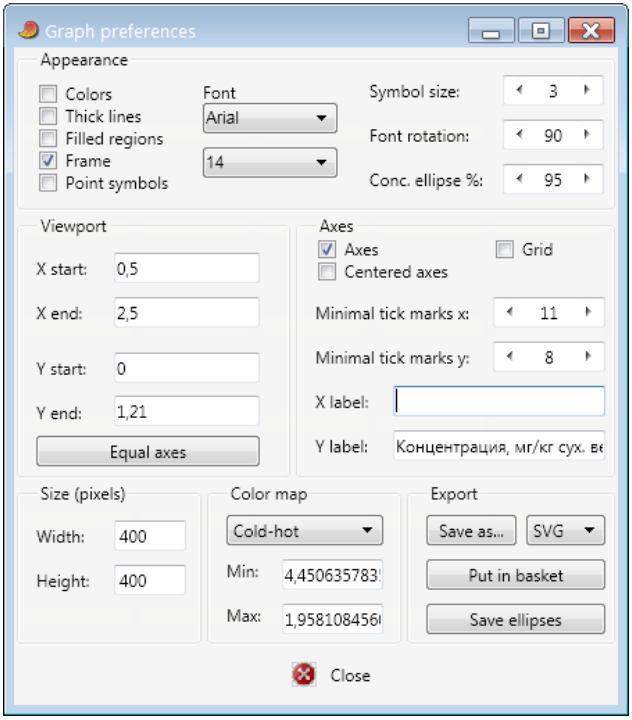
Вопрос 2: Чем плох автоматически построенный график, то есть в каких доработках он нуждается?

1) построенный график содержит много пустого пространства, то есть требует доработки по оси Y. Максимальное значение можно установить в районе 1,2, а далее настроить метки;

2) ось Y не подписана, значит, пока не понятно, что вообще изображено на рисунке;

3) в качестве десятичного разделителя стоит точка вместо десятичной запятой.





④ Редактирование графика. Входим в настройки рисунка — Graph settings.

Познакомимся с этой формой. Задание: записывайте в тетрадь английские названия, рядом — перевод на русский и пробуйте применить/отменить данную опцию.

Appearance — Внешний вид

Colors — Цвета. Снимаем галочку.

Thick lines — Толстые линии.

Filled regions — Закрашенные области. Недоступно в данном графике.

Frame — Рамка. Снимите, затем верните.

Point symbols — Символы меток. Недоступно в данном графике.

Font — Шрифт. Выберите Arial, 14 пунктов.

Symbol size — Размер символов. Недоступно в данном графике.

Font rotation — Вращение шрифта. Попробуйте уменьшить до 0, затем верните 90. Для скорости можно сразу набить цифру и нажать [Enter].

Conc. ellipce% — доверительные границы корреляционного эллипса. Недоступно в данном графике.

Viewport — Видимая область

Здесь можно задать начало (X start) и конец (X end) для оси X и ниже — для оси Y. Установите:

X start 0,5

X end 2,5

Y start 0

Y end 1,21. **ВАЖНО**: для удобства последующей разметки осей лучше брать не требуемое значение, а немного больше, то есть не 1,2, а 1,21).

Axes — Оси

Centered axes — центрированные оси.

Grid — сетка. Сетка обычно помогает восприятию графиков, поэтому её можно задать для графика в квалификационную работу и презентацию. Но, к сожалению, редакции крупных журналов требуют сетку убирать.

Minimal tick marks x — количество меток на оси x. Попробуйте разные варианты. Ничего не изменяется, поскольку у нас по этой оси находятся не количественные, а номинальные показатели — названия видов рыб.

Minimal tick marks y — количество меток на оси y. Попробуйте разные варианты, затем установите 8.

X label — название оси X. Для нашего графика оставляем поле пустым.

Y label — название оси Y. Пишем: Концентрация, мг/кг сух. вещ-ва. Подтверждаем клавишей [Enter].

Size (pixels) — Размер рисунка в пикселях

Сделаем график квадратным. Для этого выставим ширину (Width) 400, подтверждаем [Enter], и высоту (Height) 400, [Enter].

Color map — цветовая схема

Оставляем по умолчанию.

Export — Экспорт рисунка в графический формат

Мы будем использовать векторный формат по умолчанию — *.svg. Нажмите [Save as…] и сохраните файл под названием Мояфамилия_рыбы.svg.

* * *

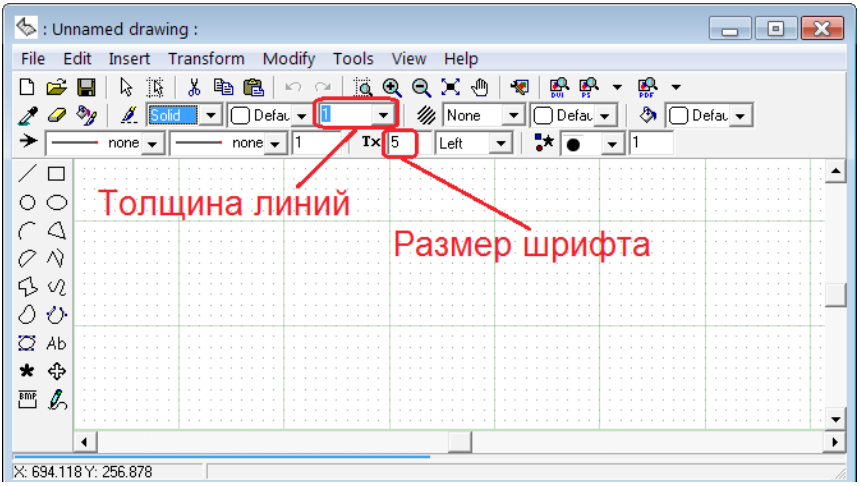
Доведём этот график до совершенства в графическом редакторе TpX. Это простой векторный редактор, очень удобный для работы именно с научной графикой. Создавался для облегчения внедрения графики в документы LaTeX — популярного макропакета системы компьютерной вёрстки T_EX. Пакт бесплатный, англоязычный. Автор — Александр Анатольевич Цыплаков, кандидат экономических наук, доцент кафедры применения математических методов в экономике и планировании Новосибирского государственного университета.

В пакете TpX

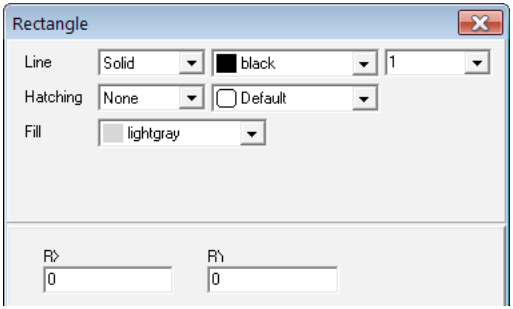
① Запустите программу. При первом запуске под указателем мыши будет передвигаться центр синего креста, который мешает работе. Поэтому зайдите в View и снимите галочку с опции ☑ Show crosshair.

② Откроем сохранённый файл: File — Open — Тип файлов: Scalable Vector Graphics (*.svg; *.svgz) — Мояфамилия_рыбы.svg.

③ Удерживая левую кнопку мыши, обведите квадратом весь рисунок для его выделения. Вы видите, что появилось много квадратиков, обозначающих отдельные элементы рисунка — их можно редактировать. Чаще всего не устраивают толщина линий рисунка и размер шрифта. Их можно изменить сразу на всём рисунке, когда он выделен. Кликните мышью в стороне от рисунка, чтобы снять выделение.



④ Кликните дважды на границе или внутри столбца для плотвы. Откроется окно редактирования этого элемента:



Line — Линия. Тип (None — нет, Solid — сплошная, Dotted — точечная, Dashed — пунктирная), цвет и толщина линии. Оставляем по умолчанию Solid.

Hatching — Штриховка. Тип штриховки и её цвет. Изменяем тип None на BDiagonal, цвет на black.

Fill — Заливка. Заменяем светло-серый lightgray на Default. Нажимаем [OK]. Для презентации можно использовать цветную заливку, в пакете достаточно большой выбор цветов и оттенков.

Задание. Точно с такими же настройками оформите столбец для окуня. Если бы групп было несколько, мы бы использовали разную штриховку или заливку цветом. В случае двух групп это будет только отвлекать.

⑤ Настройка шрифтов. Кликните дважды на названии «Плотва». Откроется меню, в котором нужно снять галочку ☑ с Custom font и нажать [OK]. Шрифт должен измениться на Times New Roman. Если этого не произошло, значит галочку нужно вернуть на место и выбрать из списка нужный шрифт (Times New Roman). Кстати, здесь же можно изменить шрифт на жирный или курсивный.

Задание. Измените весь шрифт на рисунке на Times New Roman. Также при редактировании значений оси Y изменяйте десятичную точку на запятую.

⑥ Поворот текста. Можно сделать двумя способами:

– убрать угол 90° в меню настройки текста. Кликните дважды на названии «Плотва» и в открывшейся уже знакомой форме напротив Angle (угол) сотрите число 90;

– через меню Transform. Кликните однократно на названии «Окунь». Путь: Transform — Rotate — Rotate clockwise 90 deg или можно просто нажать сочетание [Alt + Стрелка вправо].

⑦ Подровняем текст так, чтобы названия видов встали на один уровень (в этом помогают линии сетки редактора) и находились строго под центром столбцов. Для этого выделяем текст и перемещаем его стрелками на клавиатуре, удерживая клавишу [Ctrl]. Вспомнить сочетания горячих клавиш можно, зайдя в соответствующее меню, в данном случае путь: Transform — Move.

⑧ Сохраняем рисунок в формате редактора: File — Save as — Мояфамилия_рыбы.TpX, а затем в растровом формате Мояфамилия_рыбы.png. Последний рисунок готов для вставки в текстовый редактор или презентацию (рис. 3.11). Если вам покажется, что строгий чёрно-белый рисунок слишком скучен для презентации, вы всегда можете открыть файл TpX и сделать его ярче, раскрасив столбцы. Но, во-первых, не используйте слиш-

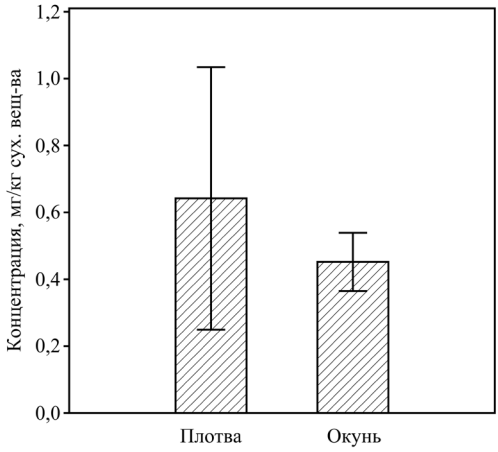


Рис. 3.11. Содержание никеля в мышечной ткани рыб оз. Чебакуль. Усы — 95% ДИ

ком много цветов, а во-вторых, эти цвета должны сочетаться со стилем шаблона презентации.

ЛАБОРАТОРНАЯ РАБОТА № 4

Анализ распределения признаков

Тема 2. Базовые понятия статистического оценивания.
Тема 4. Статистический критерий.
Количество часов: 2.
Цель: овладеть приёмами анализа распределения количественного признака с использованием графических средств (гистограмма распределения) и специальных критериев проверки на нормальность. Научиться обнаруживать гетерогенность выборки по количественному показателю и проводить разделение смеси непрерывных распределений. Работа на ПК.

► **Распределение (функция распределения)** — функция, характеризующая распределение случайной величины. Часто такие функции изображаются в виде графика распределения частот или вероятностей. Классические статистические процедуры основаны на предположении, что данные имеют эмпирическое распределение, которое близко аппроксимируется каким-либо теоретическим распределением (нормальным, логнормальным, биномиальным, пуассоновским и т. д.).

Знание о характере ***распределения*** (*distribution*) признака в популяции крайне важно:

1) для выявления резко отклоняющихся наблюдений — ***выбросов*** (*outliers*). (Но здесь нужно быть внимательным, чтобы не спутать выброс со значением в конце резко асимметричного распределения!);

2) обнаружения ***неоднородности выборки***. Выборка может быть представлена не одной группой, а несколькими подгруппами со своими средними значениями. В этом случае будет наблюдаться ***смесь распределений***;

3) для правильного описания данных и выбора способа их дальнейшего анализа при решении задач поиска различий, связей или зависимостей. Если распределение приблизительно нормальное или может быть приближено к нормальному с помощью преобразований, то в анализе можно использовать наиболее разработанные и мощные ***параметрические методы***. **ВАЖНО**: нормальное распределение должно быть не в выборке, а в популяции, откуда эта выборка извлекается — эта информация берётся из литературы и из теоретического анализа явле-

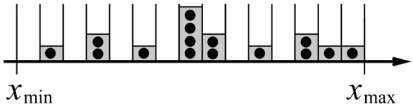
ния (см. теоретический материал). Непосредственно оценивать нормальность распределения признака в популяции по данным выборки имеет смысл при её достаточном объёме ($n \geq 30$). Последний способ справедливо критикуется некоторыми статистиками с теоретических позиций [например, 10], однако он очень распространён в исследовательской практике, поскольку часто выборка является единственным источником информации о распределении признака.

Проверка нормальности распределения признака проводится сначала графически, а затем подтверждается статистически. Рассмотрим эти этапы.

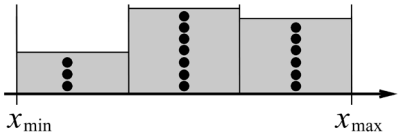
Этап I. Графический анализ распределения

Графический анализ основан на визуальной оценке формы распределения по специфическим графикам: ***гистограмме*** (*histogram*), ***полигону частот*** (*frequency polygon*) или ***кумуляте*** (графику накопленных частот). Принцип их построения одинаков для всех графиков: ***интервал*** от минимального значения $x_{\min}$ до максимального $x_{\max}$ разбивается на заданное число k интервалов и подсчитывается количество наблюдений n_k, попавших в каждый интервал. Далее эта информация откладывается на графике с интервалами признака по оси X и частотами (абсолютными или относительными) — по оси Y.

Вопрос. Как вы думаете, сколько интервалов нужно сделать, чтобы отчётливо увидеть форму распределения? Если мы выберем очень узкий интервал, то в некоторые не попадёт вообще ни одного значения и форму мы не увидим:



Если, напротив, выбрать мало интервалов, например 3, то картинка будет слишком грубая:



В действительности, удобное число интервалов зависит от объёма выборки: для больших выборок можно нарезать много узких интервалов, а для небольших выборок — мало широких. Сокал и Рольф рекомендуют [22]:

Объём выборки, n	Число классов, k
25	5–6
40–50	до 12
100 и более	более 20

Полезно ориентироваться на эмпирическое ***правило Стургеса*** (Стёрджеса, *Sturges' rule*):

$$k = 1 + 3{,}322 \times \lg n.$$

П р и м е р. Рассмотрим построение распределения на примере с длиной стопы восточноевропейской полёвки из 1-го поколения лабораторной колонии.

① Подготовительные расчёты. Из результатов предыдущего занятия имеем: $n = 49$; $x_{min} = 15$ мм; $x_{max} = 18$ мм. Точность измерения (шаг измерения) — 0,5 мм.

② Определение числа классов по Стургесу.

$k = 1 + 3{,}322 \times \lg 49 = 6{,}61$, то есть 6–7 классов; лучше взять нечётное 7 — будет лучше виден центр распределения.

③ Определение ширины ***межклассового интервала*** i. Делим расстояние от x_{min} до x_{max} на k частей, то есть в нашем случае на 7:

$$i = (x_{max} - x_{min}) / k = (18 - 15) / 7 = 0{,}43.$$

④ Вычисление границ классов. Границы классов находятся последовательным добавлением к x_{min} величин межклассового интервала i, $2i$, $3i$ и т. д. до x_{max}.

Класс	Границы
1	15,00–15,43
2	15,43–15,86
3	15,86–16,29
4	16,29–16,72
5	16,72–17,15
6	17,15–17,58
7	17,58–18,01

По договорённости верхняя граница к классу не относится, то есть первый интервал можно обозначить как [15,00; 15,43), и, если бы нам попалось значение 15,43, мы бы отнесли его уже в следующий класс.

> **К сведению**. Компьютерные программы могут проводить разметку на классы несколько иначе. Например, в качестве минимального и максимального значений ряда использовать не x_{min} и x_{max}, а значения $(x_{min} - i/2)$ и $(x_{max} + i/2)$.

⑤ Подсчёт числа попавших в классы значений. Можно заметить, что, так как в нашем случае точность измерений (0,5 мм) очень близка к межклассовому интервалу (0,43 мм), в каждом из классов окажутся строго одинаковые значения: в первом — только 15, во втором — 15,5, в третьем — 16 и т. д. Поэтому в нашем случае можно сформировать классы в соответствии с шагом измерения, приняв $i = 0{,}5$. Строим таблицу. **Задание.** Заполните все колонки таблицы. Получаем:

Номер класса	Середина класса	Частота, n_k
1	15	2
2	15,5	7
3	16	17
4	16,5	5
5	17	12
6	17,5	5
7	18	1
	$\Sigma =$	49

⑥ Построение графиков. Отложим на графике по оси абсцисс — границы класса (в нашем случае просто значение этого класса), а по оси ординат — абсолютную частоту. Если оформить график в виде столбчатой диаграммы, то получим гистограмму распределения, если соединим центры столбиков ломаной линией, получим полигон частот. **Задание**. Постройте в тетради вручную оба графика и подпишите рисунок. Помните о принципах качественной графики и обратите внимание на следующие детали:

1) обе оси должны быть обязательно подписаны с указанием единиц измерения;

2) цифр на осях не должно быть слишком много;

3) сетка на рисунке хорошо смотрится на рабочих графиках (удобно), в презентациях и квалификационных работах (удобно, солидно). Однако, как уже отмечалось ранее, центральные журналы требуют её убирать. Поэтому, если по материалам работы планируется публикация, рисунки лучше сразу строить без сетки.

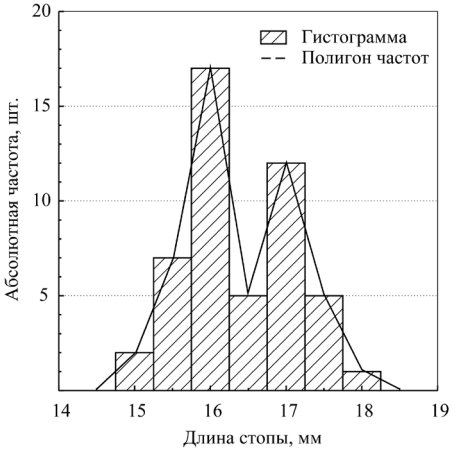


Рис. 4.1. Распределение длины стопы восточноевропейской полёвки

В пакете PAST

① Открыть файл «Длина стопы.dat» и выделить колонку значений.

② Путь: Plot — Histogram. Число интервалов Bins выставляем равным выбранному *k*, то есть в нашем примере 7. **Задание**: поизменяйте это число в бóльшую и меньшую сторону и посмотрите, как меняется график. Вернитесь к значению 7.

③ Ставим галочки:

☑ Fit normal — ***подгонка*** к нормальному распределению. Программа строит кривую нормального распределения с параметрами (среднее и стандартное отклонение), вычисленными по выборке.

☑ Kernel density — ***плотность*** распределения. Рассчитывается методом Сильвермана (см. помощь к пакету) и не зависит от выбранного нами числа классов. Видно, что, в отличие от ***унимо-***

дального (одновершинного) нормального распределения, наше распределение отчётливо ***бимодальное*** (двухвершинное).

④ Доработка графика. Опция Graph settings. Подберите Minimal tick marks x и y (минимальное количество насечек на осях x и y), чтобы график было удобно читать. Измените шрифт на Times New Roman, подобрав размер шрифта так, чтобы в окончательном документе его размер выглядел аналогично шрифту основного текста или был меньше его на 1–2 пункта.

⑤ Сохранение. Сохраните правленый рисунок в растровом формате (*.png, *.bmp, *.tif, но не *.jpg(!)). Для чистовой доработки в векторном графическом редакторе сохраните его также в формате *.svg.

Предварительный вывод 1. Графический анализ показал, что эмпирическое распределение длины стопы полёвок существенно отличалось от теоретического нормального: оно было отчётливо бимодальным.

Этап II. Статистическая проверка распределения на нормальность

Графический анализ необходимо подтверждать статистически — с помощью критериев, поскольку в зависимости от объёма выборки, компетенции и опыта исследователя разные люди могут по-разному интерпретировать график, а нам необходим объективный вывод. Для проверки распределения на нормальность предложено более 20 критериев, которые относятся или к ***критериям согласия*** (так как проверяют согласие эмпирического распределения с заданным, в данном случае — с нормальным), или непосредственно к ***критериям проверки нормальности***. Они отличаются мощностью по отношению к разным типам отклонений от нормальности: асимметрии, эксцессу и их сочетаниям [5]. В большинстве ситуаций высокую мощность демонстрирует ***критерий Шапиро — Уилка*** (критерий нормальности, *Shapiro-Wilk test*). Что такое «мощность» в статистическом смысле — см. теоретический материал. В статпакетах обычно также распространены:

1) ***критерий хи-квадрат*** (*Chi-square test*) — критерий согласия; сравнивает ряд предварительно сгруппированных наблюдаемых частот с рядом сгруппированных ожидаемых частот, вычисленных в предположении нормального распределения показателя.

Мы познакомимся с этим критерием позже и для другой задачи (см. лабораторную работу № 6);

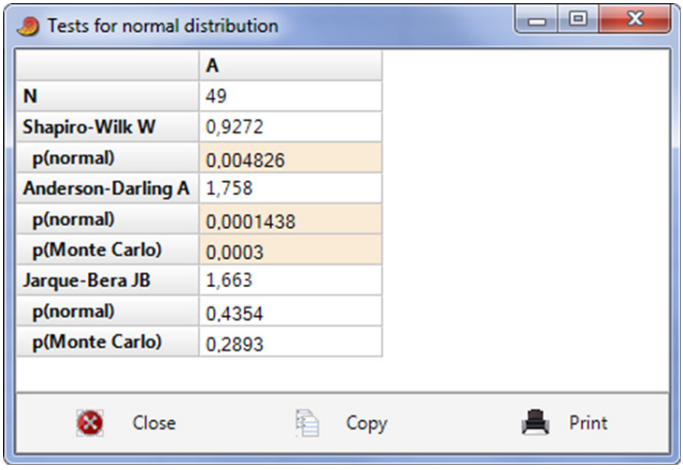
2) ***критерий Колмогорова — Смирнова*** (*Kolmogorov-Smirnov test*) — критерий согласия; сравнивает наибольшее отклонение ряда накопленных частот от ряда накопленных частот любого распределения, в том числе нормального. Модифицированный вариант этого критерия для проверки именно нормальности называется ***критерием Лиллиефорса*** (*Lilliefors test*);

> **К сведению**. Используя статистику Колмогорова, А. Н. Колмогоров и Н. В. Смирнов разработали очень близкие непараметрические подходы. В статистических пакетах в качестве критерия Колмогорова — Смирнова представлен критерий однородности Колмогорова.

3) ***критерий Андерсона — Дарлинга*** (*Anderson-Darling test*) — популярный критерий согласия, основанный на вычисляемой с использованием натуральных логарифмов весовой функции: чем дальше от центра распределения находится наблюдение, тем больший вес имеет его отклонение.

В пакете PAST

① Файл «Длина стопы.dat» открыт, колонка значений выделена.
② Путь: Univariate — Normality tests.



В окне результатов N — объём выборки, а далее следуют критерии и соответствующие *p* для них: Шапиро — Уилка, Ан-

дерсона — Дарлинга, Харке — Бера (проверяет нормальность посредством объединённой проверки асимметрии и эксцесса). Для последних двух методов значение *p* рассчитывается двумя способами: 1) асимптотически (корректно для больших выборок) — p (normal); 2) с помощью рандомизационной процедуры Монте-Карло — p (Monte Carlo), что более предпочтительно.

Интерпретация. Мы впервые столкнулись с результатом статистического критерия, поэтому рассмотрим подробнее цепочку наших рассуждений для принятия вывода.

► **Значение *P*** (*p-value*) — вероятность наблюдать имеющееся и ещё более экстремальное значение статистики при условии справедливости ***нулевой гипотезы*** H_0. То есть *P* — непрямая оценка вероятности H_0. Нулевая гипотеза — гипотеза об отсутствии проверяемых предположений, в данном случае H_0: эмпирическое распределение не отличается от нормального. Если её вероятность, оцениваемая по *P*, мала, например 0,05 и менее ($P \leq 0{,}05$), то велика вероятность ***альтернативной гипотезы*** H_A: распределение отличается от нормального. В нашем случае для критерия Шапиро — Уилка $p = 0{,}004826$. По существующей в рамках ***частотного подхода*** (*frequentist approach*) к принятию статистических решений договорённости для большинства ситуаций малым считается значение $P \leq 0{,}05$ (см. теоретический материал). Поскольку $p = 0{,}004826$ — очень малая вероятность для нулевой гипотезы, поэтому такую маловероятную H_0 мы отклоняем и делаем вывод о ***статистической значимости*** (*statistical significance*) отличия распределения от нормального или даже о ***высокой*** статистической значимости, так как $P < 0{,}01$. В ситуации $P > 0{,}10$ мы бы оставили H_0 в силе: отличий от нормального распределения нет. В промежуточных ситуациях, когда $0{,}05 < P \leq 0{,}10$, принять однозначное решение сложнее: при небольшом увеличении объёма выборки, скорее всего, *P* станет меньше 0,05, но пока формально — больше. Учитывая договорной характер граничного значения 0,05, в таких ситуациях можно обсуждать тенденцию к наличию обсуждаемого эффекта (хотя, возможно, некоторые редакторы статей с этим и не согласятся).

> **Комментарий 1**. Как правильно писать: *p* или *P*? В литературе можно встретить оба варианта написания *p*-значения: строчное *p* и прописное *P*. Стандарты стилей American Medical Assosiation и American Psychological Assosiation рекомендуют прописное написание: *P*. Прописное написание несколько

чаще встречается в книгах, включая лучшие учебники по биостатистике Сокала и Рольфа и Зара [21; 22]. Карл Пирсон, впервые предложивший использовать *P* для оценки гипотез, использовал прописную P, а Рональд Фишер, разработавший концепцию проверки статистических гипотез, в посвящённой этому статье использовал строчную *p*. Строчное написание чаще встречается в журнальных публикациях и статистических выкладках самых известных статистических пакетов. Поэтому в прописном или строчном написании нет ошибки, но более академическим является прописное написание курсивом: *P*. В настоящем пособии используются оба варианта написания: *p* — при описании работы со статистическим пакетом PAST, где принято именно строчное написание, и *P* — в разделах, посвящённых оформлению результатов в публикациях и квалификационных работах.

Комментарий 2. Как правильно писать: «статистическая значимость» или «достоверность»? Во-первых, математическая статистика базируется на теории вероятности, где ***достоверным*** называется событие, вероятность которого равна 1. При принятии статистических решений мы не имеем ни невозможных событий с $P(E) = 0$, ни достоверных событий с $P(E) = 1$, поскольку всегда $0 < P < 1$. Мы можем лишь относиться к ним как к практически достоверным или практически невозможным исходя из выбранного ***уровня значимости*** α («альфа»). Использование уже занятого термина «достоверность» с другим смыслом некорректно. Во-вторых, концепция уровня значимости — исключительно английская разработка, а в английском языке используется термин «significance» — «значимость». Поэтому грамотным является употребление сочетания «статистическая значимость». Наукометрический, лингвистический и семиологический анализ некорректного использования термина «достоверность» — см. в работе [4].

③ Выписываем значение критерия и округляем его до сотых: $W = 0{,}93$. Какую букву использовать для критерия, программы нам часто подсказывают. Также выписываем объём выборки: $n = 49$. Далее выписываем соответствующее значение p и округляем до тысячных, поскольку **трёх знаков после запятой достаточно для самых строгих выводов**: $p = 0{,}005$. Если число p очень маленькое, например, как в критерии Андерсона — Дарлинга и его рандомизационном варианте (соответственно: 0,0001438 и 0,0003), то обычно достаточно записать просто $p<0{,}001$ (хотя это и не совсем правильно — см. теоретическую часть о проблемах ***синтетического подхода*** к проверке статистических гипотез).

④ **Предварительный вывод 2**: обнаружено высоко статистически значимое отличие распределения длины стопы восточноевропейской полёвки от нормального: критерий Шапиро — Уилка $W_{(49)} = 0{,}93$; $p = 0{,}005$.

* * *

Таким образом, и графический анализ, и использование статистического критерия указало на отличие распределения нашего признака от нормального. **Задание**. Подумаете, чем может быть вызвано это отличие. Порассуждайте как биологи: исходя их своих знаний о полёвках, длине стопы и т. д.

Обычно студенты высказывают следующие гипотезы:

1. Истребление хищниками наиболее активных средневозрастных животных. Например, самые маленькие (малая длина стопы) и самые старые (большая длина стопы) животные сидят преимущественно в норах, а бегают и истребляются хищниками преимущественно животные среднего возраста со средней длиной стопы. Гипотеза отпадает, поскольку, во-первых, эти животные были получены в виварии, а во-вторых, образ жизни полёвок иной.

2. Сильные возрастные различия: первый пик распределения — молодые животные, второй пик — более возрастные. Возраст действительно может обусловить резкую неоднородность размеров признака, но только в случае развития с метаморфозом. То есть если бы полёвки осенью окукливались, а по весне скидывали старую шкуру, мы бы наблюдали резкий скачок в размерах, в том числе и в размерах стопы. Но у млекопитающих развитие протекает без метаморфоза.

3. Мутация, обусловливающая неоднородность в размере стопы. Родители этих животных действительно были завезены в виварий из района Тоцкого радиоактивного следа (место в Красногвардейском районе Оренбургской области, где в 1954 г. были проведены тактические общевойсковые учения «Снежок» с реальным применением атомного оружия). Однако крайне маловероятно, что в этой популяции закрепилась такая странная мутация.

4. Половой диморфизм длины стопы. Может оказаться, что самцы и самки отличаются размерами стопы и наблюдаемое распределение — смесь распределений животных разного пола. Из всех возможных гипотез наиболее разумной является именно эта гипотеза: она вполне укладывается в наши представления о развитии млекопитающих, половом диморфизме размеров стопы, свойственного, кстати, и человеку. А значит, в отсутствие дополнительных данных рационально придерживаться именно такой гипотезы.

На самом деле наша выборка действительно состояла из самцов и самок, а полёвкам свойствен половой диморфизм размеров стопы: как и у человека, женские особи имеют меньшие размеры стопы. Этот пример был выбран для того, чтобы показать, насколько важно проводить анализ распределения, какую информацию он может дать. На предыдущих занятиях мы научились рассчитывать показатели описательной статистики и строить графики. Но можно ли отнести эти результаты к восточноевропейской полёвке? Предположим, мы захотим сравнить этот вид с другим видом, но в нашей выборке окажется больше самок, а в выборке другого вида — самцов. Различия в средних значениях будут обусловлены не только, а может быть, и не столько межвидовыми различиями размеров, но и соотношением полов в выборках. Таким образом, не разбив выборки по полу (расслоение, или ***стратификация*** выборки) и не сравнив самок с самками, а самцов с самцами, мы не сможем сделать никаких определённых выводов о различиях видов. **ВАЖНО!** Поэтому, если объём выборки позволяет пытаться строить распределение (30 и более наблюдений), это всегда необходимо делать, чтобы обнаружить возможную неоднородность, попытаться её объяснить и, по возможности, устранить для дальнейшего анализа.

Этап III. Разделение смеси распределений

Поскольку неоднородность распределения признака получила хорошее биологическое объяснение, можно попытаться разделить распределения самцов и самок статистически. В пакете PAST для этого есть ***передовая****, или* ***продвинутая*** (*advanced*) процедура, выполняемая по современному ***EM-алгоритму***.

> **К сведению**. ► **EM-алгоритм** (*Expectation-maximization (EM) algorithm*) — алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, зависящих от некоторых скрытых переменных. Каждая ***итерация*** алгоритма состоит из двух шагов. На E-шаге (expectation) вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. На M-шаге (maximization) вычисляется оценка максимального правдоподобия, таким образом, увеличивается ожидаемое правдоподобие, вычисляемое на E-шаге. Затем это значение используется для E-шага в следующей итерации. Алгоритм выполняется до сходимости.

В пакете PAST

① Файл «Длина стопы.dat» открыт, колонка значений выделена.

② Путь: Model — Mixture analysis (Анализ смеси).

По умолчанию: Distribution — Normal (нормальное распределение).

По умолчанию: Groups=2, то есть алгоритм постарается разделить распределение на две группы. В данном случае нас это устраивает.

Выставляем число классов Bins=7 — только для удобства графического восприятия, поскольку разделение смеси не зависит от этого нашего выбора. Можно также поставить галочку в ☑ Kernel density, чтобы убедиться в бимодальности распределения. Если число мод будет больше — следует пытаться разделить смесь на большее число распределений, изменяя количество групп (Groups). Рисунок можно сохранить, доработать через [Graph settings] и/или TpX и вставлять в работу.

③ Переходим на закладку Numbers и смотрим ***параметры*** разделённых программой распределений (среднее — Mean и стандартное отклонение St dev) и вероятную (Prob) долю этого распределения в смеси (доли единицы удобнее умножить на 100 и получить результат в процентах).

Теперь мы можем обоснованно предполагать, что выборка состояла из 59,9 % самок, со средней длиной стопы 15,9 мм, и 40,1 % самцов, со средней длиной стопы 17,1 мм.

④ **Оформление в квалификационной работе (вариант)**.

4.1. Статистическая часть раздела «Материал и методы».

Для оценки однородности выборки и проверки распределения на нормальность использовали графический анализ гистограмм распределения и статистический критерий Шапиро — Уилка. Количество классов для построения гистограмм рассчитывали по формуле Стургеса, плотность распределения определяли методом Сильвермана, а разделение смеси распределений проводили по EM-алгоритму. Статистически значимым считали отклонение от нормального распределения при $P \leq 0{,}05$. Расчёты и графические построения выполнены в пакетах PAST (version 3.19, Hammer et al., 2001) и TpX (Дать ссылку на источник).

4.2. Раздел «Результаты и обсуждение».

Распределение длины стопы восточноевропейских полёвок высоко статистически значимо отличалось от нормального: критерий Шапиро — Уилка $W_{(49)} = 0{,}93$; $p = 0{,}005$. Как видно из рис. 4.2, оно было бимодальным и, вероятно, представляло собой смесь двух близких к нормальному распределений.

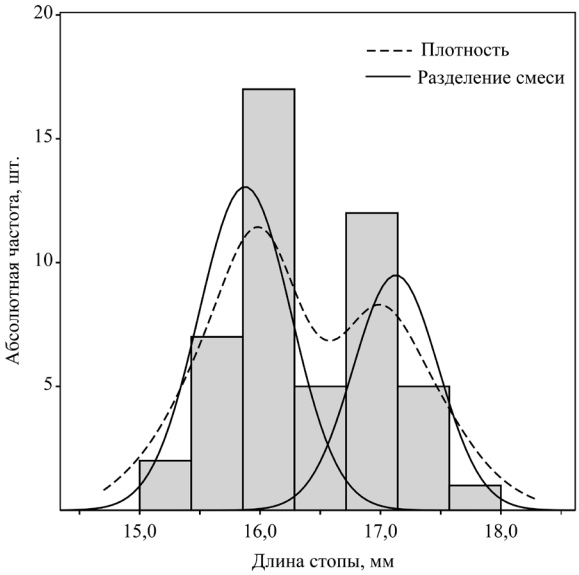


Рис. 4.2. Гистограмма, плотность распределения и разделение смеси распределений длины стопы восточноевропейской полёвки

Разделение смеси распределений позволило установить, что выборка состояла из двух групп животных: 59,9 % её составляли животные со средней длиной стопы 15,9 мм, а 40,1 % — с длиной стопы 17,1 мм. Полагаем, что первая группа была сформирована самками, а вторая — самцами полёвок. Действительно, из литературы известно, что… (далее про половой диморфизм размеров у млекопитающих, желательно — у полёвок, желательно — конкретно у восточноевропейских полёвок).

4.3. Раздел «Выводы».

Распределение длины стопы полёвок было отчётливо бимодальным и высоко статистически значимо отличалось от нормально-

го: критерий Шапиро — Уилка $W_{(49)} = 0{,}93$; $p = 0{,}005$. Наиболее вероятной причиной этого был половой диморфизм размеров стопы восточноевропейской полёвки.

> **Комментарий 3**. В научных публикациях выводы иногда подкрепляют статистическими выкладками с указанием *p*-значения, иногда — нет. Чаще информация о результатах статпроверки приводится и обсуждается в разделах, предшествующих выводам, исходя из цели исследования. Здесь и далее в практикуме выводы всегда содержат статистические выкладки, поскольку наша цель — изучение методов биостатистики, тогда как в предметных областях, откуда выбраны примеры, мы специалистами не являемся.

ЛАБОРАТОРНАЯ РАБОТА № 5

Сравнение двух независимых выборок по количественным и порядковым показателям

Тема 7. Выборочные сравнения для случая двух групп.
Количество часов: 2.
Цель: Освоить стратегию выбора статистических критериев для сравнения двух групп с применением критериев Снедекора — Фишера и Левене. Научиться использовать *t*-критерий Стьюдента (в том числе в модификации Уэлча) и критерий Манна — Уитни. Работа на ПК, решение задач.

Сравнение двух выборок — очень распространённая в исследовательской практике задача. Обычно одна выборка является экспериментальной или опытной (в медицине — «основная группа») и сравнивается со второй — контрольной (в медицине — «группа сравнения»). Также это могут быть выборки особей разного пола, разных видов и т. д. Отметим, что методы для сравнения двух выборок не подходят для ***попарных сравнений*** нескольких выборок (см. теоретический материал); также нужно отличать ***независимые выборки*** от ***зависимых***. Методы сравнения нескольких независимых выборок, а также зависимых выборок будут рассмотрены позже.

Объёмы сравниваемых выборок могут отличаться — вопреки распространённому заблуждению это вовсе не является препятствием для анализа. Иногда исследователи специально делают большую контрольную выборку, так как: а) её проще набрать, б) контроль может пригодиться для дальнейших исследований. За счёт большого контроля увеличивается мощность анализа и становится возможным обнаружить изменения в небольшой экспериментальной группе. Иногда, напротив, небольшой контроль призван служить лишь ориентиром границ условной нормы, а большая экспериментальная группа позволяет исследовать явление во всём многообразии контролируемых и неконтролируемых факторов (например, в медицине: пол, возраст, сопутствующие заболевания, профессиональные вредности и др., в экологии — химические и физические факторы среды).

Все методы для сравнения двух выборок делятся на ***параметрические*** (*parametric*), которые задействуют в расчётах парамет-

ры нормального распределения (математическое ожидаемое μ и стандартное отклонение σ), и ***непараметрические*** (*nonparametric*). Также важно помнить, что сравнение мы можем проводить с целью обнаружения различий: 1) центральной тенденции (наиболее частая задача); 2) рассеяния; 3) формы распределения. На этом лабораторном занятии мы познакомимся с критериями оценки только центральной тенденции для разных шкал данных.

I. Количественные признаки с нормальным распределением

Информация о нормальности распределения берётся из литературы, предыдущих исследований или проверяется непосредственно по данным, если позволяет объём выборки ($n \geq 30$). Если данные распределены ненормально, можно попытаться их нормализовать с помощью подходящих преобразований (логарифмирование, преобразование Бокса — Кокса, угловые преобразования для частот и др.).

Для сравнения средних значений показателя в выборках, извлечённых из популяций с нормальным распределением признака, используется параметрическая техника — варианты ***t-критерия Стьюдента*** (*Student's t-test*):

1. Классический или обычный *t*-критерий для независимых выборок. Требует равных дисперсий признака в популяциях.

2. *t*-критерий в модификации Уэлча (***критерий Уэлча***, *Welch's t-test*). Используется для сравнения средних значений независимых выборок в случае различия дисперсий.

3. *t*-критерий для сравнения единственного наблюдения с выборкой.

П р и м е р. Изучалось генотоксическое действие нового инсектицида (опыт) по сравнению с препаратом предыдущего поколения (контроль). Мух обрабатывали препаратами, рассаживали 16 пар (8 — опыт, 8 — контроль) в изолированные пробирки и подсчитывали число живых потомков. Получены следующие данные (в шт.):

Контроль	36	7	49	14	52	22	40	48
Опыт	10	6	3	17	18	22	5	39

Задача. Определить, отличаются ли препараты генотоксическим воздействием, то есть различаются ли они средним числом выживших потомков мух.

Решение. Поскольку данные представляют собой численности, есть основания сомневаться в нормальном распределении признака. Такие данные обычно распределены приблизительно логарифмически нормально, а потому *t*-критерий лучше использовать для логарифмов численностей. Проведём расчёт сначала для исходных данных, а затем самостоятельно — для преобразованных.

В пакете PAST

① Данные для разных групп вбиваются в соседние столбцы, столбцы именуются («Контроль» и «Опыт»), область значений выделяется.

② Путь: Univariate — Two-sample tests (F, t, …).

③ Сначала нам нужно выбрать, какой вариант *t*-критерия использовать: обычный для равных дисперсий или подход Уэлча для неравных дисперсий. Поэтому переходим на закладку $\overbrace{\text{F test}}$ и проверяем равенство дисперсий (*variance*) *F*-критерием Снедекора — Фишера (*Snedecor's F-test, Fisher's F-test, Fisher-Snedecor distribution*). Выписываем *F*, *p* для него, рассчитываем степени свободы как $df_1 = n_1 - 1$; $df_2 = n_2 - 1$, оформляем как $F_{(df_1;\, df_2)}$. Если *p* для *F*-критерия $\leq 0{,}05$, значит дисперсии отличаются статистически значимо и нужно использовать подход Уэлча; если $p > 0{,}05$, будем использовать классический *t*-критерий.

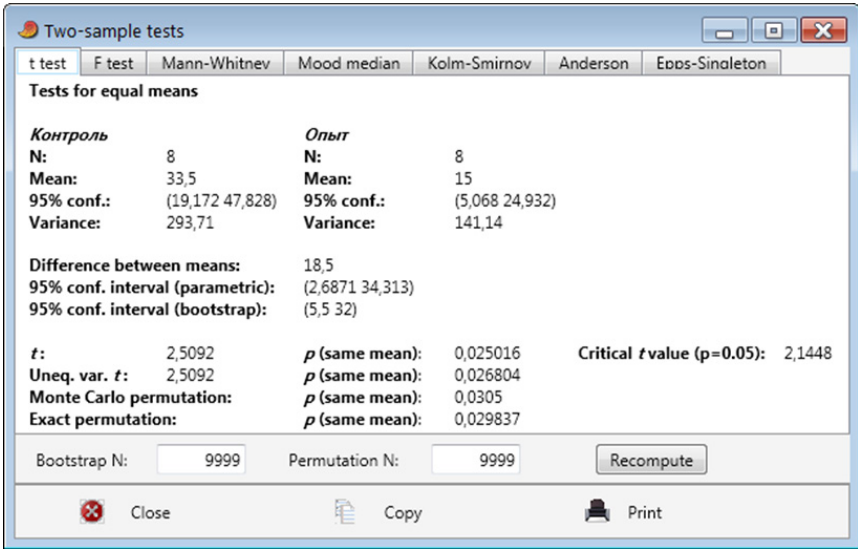
В нашем случае $F_{(7;\,7)} = 2{,}08$; $p = 0{,}355$. Поскольку $p > 0{,}05$, значит, дисперсии не различаются статистически значимо (**Внимание**! Не различаются именно значимо, хотя по самим значениям они отличаются более чем в два раза: 293,71 для контроля / 141,14 для опыта = 2,08). Поэтому для сравнения средних значений будем использовать обычный *t*-критерий.

④ Переходим на закладку $\overbrace{\text{t test}}$. Для обычного *t*-критерия выписываем *t* и *p* для него; степень свободы рассчитываем как $df = n_1 + n_2 - 2$.

Для модификации Уэлча выписываем Uneq.var.t (*Unequal variance t*), *p* для него; степень свободы рассчитываем как

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 + 1}} - 2,$$

где s^2 — дисперсия, n — объём выборки 1 (контроль) или 2 (опыт). Возможно, последующие версии PAST будут выдавать в результатах и ***степени свободы*** (*degree of freedom, df*), но пока (версия 3.19) их приходится считать вручную.



Если p для t-критерия $\leq 0{,}05$, значит средние значения отличаются статистически значимо; если $p > 0{,}10$ — не отличаются. В промежуточных случаях ($0{,}05 \leq p < 0{,}10$) можно обсуждать тенденцию к различиям или ориентироваться на результаты точного рандомизационного критерия — смотреть p для Exact permutation.

В нашем случае $df = 8 + 8 - 2 = 14$. $t_{(14)} = 2{,}51$; $p = 0{,}025$, то есть выборки отличаются статистически значимо.

Видно, что, кроме описательной статистики и t-критериев, пакет выдаёт также различия между средними с 95% ДИ — Difference

between means, которые можно использовать в качестве показателя ***величины эффекта*** (*effect size*). 33,5 – 15 = 18,5, то есть число потомков у мух, обработанных новым препаратом, было в среднем на 18,5 меньше, а значит, новый препарат был эффективнее, хотя — нужно признать — не намного.

⑤ **Оформление в квалификационной работе (вариант)**.

5.1. Раздел «Материал и методы».

Сравнение двух выборок по количественным признакам с нормальным распределением проводили с помощью *t*-критерия Стьюдента. В случае различий выборочных дисперсий использовался метод Уэлча. Эффекты считали статистически значимыми при при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к различиям. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

5.2. Раздел «Результаты и обсуждение».

Сравнение среднего числа потомков мух, обработанных новым и старым препаратами, проводили с помощью *t*-критерия Стьюдента. Для выбора нужного варианта этого критерия на первом этапе анализа проверяли равенство дисперсий в выборках. Было установлено, что они не различались статистически значимо: критерий Снедекора — Фишера $F_{(7;\,7)} = 2{,}08$; $P = 0{,}355$. Поэтому на втором этапе анализа для сравнения средних использовали классический вариант *t*-критерия для равных дисперсий. Также в разделе результатов следует дать таблицу с описательной статистикой и/или график.

5.3. Раздел «Выводы».

Обнаружено статистически значимое снижение числа потомков у мух, обработанных инсектицидом нового поколения: $t_{(14)} = 2{,}51$; $P = 0{,}025$. Новый препарат снижал число потомков в среднем на 18,5 (95% ДИ: от 5,5 до 32,0) мух больше, чем старый.

II. Количественные признаки с ненормальным распределением и порядковые признаки

Вариантов анализа в этом случае много, рассмотрим наиболее популярные и современные.

Способ 1. *t*-критерий Стьюдента после нормализующего преобразования. В большинстве случаев оптимальным преобразо-

ванием является преобразование Бокса — Кокса из семейства степенных преобразований. В пакете PAST путь: Transform — Box-Cox. Для данных, выраженных частотами, используют ***угловые преобразования*** (φ-преобразование арксинуса и др.), которых пока нет в PAST.

Способ 2. Рандомизационный вариант *t*-критерия Стьюдента. В пакете PAST путь тот же, закладка t test, в результатах нужно смотреть *p*-значение точного рандомизационного критерия — *Exact permutation*. Если пакет не выдаёт его значений — смотрим результаты рандомизационного критерия Монте-Карло — Monte Carlo permutation; при этом число перестановок можно увеличить с 9 999 до 99 999 или даже 999 999: при последовательных нажатиях на кнопку [Recompute] третий знак *p*-значения не должен изменяться. Само значение статистики *t* можно не приводить. При этом нужно знать философию рандомизационных критериев и уметь объяснить, почему такой вариант параметрического *t*-критерия может использоваться для сравнения и ненормально распределённых данных. В нашем случае $p = 0{,}030$.

> **К сведению**. ► Метод Монте-Карло — общее название группы численных методов, основанных на получении большого числа реализаций *стохастического* (случайного) процесса, который формируется таким образом, чтобы его вероятностные характеристики совпадали с аналогичными величинами решаемой задачи. В случае сравнения двух выборок алгоритм будет следующим. На этапе 1 рассчитывается интересующая статистика — например, *t*-критерий — для исходных выборок 1 и 2, объёмов n_1 и n_2. На этапе 2 значения обеих выборок смешиваются, и n_1 значений случайным образом назначаются в выборку 1, а оставшиеся n_2 значений — в выборку 2. (Формировать случайные выборки мы научимся на последней лабораторной работе № 18, см. рандомизацию.) Таким образом, сами числовые значения в анализе остаются такими же, как были в исходных данных, но их распределение между выборками изменяется на случайное. На этапе 3 для сгенерированных в результате случайных перестановок выборок рассчитывается интересующая статистика. Далее этапы 2 и 3 повторяются многократно, например, 9 999 раз. На последнем, этапе 4 проводится расчёт *p*, как доли случаев *k* среди $N = 9\,999$ значений, когда статистика была меньше или равна вычисленной по исходным данным на этапе 1: $p = k/N$ или по скорректированной формуле $p = (k+1)/(N+1)$, исключающей возможность $p = 0$. Это и есть *p*-значение, вычисленное методом Монте-Карло.
>
> Точное рандомизационное значение *p* получается сходным образом, однако генерируется не просто большое число различных случайных разбиений данных на 2 группы, а в точности все возможные разбиения. Для больших *n* это может оказаться непосильной задачей даже для современных компьютеров. В нашем примере для двух групп по 8 наблюдений таких вариантов

разбиения будет $16!/(8! \times 8!) = 12\,870$. Пакет PAST проводит Exact permutation вплоть до $(n_1 + n_2) < 27$. (см. Руководство к пакету).
Рандомизационные критерии можно считать непараметрическими, поскольку независимо от рассчитываемой статистики, для расчёта p вид распределения значения не имеет. Вместо t-критерия мы могли бы использовать другую статистику, например, просто разность средних значений, и получили бы близкое значение p.

Способ 3. По доверительному интервалу для разности средних, рассчитанному бутстрепом: если этот ДИ содержит 0, значит разность между средними может быть нулевая, то есть различий нет. Если 95% ДИ разности не содержит 0, средние отличаются статистически значимо ($p < 0{,}05$). Путь такой же, закладка t test, смотрим 95% conf. interval (bootstrap). Число выборок бутстрепа Bootstrap N можно увеличить.

Способ 4. Классические непараметрические критерии, которых разработано очень много. Наибольшей мощностью обладает ***критерий нормальных меток ван дер Вардена*** (*van der Waerden normal scores test*), однако наиболее популярен U-критерий Уилкоксона — Манна — Уитни, чаще называемый просто ***критерием Манна — Уитни*** (*Wilcoxon-Mann-Whitney test, Wilcoxon rank sum test, Mann-Whitney U-test*). Это прямой ранговый эквивалент t-критерия Стьюдента: если от параметрической статистики перейти к порядковой, то формула t-критерия станет монотонной функцией U-критерия [23]. Хотя он может быть выведен и из других теоретических построений:

1) из вероятностей отнесения наблюдения к одной из двух групп — как частный случай ***ридит-анализа*** (*ridit analysis*) для упорядоченных категорий;

2) из ROC-анализа диагностической эффективности с расчётом площади *AUC* под ROC-кривой (см. лабораторную работу № 14): $U = n_1 n_2 \times \text{AUC}$, если средний ранг первой выборки больше, чем второй ($\overline{R}_1 > \overline{R}_2$) или $U = n_1 n_2 \times (1 - \text{AUC})$, если $\overline{R}_1 < \overline{R}_2$.

Критерий обладает высокой мощностью: ***асимптотическая эффективность*** критерия составляет $3/\pi$, то есть около 95 %. Это означает, что он только на 5 % уступает в мощности t-критерию, однако не требует нормального распределения в популяции. Требования: 1) случайные независимые выборки и 2) распределения в обеих группах относятся к одинаковому типу (на практике обычно не проверяется).

П р и м е р. Рассчитаем критерий Уилкоксона — Манна — Уитни для тех же данных по выживаемости мух двумя способами: вручную — для лучшего понимания философии порядковых статистик — и в пакете PAST. Заполним следующую таблицу и опишем алгоритм расчётов:

Значение
Ранг R
Группа (К или О)
$\sum R_К$
$\sum R_О$

Алгоритм:

1. Значения из обеих групп одновременно выписываются в порядке возрастания; при этом отмечается ранг наблюдения и его принадлежность к группе контроля (К) или опыта (О). Например, минимальное значение было 3, оно стоит на первом месте (ранг 1) и относится к группе опыта (О).

Значение	3	5	6	7	10	14	17	18	22	22	36	39	40	48	49	52	
Ранг R	1	2	3	4	5	6	7	8	9,5	9,5	11	12	13	14	15	16	
Группа (К или О)	О	О	О	К	О	К	О	О	О	К	К	О	К	К	К	К	
$\sum R_К$				4		+6				+9,5	+11		+13	+14	+15	+16	=88,5
$\sum R_О$	1	+2	+3		+5		+7	+8	+9,5			+12					=47,5

Одинаковые значения получают средний ранг. Так, два значения 22 делят между собой 9 и 10 места, то есть получают ранг (9 + 10) : 2 = 9,5.

2. Рассчитываются суммы рангов контроля $\sum R_К$ и опыта $\sum R_О$. Таким образом, на этом этапе мы заменяем сами значения их порядковыми местами — рангами.

3. Рассчитывается U-статистика Манна — Уитни:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1 = 8\times 8 + \frac{8\times 9}{2} - 88{,}5 = 100 - 88{,}5 = 11{,}5;$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2 = 8\times 8 + \frac{8\times 9}{2} - 88{,}5 = 100 - 47{,}5 = 52{,}5.$$

Проверка: $U_1 + U_2 = n_1 \times n_2$;

$11{,}5 + 52{,}5 = 8 \times 8;$
$64 = 64$ — верно.

Меньшее из двух U является искомой статистикой: $U = 11{,}5$. Подчеркните или обведите его.

4. Сравнение полученного значения критерия с критическими значениями U-статистики для разных уровней значимости. Воспользуемся таблицей из учебника Л. Закса «Статистическое оценивание» (с. 272 и далее). Объёмы наших выборок 8 и 8. Открываем первую таблицу, смотрим двусторонний критерий, $\alpha = 0{,}20$. На пересечении значений 8 и 8 находим критическое значение — 19. Далее находим и выписываем значения для других уровней значимости α.

Уровень значимости, α	U критическое
0,20	19
0,10	15
0,05	13
0,02	9
0,01	7
0,002	4

$< U = 11{,}5$

Наше фактическое значение U находится между 9 и 13, а значит, соответствующее значение p находится между 0,02 и 0,05, то есть $0{,}02 < p < 0{,}05$. Поскольку $p < 0{,}05$, можем констатировать статистическую значимость различий.

В пакете PAST

① Данные уже введены и выделены.

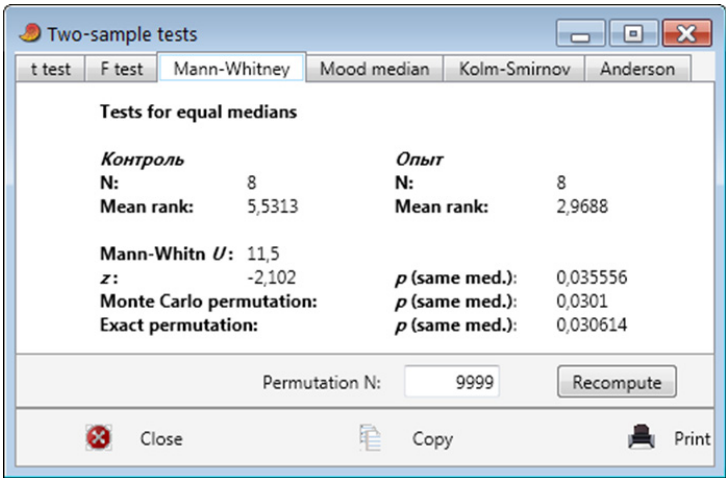
② Путь: Univariate — Two-sample tests (F, t, …).

③ Закладка Mann-Whitney. Выписываем значение U, округляем до десятых. Выписываем значение p, округляем до тысячных. Лучше взять Exact permutation или Monte Carlo permutation. Также выписываем объёмы выборок: 8 и 8; их поместим рядом с U.

Оформляем результат: $U_{(8,\,8)} = 11{,}5$; $p = 0{,}031$.

④ **Оформление в квалификационной работе**.

Проводится аналогично критерию Стьюдента. В работу можно дать таблицу с описательной статистикой и/или график. В зависимости от области биологии и медицины, а также личных предпочтений автора можно представить либо график средних



значений с ДИ, вычисленными бутстрепом, либо коробчатую диаграмму с порядковыми мерами (см. рис. 8.1).

Вывод: Обнаружено статистически значимое снижение числа потомков у мух, обработанных инсектицидом нового поколения: критерий Уилкоксона — Манна — Уитни $U_{(8,\,8)} = 11{,}5$; $P = 0{,}031$.

Домашнее задание. Как мы уже отметили, имеются теоретические основания сомневаться в нормальном распределении признака. Поэтому перед использованием параметрического *t*-критерия лучше исходные данные преобразовать. Поскольку численности обычно распределены приблизительно логарифмически нормально, логично использовать преобразование логарифма. Для этого нужно выделить данные и пройти по пути: Transform — Log. Если способ нормализации данных неясен из теории, используют степенное ***преобразование Бокса — Кокса*** (*Box-Cox transformation*), которое нормализует данные настолько, насколько они сами это позволяют (см. теоретический материал): Transform — Box-Cox.

Преобразуйте данные примера с мухами с помощью обоих преобразований. Посмотрите, как изменилась оценка равенства дисперсий в *F*-критерии и результаты сравнения *t*-критерием. Оформите результаты.

ЛАБОРАТОРНАЯ РАБОТА № 6

Сравнение двух независимых выборок по качественным показателям

Тема 7. Выборочные сравнения для случая двух групп.
Количество часов: 2.
Цель: Овладеть методами анализа различий между выборками по качественным показателям в ходе анализа таблиц сопряжённости. Научиться представлять результаты анализа с использованием статистики типа хи-квадрат, относительных рисков и отношений шансов.

Качественные признаки часто встречаются в биологии и медицине. Такие признаки представлены категориями, например: вид животного, форма листа, цвет венчика, наличие или отсутствие симптома. Чаще всего их удобно подсчитывать и представлять в процентах от общего числа. Варианты анализа:

I. Если категории можно упорядочить (например: мало — средне — много), то две группы можно сравнить критерием Уилкоксона — Манна — Уитни. Для этого категории «мало» присваивается ранг 1, «средне» — 2, «много» — 3. Далее таблица частот разворачивается в длинную форму.

Например, такую таблицу

	Мало	Средне	Много
Группа 1	2	2	2
Группа 2	0	1	5

следует переписать в такую:

Группа 1	Группа 2
1	2
1	3
2	3
2	3
3	3
3	3

Далее проводится сравнение, как мы делали на предыдущем занятии.

ВАЖНО! В результате такого анализа будет учтено наличие упорядоченности и статистическое сравнение получится более мощным. Именно таким образом сравнивает упорядоченные категории лидирующий по точным вычислениям пакет StatXact от компании Cytel.

II. Если категории упорядочить нельзя, то есть если данные представлены номинальной шкалой, анализ проводят ***критериями согласия*** или современными ***рандомизационными критериями*** в ходе анализа ***таблиц сопряжённости*** (*ТС*, *contingency table*). Методов анализа ТС предложено много; перечислим основные из них:

1) ***критерий хи-квадрат Пирсона*** (*Pearson's Chi-square test*), обозначается χ^2 Пирсона или просто χ^2. В некоторой статистической литературе обозначается как X^2 («икс-квадрат») — для подчёркивания отличий от теоретического статистического распределения хи-квадрат. Предложен Карлом Пирсоном ещё в 1901 г., но до сих пор популярен. Есть во всех статистических пакетах;

2) ***критерий Фримана — Тьюки*** (*Freeman-Tukey test*). Сам критерий малоизвестен, но ***отклонения Фримана — Тьюки*** (*Freeman-Tukey deviations*), основанные на той же статистике, используются для углублённого анализа больших таблиц сопряжённости;

3) ***критерий отношения правдоподобия*** (иногда обозначают Θ, Λ, *likelihood ratio test*). Также встречается в литературе под другими названиями: *G*-критерий Вулфа, критерий G^2 («джи-квадрат»), информационный критерий Кульбака I^2, хи-квадрат максимального правдоподобия χ^2_{ML} и др. **Вопрос**: почему один и тот же критерий имеет столько названий? Данный критерий многократно переоткрывался, причём исходя из разных теоретических построений. Таким образом, в отличие от χ^2 Пирсона он отлично обоснован теоретически и является его более современным аналогом. Сокал и Рольф — авторы одного из лучших в мире учебников по биостатистике — рекомендуют всегда использовать *G*-критерий вместо χ^2 Пирсона.

Все три перечисленных критерия имеют ***теоретическое распределение χ^2*** (см. теоретический материал). Для всех трёх критериев существует проблема *допустимого минимального ожидаемого:* если в таблице есть ячейки с малыми ожидаемыми (примерно меньше 4), статистика критериев плохо аппроксимируется

распределением χ^2. На практике, если конкретный статпакет не выдаёт в результатах таблицу ожидаемых частот, то можно ориентироваться так: если в таблице есть значения от 0 до 5 включительно, то использовать эти критерии некорректно. Раньше для анализа таких ***слабонасыщенных таблиц*** применялся точный метод Фишера;

4) ***точный метод Фишера*** (ТМФ, *Fisher's exact test*) предложен Р. Фишером в 1954 г. для анализа слабонасыщенных таблиц и до сих пор популярен. Однако теоретически он не очень хорош: критерий основан на гипергеометрическом распределении, хотя используется для анализа ТС с данными, имеющими биномиальное или полиномиальное распределение. В настоящее время вместо него корректнее пользоваться рандомизационными критериями;

5) ***рандомизационный критерий Монте-Карло*** (*permutation test*, Monte Carlo test), случайным образом генерирует большое число (десятки и сотни тысяч) ТС с такими же краевыми частотами, как у исходной. Доля таблиц со значением статистики, меньшим или равным наблюдаемой от общего числа сгенерированных таблиц, и есть p-значение: $p = k / N$; или по скорректированной формуле $p = (k + 1) / (N + 1)$;

6) ***точный рандомизационный (перестановочный) критерий*** (*Exact permutation test*) — похож на 5), но генерируются не случайные таблицы с такими же краевыми частотами, а в точности все возможные. Для ТС с большим числом наблюдений это может быть непосильной задачей даже для современных компьютеров, и тогда приходится использовать предыдущий критерий. Точный рандомизационный критерий — наиболее точный и современный метод, который рекомендуется использовать во всех случаях, а особенно — для анализа слабонасыщенных таблиц. Он есть в продуктах компании Cytel (StatXact и LogXact); также по лицензии их алгоритм расчёта используется в пакете SPSS. В пакете PAST есть 1-й, 4-й и 5-й методы. Лучший из них — 5-й: рандомизационный критерий Монте-Карло.

Если с помощью перечисленных критериев обнаруживаются различия, то далее обычно рассчитываются показатели силы различий (величины эффекта): ***разность рисков***, ***относительный риск*** или ***отношение шансов***.

П р и м е р. У пациентов клиники определялся уровень общего холестерина в крови. Все измерения были разбиты на две категории: 1) до 6,72 ммоль/л (260 мг/дл) включительно — «норма»; 2) свыше 6,72 ммоль/л — «повышенный» уровень. Параллельно отмечалось наличие заболеваний сердечно-сосудистой системы (ССС). **Вопросы**: отличаются ли лица с высоким и нормальным холестерином частотами заболеваний ССС? Если отличаются, то насколько сильно?

Данные:

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Зарисуйте в тетради эту таблицу и выделите четыре центральные ячейки собственно данных. Такая простейшая ТС называется *таблицей 2×2* («два на два») или *четырёхпольной таблицей*. В ней суммы по столбцам и строкам называются *краевыми частотами*, а общее число наблюдений — *общей суммой*.

1. Расчёт относительных частот

Повышенный холестерин. Доля больных равна: 41 / 286 = 0,143, или 14,3 %.

Нормальный холестерин. Доля больных равна: 51 / 1 043 = 0,049, или 4,9 %.

Доля каких-либо интересующих событий в выборке называется в биостатистике ***риском***; то есть можно сказать, что риск заболеваний ССС в группе с повышенным холестерином составил 0,143, а в группе с нормальным — 0,049.

Таким образом, доля пациентов с заболеваниями ССС была выше в группе с повышенным уровнем холестерина. Необходимо убедиться, что эти два значения различаются статистически значимо, то есть **речь идёт о сравнении двух процентов**. **ВАЖНО!** Если необходимо сравнить два процента, а абсолютные частоты не заданы, эти частоты нужно рассчитать из процентов и объёмов выборок, а далее для анализа свести в ТС.

2. Сравнение двух частот с помощью критерия

Познакомимся подробнее с критерием χ^2 Пирсона: рассчитаем его вручную и в пакете PAST.

Алгоритм:

2.1. Расчёт ***ожидаемых частот*** *(expected frequencies).*

Проводится в предположении отсутствия различий между группами, то есть считается, что данные в ячейках таблицы 2×2 являются простым наложением двух отношений: доли больных и здоровых людей в популяции и доли людей с высоким и нормальным холестерином.

Так, по краевым суммам вычислим долю больных людей в популяции как 92 / 1329. Значит, в группе с повышенным холестерином должно наблюдаться 286 × 92 / 1 329, а в группе с нормальным холестерином — 1043 × 92 / 1329 больных людей. На практике расчёт вручную удобно проводить по формуле

$$\hat{f} = \frac{\Sigma \text{ по строке} \times \Sigma \text{ по столбцу}}{\Sigma \text{ общая}}.$$

Значок «крыша» означает, что данное теоретическое значение вычислено по выборке. Для первой ячейки таблицы (строка 1, столбец 1) и далее имеем:

$\hat{f}_{11} = 286 \times 92 / 1\,329 = 19{,}79834462 \approx 19{,}8$ (округлим до десятых);

$\hat{f}_{21} = 1\,043 \times 92 / 1\,329 = 72{,}2$;

$\hat{f}_{12} = 286 \times 1\,237 / 1\,329 = 266{,}2$;

$\hat{f}_{22} = 1\,043 \times 1\,237 / 1\,329 = 970{,}8$.

Сводим полученные значения в таблицу ожидаемых частот:

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	19,8	266,2	286
Норма	72,2	970,8	1043
Всего	92	1237	1329

Сравните полученную таблицу с исходной. Обратите внимание, что таблица ожидаемых частот имеет такую же общую сумму и такие же краевые частоты, как исходная, однако сами частоты внутри соответствуют нулевой гипотезе — отсутствию различий между выборками.

2.2. Вычисление критерия χ^2 Пирсона.

Критерий оценивает согласие наблюдаемых и ожидаемых частот. Вы, вероятно, уже знакомились с ним в курсе генетики, когда оценивали согласие расщепления менделеевского признака по фенотипам во втором поколении (3 : 1 или 9 : 3 : 3 : 1). Ожидаемые частоты вы рассчитывали иначе, но сам критерий — тот же самый, формула — та же:

$$\chi^2 = \sum \frac{(f_{\text{наблюдаемая}} - \hat{f}_{\text{ожидаемая}})^2}{\hat{f}_{\text{ожидаемая}}}.$$

$\chi^2 = (41 - 19{,}8)^2 / 19{,}8 + (245 - 266{,}20)^2 / 246{,}2 + (51 - 72{,}2)^2 / 72{,}2 + + (992 - 970{,}8)^2 / 970{,}8 = \underline{22{,}70} + 1{,}69 + 6{,}22 + 0{,}46 = 31{,}07.$

Видно, что в значение критерия наибольший вклад внесла первая ячейка (подчеркните значение 22,70), то есть у людей с высоким холестерином больных было намного больше, чем ожидалось в соответствии с нулевой гипотезой.

2.3. Расчёт степеней свободы:

$$df = (n_{\text{строк}} - 1)(n_{\text{столбцов}} - 1);$$
$$df = (2 - 1)(2 - 1) = 1.$$

2.4. Оценка статистической значимости.

Полученное значение χ^2 при нужном числе степеней свободы сравнивается с табличным [1. С. 134].

Уровень значимости α (двусторонний)	χ^2 критическое
0,05	3,84
0,01	6,63
0,001	10,83
	31,07

Наше значение оказалось намного больше 10,83, а значит, P намного меньше 0,001. Для таких случаев можем воспользоваться значком *«много меньше»* <<. Таким образом, имеем:

$\chi^2_{(1)} = 31{,}07$; $P << 0{,}001$ (различия высоко статистически значимы).

> **К сведению**. ***Поправка Йейтса на непрерывность*** (*Yates' continuity correction*). При расчёте критерия хи-квадрат Пирсона задействуются дискретные величины — частоты, однако теоретическое статистическое рас-

пределение хи-квадрат — непрерывное. Это приводит к неточности, которая будет тем больше, чем меньше объём выборки. Для её коррекции в таблицах с общей суммой ≤20 ранее использовали поправку, предложенную Фрэнком Йейтсом: уменьшали каждую разность между наблюдаемой и ожидаемой частотами в формуле на 0,5. Данный подход всегда критиковался за излишнюю консервативность. Менее консервативной является ***поправка Уильямса*** (*Williams' correction*), которая применяется обычно к *G*-критерию (в некоторых пакетах — по умолчанию). В настоящее время подход с введением поправок можно считать устаревшим, поскольку современные рандомизационные техники, рекомендуемые для анализа слабонасыщенных таблиц, не задействуют теоретическое распределение хи-квадрат при расчёте *P*, а следовательно, не нуждаются в поправках.

В пакете PAST

① Ввести четыре значения данных в соседние ячейки и выделить:

41	245
51	992

② Путь: Univariate — Contingency table. (Если общая сумма велика, пакет не может вычислить точный критерий Фишера, о чём сообщает в окне предупреждения; закройте его.)

③ Выписываем значение критерия (Chi^2), степени свободы (degrees of freedom), *p*. Если в таблице есть значения 5 и менее — выписываем *p*, вычисленное рандомизационным критерием Монте-Карло.

ВАЖНО! Во многих статистических пакетах используется ***экспоненциальная форма*** записи чисел. $p = 2{,}4738\text{E}{-}08$ значит $2{,}4738 \times 10^{-8}$. Столь малое число можно записать как $p < 0{,}001$ или $p << 0{,}001$. Видим, что, несмотря на ошибки округления, ручной расчёт мы провели достаточно точно: 31,07 против вычисленного на компьютере 31,08.

Вывод краткий: пациенты с повышенным и нормальным уровнем холестерина в сыворотке крови высоко статистически значимо различались частотами заболеваний сердечно-сосудистой системы: критерий хи-квадрат Пирсона: $\chi^2_{(1)} = 31{,}08$; $P << 0{,}001$.

3. Оценка величины различий

Относительные частоты, рассчитанные в п. 1, указали нам, в какой группе частота заболеваний была выше. Критерий хи-квадрат указал на то, что различия между группами пациентов

были статистически значимы, то есть, вероятно, неслучайны. Теперь необходимо оценить, насколько же сильны обнаруженные различия. В качестве показателей ***величины эффекта*** (*effect size*) для различий частот используется несколько мер.

3.1. ***Разность рисков*** (*Risk difference*).

Показывает, насколько риск события в одной группе больше или меньше по сравнению с риском в другой. Рассчитывается как простая арифметическая разность рисков, рассчитанных в п. 1. В нашем случае она равна: 0,143 – 0,049 = 0,094.

3.2. ***Отношение рисков*** (или относительный риск, *Risk ratio*, *Relative risk — RR*).

Показывает, во сколько раз риск (частота) события в одной группе больше или меньше по сравнению с риском в другой. Для равных рисков $RR = 1$. В нашем случае

$$RR = 0{,}143 / 0{,}049 = 2{,}92.$$

Это очень удобная для понимания и интерпретации мера: с увеличением содержания холестерина в сыворотке крови до 6,72 ммоль/л риск заболеваний ССС увеличивается в 2,92 раза.

3.3. ***Отношение шансов*** (*Odds ratio* — OR).

Показывает, во сколько раз шанс события в одной группе больше или меньше по сравнению с шансом в другой. ► **Шанс** — отношение вероятности события к его альтернативе. В нашем случае при повышенном холестерине вероятность иметь заболевания ССС составляет 41 / 1 329, а не иметь (альтернатива) — 245 / 1 329. Таким образом, шанс иметь заболевания ССС при высоком холестерине составляет

$$\frac{41}{1\,329} : \frac{245}{1\,329} = \frac{41}{245} = 0{,}16735.$$

Знак «:» читается «к», то есть шанс составляет сорок один к двумстам сорока пяти. Аналогично шанс иметь заболевания ССС при нормальном холестерине составляет пятьдесят один к девятистам девяноста двум: 51 / 992 = 0,0514. Следовательно, отношение шансов составляет:

$$OR = 0{,}16735 / 0{,}05141 = 3{,}26.$$

Интерпретация: с увеличением содержания холестерина в сыворотке крови до 6,72 ммоль/л шансы заболеваний ССС увеличиваются в 3,26 раза.

Данная мера не столь понятна, как отношение рисков, но в последние два десятилетия стала очень популярной благодаря использованию в другом статистическом методе — ***множественной логистической регрессии***, где коэффициенты регрессии легко пересчитываются в отношения шансов. С логистической регрессией мы будем знакомиться на лабораторной работе № 13.

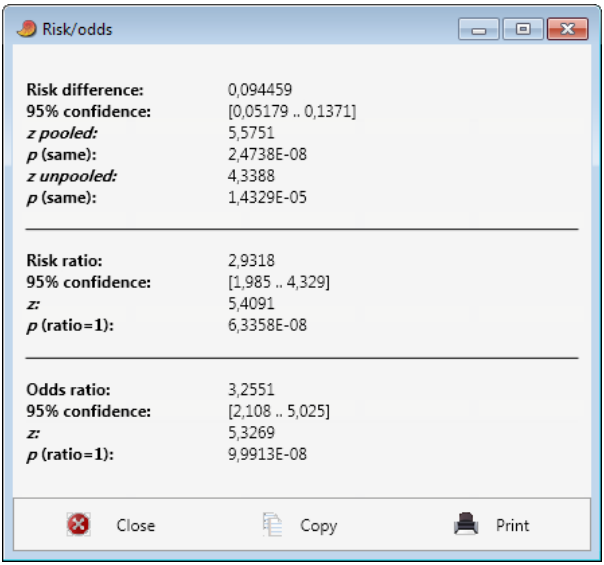
В пакете PAST

① Четыре значения введены в соседних ячейках и выделены:

41	245
51	992

② Путь: Univariate — Risk/Odds.

③ В окне результатов видим все три меры, а также 95% ДИ для них. Пакет выдаёт ещё и значения *p*, расчёт которых возможен без опоры на статистические критерии типа хи-квадрат, а с использованием стандартного нормального распределения (*z*-критерий). Для нас они не важны, поскольку вывод о различии частот заболеваемости мы проводили не по данным оценкам величины эффекта, а с использованием критерия хи-квадрат Пирсона. Поэтому выписываем только необходимую меру (все три приводить не следует) с 95% ДИ.



Внимание! Очень ВАЖНО! Для правильного расчёта пакетом рисков и шансов необходимо, чтобы данные были организованы в таблице именно так, как у нас:
в строке 1 — группа для которой проводится оценка рисков/шансов,
в строке 2 — контрольная группа, относительно которой проводится оценка;
в колонке 1 — наличие интересующего признака,
в колонке 2 — его отсутствие.

При другом расположении значения в таблице будут соотнесены неправильно!

④ **Оформление в квалификационной работе (вариант)**.

4.1. Статистическая часть раздела «Материалы и методы».

Сравнения двух групп по качественным номинальным показателям проводили в ходе анализа таблиц сопряжённости критерием хи-квадрат Пирсона. Для слабонасыщенных таблиц (имелись ячейки со значениями ≤5), оценку статистической значимости проводили с помощью рандомизационной процедуры Монте-Карло. В качестве показателя величины эффекта рассчитывали относительные риски *RR* с 95% ДИ.

Различия считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к различиям. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

4.2. Раздел «Результаты и обсуждение».

Даются таблицы с абсолютными (в штуках) и относительными (в процентах) частотами. Последние желательно снабдить 95% ДИ, вычисленными по Джеффрису, Вилсону, Агрести — Коулу или Клопперу — Пирсону (см. лабораторную работу № 2). В квалификационную работу нужно включить и результаты статистического сравнения. Можно сделать столбчатые диаграммы с 95% ДИ.

4.3. Раздел «Выводы».

Пациенты с повышенным и нормальным уровнем холестерина в сыворотке крови высоко статистически значимо различались частотами заболеваний сердечно-сосудистой системы: критерий хи-квадрат Пирсона: $\chi^2_{(1)} = 31{,}08$; $P << 0{,}001$. Для лиц с содержанием холестерина в сыворотке крови 6,72 ммоль/л и выше относительный риск заболеваний ССС составил 2,93 (95% ДИ: от 1,99 до 4,33).

ЛАБОРАТОРНАЯ РАБОТА № 7

Сравнение двух зависимых выборок

Тема 7. Выборочные сравнения для случая двух групп.
Количество часов: 2.
Цель: Овладеть методами анализа различий между зависимыми выборками по количественным, порядковым и качественным показателям с помощью парных критериев Стьюдента, Уилкоксона и Макнемара. Познакомиться с работой онлайновых статистических калькуляторов. Работа на ПК.

В теоретической части курса мы рассматривали такую характеристику выборок, когда по способу включения объектов они могут быть ***независимыми*** (*independent samples*) или ***зависимыми*** (*paired samples, dependent samples*). На предыдущих лабораторных работах мы имели дело только с независимыми выборками, когда объекты в двух сравниваемых группах не были никак связаны друг с другом. На этом занятии познакомимся с анализом зависимых выборок.

Чаще всего зависимые выборки образуются одними и теми же объектами, изученными в разное время и/или в разных условиях. Например, одни и те же лабораторные животные, изученные до воздействия и после воздействия. В таком экспериментальном плане каждое животное будет иметь своё собственное контрольное значение. Другой распространённый пример зависимых выборок — части одного образца, исследованные разными методами. Для проверки некоторых гипотез пары могут образовывать разные объекты, например, близнецы, братья и сёстры, мужья и жёны, а также специально подобранные сходные индивиды для исследований типа «случай—контроль».

Организация зависимых выборок позволяет провести более экономное и/или мощное исследование с возможностями более широкой интерпретации данных (см. теоретический материал и лабораторную работу № 18).

I. Количественные признаки с нормальным распределением

Информация о нормальности распределения берётся из литературы, предыдущих исследований или проверяется непо-

средственно по данным, если позволяет объём выборки ($n \geq 30$). Если данные распределены ненормально, можно попытаться их нормализовать с помощью подходящих преобразований (логарифмирование, преобразование Бокса — Кокса, угловые преобразования для частот и др.).

Для сравнения средних значений показателя в двух зависимых выборках для признаков, изменяющихся по закону нормального распределения, используется параметрическая техника — ***парный t-критерий Стьюдента*** (*matched-pair t-test*, *paired sample t-test*).

П р и м е р. Для большинства показателей, используемых в медицинской диагностике, стандартные методики предусматривают анализ венозной крови. Вместе с тем современное аналитическое оборудование позволяет работать с очень небольшими объёмами образцов, которые можно получить из капиллярной крови, взятой из пальца пациента. В силу простоты и удобства для пациента последнее было бы предпочтительным, если бы удалось доказать, что результаты анализов венозной и капиллярной крови не различаются.

В ходе небольшого эксперимента у 16 пациентов были отобраны образцы венозной и капиллярной крови, в которой определялся ряд биохимических показателей. Мы проанализируем данные по содержанию общего билирубина (ОБ) в сыворотке (в мкмоль/л). Данный показатель характеризует сумму промежуточных продуктов метаболизма гемоглобина и позволяет диагностировать различные заболевания, прямо или косвенно связанные с нарушением процессов кроветворения, функции печени и желчевыводящих путей.

Данные:

Венозная кровь	Капиллярная кровь
10,5	5,7
6,6	2,1
18,3	9,8
16,6	9,3
26,8	17,8
6,3	3,9
6,9	3,0
7,9	3,2

Окончание таблицы

Венозная кровь	Капиллярная кровь
6,9	4,1
14,8	11,7
28,4	17,2
11,5	5,6
39,6	26,0
19,8	9,9
27,6	14,9
19,6	12,2

Задание. Оценить статистическую значимость, а также величину различий между содержанием билирубина общего в сыворотке венозной и капиллярной крови. Можно ли для анализа на этот показатель использовать кровь не из вены, а из пальца пациента?

В пакете PAST

① Внести данные и сохранить файл «Билирубин.dat». Выделить область данных.

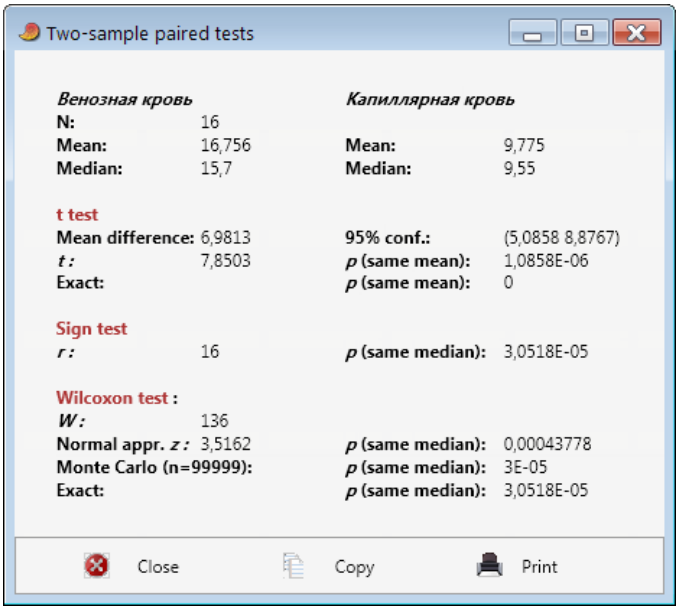
② Путь: Univariate — Two-sample paired tests (парные критерии для двух выборок).

③ В таблице результатов смотрим раздел *t*-test. Выписываем значение *t*-критерия (округляем до сотых) и соответствующее значение *p* (округляем до тысячных). До $n = 27$ программа рассчитывает точное (Exact) значение *p*, предпочтительнее взять его. В нашем случае $p = 0$, значит, пишем $p < 0{,}001$ или даже $p << 0{,}001$. Различия высоко статистически значимы. **Вопрос**: объясните, почему нельзя написать $p = 0$?

④ Рассчитываем степени свободы парного *t*-критерия как $df = n_{пар} - 1$. В нашем случае $df = 16 - 1 = 15$.

⑤ Оформляем результат: $t_{(15)} = 7{,}85$; $p < 0{,}001$. **Вопрос**: можно ли для анализа на общий билирубин использовать сыворотку капиллярной крови, а не венозной?

⑥ При сравнении зависимых выборок оценкой величины эффекта является ***средняя разность*** (*mean difference*). Она рассчитывается как среднее значение разности, вычисленное по *n* значениям разностей всех *n* пар, и совпадает со значением *разности средних*: 16,756 – 9,775 = 6,981. Пакет снабжает среднюю



разность 95% ДИ, вычисленным по формуле для нормального распределения. Эти значения округляем с точностью среднего значения. **Интерпретация**: в капиллярной крови содержание общего билирубина было в среднем меньше на 6,98 (95% ДИ от 5,09 до 8,88) мкмоль/л. Если 95% ДИ средней разности не включает ноль, значит различия статистически значимы на 5%-ном уровне — это другой способ оценки значимости различий.

ВАЖНО! Как и в случае разности средних независимых выборок, среднюю разность зависимых выборок часто приводят не в единицах шкалы признака, а в относительных единицах — в процентах от исходного или референтного значения (иногда их называют «дельта-процент»). В нашем случае стандартная методика предусматривает анализ венозной крови, а значит, референтным значением будет концентрация показателя в венозной крови: 16,756 мкмоль/л. Относительно него в капиллярной крови значение было меньше на 6,981 мкмоль/л, или на 6,981 / 16,756 × 100 % = 41,7 %. Следует пересчитать в процентах и границы ДИ для разности: 5,0858 / 16,756 × 100 % = 30,4 % и 5,0858 / 16,756 × 100 % = 53,0 %.

⑦ **Вывод**. Концентрации общего билирубина в венозной и капиллярной крови различались высоко статистически значи-

мо: парный критерий Стьюдента $t_{(15)} = 7{,}85$; $P < 0{,}001$. В капиллярной крови концентрация была ниже в среднем на 41,7 % (95% ДИ от 30,4 до 53,0 %). Таким образом, анализ венозной крови не может быть заменён на анализ капиллярной крови.

⑧ График. Можно представить средние с 95% ДИ для двух групп — как для независимых выборок (ДИ потребуется рассчитать специально). Но лучше представить среднюю разность с 95% ДИ (если важнее относительные единицы — в процентах от референтного значения). Пакет PAST пока не позволяет создать такие графики (для двух и более значений, и только если ДИ симметричен, график можно сделать в Plot — XY with error bars).

II. Количественные признаки с ненормальным распределением и порядковые признаки

Для количественных признаков с ненормальным распределением можно использовать парный *t*-критерий Стьюдента после нормализующих преобразований. Но чаще от количественных шкал (интервальная шкала и шкала отношений) переходят к порядковой шкале и рассчитывают значение ***критерия Уилкоксона для разностей пар*** (синоним: парный критерий Уилкоксона, *Wilcoxon matched pairs test*, *Wilcoxon signed rank test*). Это прямой ранговый аналог парного критерия Стьюдента, причём весьма мощный: асимптотическая эффективность критерия составляет $3/\pi$, то есть около 95 %. Рассчитаем его для этих же данных.

В пакете PAST

① Файл «Билирубин.dat» открыт, область данных выделена.

② Путь: Univariate — Two-sample paired tests (парные критерии для двух выборок).

③ В таблице результатов смотрим раздел Wilcoxon test. Выписываем значение *W* (округляем до десятых) и соответствующее *p*-значение (округляем до тысячных). До $n = 27$ программа рассчитывает точное (Exact) значение *p*, предпочтительнее взять его. Если нет точного значения — лучше использовать *p*, вычисленное методом Монте-Карло. Для больших выборок можно использовать нормальную аппроксимацию *W*-статистики: *p* из строки Normal appr. z; само *z*-значение не приводим.

④ Оформляем результат: $W_{(16)} = 136{,}0$; $p < 0{,}001$. Для W-критерия степени свободы не используются, поэтому в скобках просто указываем число пар $n = 16$.

⑤ В качестве величины эффекта можно также использовать среднюю разность, однако для ненормально распределённых данных мы не имеем права приводить эту разность с 95% ДИ из раздела парного *t*-критерия. Однако мы можем рассчитать эту разность непосредственно и построить для неё непараметрический ДИ методом бутстрепа. Сделаем это.

В пакете Excel

5.1. Скопируйте выделенные данные и вставьте в Excel, начиная с ячейки A1. В третьем столбце рассчитаем разность между содержанием ОБ. Для этого в ячейке C1 введите формулу: =A1-B1 (метки ячеек лучше вводить не с клавиатуры, а кликая на соответствующей ячейке). По нажатии на [Enter] получим результат (4,8), который нужно скопировать в оставшиеся ячейки столбца C (можно через буфер обмена, можно «протяжкой»). Полученную колонку значений разности копируем в буфер и вставляем в PAST.

5.2. В PAST рассчитываем для разности среднее значение и 95% ДИ бутстрепом (метод BCa): 6,98 (95% ДИ от 5,31 до 8,68). **Задание**: пересчитайте самостоятельно границы ДИ в процентах от среднего значения ОБ в венозной крови.

⑥ **Вывод** (вариант). Концентрации общего билирубина в венозной и капиллярной крови различались высоко статистически значимо: критерий Уилкоксона для разностей пар $W_{(16)} = 136{,}0$; $P < 0{,}001$. В капиллярной крови концентрация была ниже в среднем на 41,7 % (95% ДИ от 31,7 до 51,8 %).

III. Качественные номинальные признаки

В случае качественных номинальных признаков две зависимые выборки сравнивают обычно ***критерием Макнемара*** (*McNemar test of symmetry*). Для не слишком малых выборок статистика критерия имеет распределение хи-квадрат с одной степенью свободы. В случае малых выборок (см. далее) критерий становится слишком либеральным, поэтому вводится ***поправка Эдвардса на непрерывность*** (*Edwards' continuity correction*). Более точным и предпочтительным является использование ***точного биномиального критерия*** (*Binomial exact test*, реже он называ-

ется критерием Лидделла — *Liddell's test*). В случае его использования достаточно привести только *p*-значение.

П р и м е р. В клинических испытаниях широко используется схема с назначением плацебо. Она заключается в том, что часть пациентов получают лекарственное средство, а часть — плацебо, то есть пустышку, без явных лечебных свойств (лактоза, мел). При этом пациент не знает, что именно он получает (*простой слепой метод*), а чаще также и медицинский персонал, дающий препарат, не знает, что они дают пациенту (*двойной слепой метод*). Это позволяет исключить из результата исследования психологический компонент, связанный с верой пациента в эффективность лекарственного средства.

В небольшом эксперименте участвовало 40 пациентов, оценивавших эффективность двух препаратов, один из которых в действительности являлся плацебо. Пациентам случайным образом (рандомизация; см. лабораторную работу № 18) назначался первый или второй препарат. После паузы в лечении, достаточной для обеспечения независимости оценок препаратов, давался другой препарат. На основании высказываний пациентов врач определял действие препарата как «сильное» или «слабое». Данные находятся в файле «Плацебо.dat»; в строках — пациенты, в колонках — результат: 1 — сильное действие, 0 — слабое действие.

Задание: определить, обладает ли препарат лечебным эффектом? Если да, то какова его сила?

В пакете PAST

① Открыть файл «Плацебо.dat» и выделить область данных.

② Путь: Edit — Rearrange — Observations to contingency table (Наблюдения в таблицу сопряжённости).

Мы получили таблицу частот, для которой пока в пакете PAST нет нужного критерия, поэтому просто перепишем её в понятном виде, а далее рассчитаем критерий Макнемара вручную или в онлайновом калькуляторе.

		Действие плацебо	
		Сильное (1)	Слабое (0)
Действие препарата	Сильное (1)	8	16
	Слабое (0)	5	11

Вопрос: помогают ли нам находящиеся на диагонали таблицы (пунктирная линия) значения 8 и 11 определиться с тем, действует препарат или нет?

Восьми испытуемым помог как препарат, так и плацебо, а одиннадцати — не помогло ничего. Поэтому эти стоящие на *главной диагонали* таблицы ячейки бесполезны для сравнения: они не несут никакой информации о различиях. 16 человек оценили действие препарата как сильное, а плацебо — как слабое, а 5 человек — наоборот. Если бы мы имели числа 16 и 16, то очевидно, что действие препарата не отличалось бы от плацебо: 16 человек «проголосовали» за препарат, 16 — за плацебо. То есть мы наблюдали бы симметрию значений в ячейках над и под диагональю. Именно поэтому критерий Макнемара, а также ***критерий Боукера*** (*Bowker's test*, см. лабораторную работу № 9) для таблиц больше чем 2×2, называются ***критериями симметрии***. В нашем случае симметрия нарушена: 16 и 5; именно эти числа и будут использоваться для сравнения.

③ Расчёт по формуле. Обозначим ячейки буквами:

a	*b*
c	*d*

8	16
5	11

$$\chi^2_{McNemar} = \frac{(b - c - \text{constant})^2}{b + c},$$

где constant — константа, используемая для поправки на непрерывность. Обычно constant = 1 (поправка Эдвардса), но в некоторых пакетах constant = 0,5, и такая поправка может называться поправкой Йейтса. Если объём выборки не слишком мал, то есть $(b + c) > 25$, то поправка не нужна (constant = 0), поскольку делает критерий излишне консервативным.

В нашем случае $b + c = 16 + 5 = 21$ (<25), поэтому будем использовать поправку:

$$\chi^2_{McNemar} = \frac{(b - c - 1)^2}{b + c} = \frac{(16 - 5 - 1)^2}{16 + 5} = 4{,}76.$$

Число степеней свободы $df = 1$. Поскольку полученное значение больше критического для 5%-ного уровня значимости ($\chi^2_{(1;\,\alpha=0{,}05)} = 3{,}84$), делаем вывод о статистической значимости различий.

К сведению. Как и в случае критериев типа хи-квадрат, использование поправок оправдано лишь в случае, если для расчёта *P*-значения будет использоваться непрерывное статистическое распределение хи-квадрат. Поскольку современные программы в состоянии рассчитать точное значение *P,* минуя статистическое распределение, именно такой подход будет наиболее точным и современным. Поэтому следует искать ресурсы, которые позволяют провести Binomial exact test или Liddell's test. С одним таким ресурсом вы познакомитесь в процессе выполнения домашнего задания.

④ Оценка силы различий. В качестве показателя величины эффекта используется отношение шансов. Оно рассчитывается как отношение *наддиагонального* и *поддиагонального* элементов таблицы: $OR = b / c$.

В нашем случае $OR = 16 / 5 = 3{,}20$ (округляем до сотых), то есть шансы выраженного лечебного эффекта препарата в 3,2 раза выше, чем плацебо. Данное значение желательно снабдить 95% ДИ, которые можно рассчитать в онлайновых калькуляторах.

⑤ **Вывод** (неполный вариант). Препарат оказывал статистически значимый лечебный эффект по сравнению с плацебо: критерий Макнемара $\chi^2_{McNemar} = 4{,}76$; $P < 0{,}05$; отношение шансов $OR = 3{,}20$.

⑥ Расчёт в онлайновом калькуляторе.

Домашнее задание

В браузере

Введите в строке поисковика браузера: «McNemar test calculator». Наиболее популярные ресурсы для онлайновых расчётов будут представлены на первой странице. Опробуйте 3–4 калькулятора. Обратите внимание на тот, который позволяет рассчитать: 1) 95% ДИ для *OR*, 2) *p*-значение точным биномиальным методом: это лучше, чем использовать поправку Эдвардса. Выпишите его название и адрес в тетрадь для практических занятий, а также выдаваемые им результаты. Сформулируйте и оформите полный вывод в тетради.

ЛАБОРАТОРНАЯ РАБОТА № 8

Сравнение трёх и более выборок по количественным и порядковым показателям

Тема 8. Выборочные сравнения для случая трёх и более групп и одного действующего фактора.

Количество часов: 2.

Цель: освоить стратегию выбора статистических критериев для сравнения трёх и более групп. Научиться использовать однофакторный дисперсионный анализ, критерий Краскела — Уоллиса и соответствующие апостериорные критерии. Работа на ПК, решение задач.

Сравнение трёх и более выборок — распространённая задача в практике исследователя. При этом часто одна выборка служит контролем («контрольная группа», в медицине — «группа сравнения»), в то время как несколько других являются различными вариантами опыта («экспериментальные группы», в медицине — «основные группы»).

Методы статистического анализа в случае двух выборок и в случае трёх и более выборок различны. Наиболее частой ошибкой анализа данных в случае нескольких выборок является их попарное сравнение методами, разработанными для анализа двух выборок, например, *t*-критерием или критерием Уилкоксона — Манна — Уитни. Такое сравнение статистически некорректно, поскольку увеличивает ошибку I рода: чем больше гипотез проверяется, тем выше вероятность ложноположительных «открытий» (см. теоретический материал). Чтобы обойти эту проблему, можно использовать методы для сравнения двух групп, но применять специальные поправки на множественность сравнений типа ***поправки Бонферрони*** (*Bonferroni correction*) (см. теоретический материал). Однако такие поправки, напротив, слишком консервативны и увеличивают ошибку II рода. Поэтому при наличии нескольких выборок рационально использовать другой — двухэтапный подход к проверке гипотезы:

Этап 1. ***Омнибусный критерий*** (*omnibus test*), проверяющий весь набор («омнибус») гипотез. Если нулевая гипотеза H_0 об отсутствии различий принимается ($p > 0{,}10$), то констатируем от-

сутствие межгрупповых различий. Если H_0 отклоняется ($p \leq 0{,}05$), то далее:

Этап 2. Проводят ***запланированные сравнения*** (*planned comparisons*) или незапланированные ***множественные апостериорные сравнения*** (*post hoc comparisons*), призванные обнаружить, за счёт различий каких пар групп или их сочетаний значимым оказался омнибусный критерий.

I. Количественные признаки с приблизительно нормальным распределением

Информация о нормальности распределения берётся из литературы, предыдущих исследований или проверяется непосредственно по данным, если позволяет объём выборки ($n \geq 30$).

В качестве омнибусного критерия используется ***однофакторный дисперсионный анализ*** (*One-way **An**alysis **of Va**riance, One-way ANOVA*). Далее для ***модели I*** дисперсионного анализа (ДА), которая применяется для сравнения средних в группах, проводят запланированные или незапланированные множественные апостериорные сравнения. Критериев для таких сравнений предложено много. В статистических пакетах распространены:

- ***метод наименьшей значимой разности Фишера*** (*Fisher's LSD*) — слишком либеральный и даже некорректный метод;
- ***метод Тьюки*** (*Tukey's HSD*) — строгий и даже несколько консервативный критерий. Есть в пакете PAST;
- ***метод Бонферрони*** (не путать с поправкой Бонферрони);
- ранговые ***методы Дункана*** (*Duncan's Multiple Range Test*) и ***Ньюмена — Кёйлса*** (*Newman-Keuls test, Student-Newman-Keuls (SNK) test*) — хорошо сбалансированные и популярные методы и др.

Для ***модели II*** ДА — задача разложения общей изменчивости признака на компоненты — рассчитывают и интерпретируют компоненты дисперсии.

П р и м е р. Пойманы четыре зайца. С них собраны все личинки заячьего клеща, и у личинок измерена длина щитка (в мкм).

Данные:

заяц 1	заяц 2	заяц 3	заяц 4
380	350	354	376
376	356	360	344
360	358	362	342
368	376	352	372
372	338	366	374
366	342	372	360
374	366	362	
382	350	344	
	344	342	
	364	358	
		351	
		348	
		348	

Задание: определить, различаются ли средней длиной щитка личинки, собранные с разных хозяев? Почему?

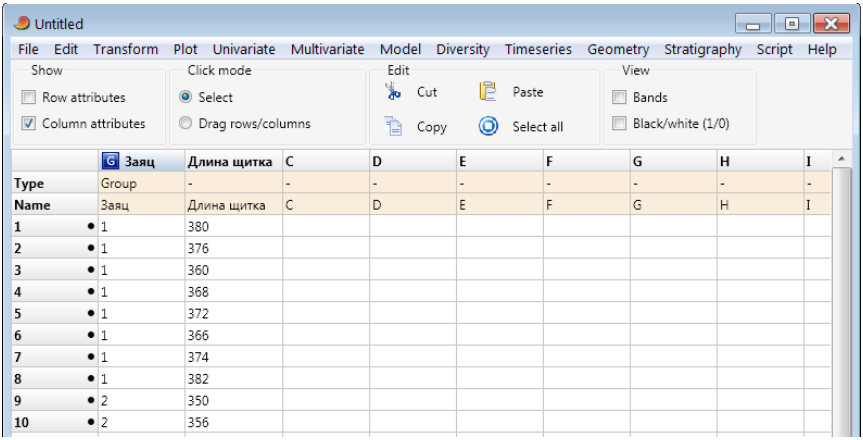
В пакете PAST

① Данные могут быть внесены в пакет двумя способами:

а) данные для разных групп вбиваются в соседние столбцы и выделяются так, чтобы выделенными оказались все значения. На предыдущих занятиях мы так и поступали, однако когда групп много, более удобен следующий способ;

б) все данные вбиваются в один столбец, и дополнительно создаётся столбец с меткой принадлежности значения к группе. Для этого нужно зайти в свойства колонки: ☐ Column attributes, ввести названия столбцов в строке Name и дважды кликнуть в ячейке Type, которая находится выше. При этом появляется выпадающее меню, в котором нужно выбрать Group. После этого рядом с названием колонки появится синий значок  , сигнализирующий о том, что данная колонка содержит метки ***группирующей переменной*** *(grouping variable)*. Далее галочку ☑ Column attributes можно снять, сохранить файл и выделить обе колонки.

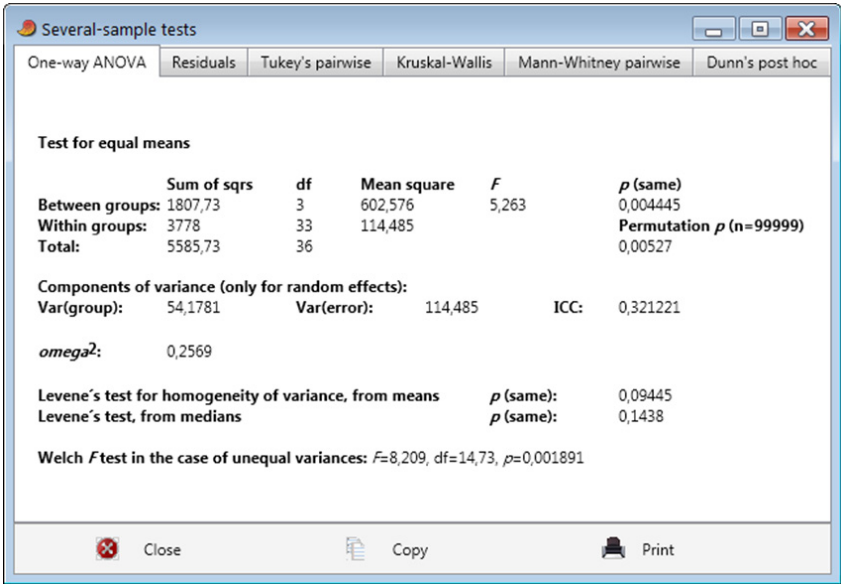
Такой способ организации данных является предпочтительным, поскольку ускоряет обработку больших массивов данных.



	Заяц	Длина щитка	C	D	E	F	G	H	I
Type	Group	-	-	-	-	-	-	-	-
Name	Заяц	Длина щитка	C	D	E	F	G	H	I
1	1	380							
2	1	376							
3	1	360							
4	1	368							
5	1	372							
6	1	366							
7	1	374							
8	1	382							
9	2	350							
10	2	356							

② Путь: Univariate — ANOVA etc. (several samples) — Several-sample tests (ANOVA, Kruskal-Wallis).

По умолчанию открывается форма на закладке $\overline{\text{One}-\text{wayANOVA}}$ (Однофакторный дисперсионный анализ):



③ Проверка требований модели дисперсионного анализа:

3.1. ***Однородность дисперсий*** (*homoscedasticity*). Аналогично

тому, как *t*-критерий Стьюдента требует равенства дисперсий, дисперсионный анализ требует их взаимного равенства, то есть однородности. В пакете она проверяется ***критерием Левена*** (Ливина) — *Levene's test …* from means. В нашем случае есть лишь тенденция к ***неоднородности дисперсий*** (*heteroscedasticity*): p = 0,094. Поэтому формально можно доверять результатам обычного дисперсионного анализа.

3.2. Нормальное распределение ошибки. Закладка $\widehat{\text{Residuals}}$ (Остатки). Это ошибка модели, то есть остатки после последовательного выражения и вычитания из каждого значения в наборе данных всех эффектов модели (общего среднего, группирующих факторов и их взаимодействий; см. теоретический материал). Ошибка должна быть нормально распределена со средним равным нулю. Нормальность распределения проверяется в пакете критерием Шапиро — Уилка, и в нашем случае нет оснований отвергать гипотезу о нормальности распределения остатков: $W_{(37)}$ = 0,98; p = 0,850. **Задание**. Посмотрите в этом же разделе распределение остатков в форме гистограммы и кривую плотности распределения на фоне кривой нормального распределения.

В случае невыполнения требований дисперсионного анализа возможны следующие варианты.

1) если распределение ошибки не отличается от нормального (для критерия Шапиро — Уилка $p > 0{,}05$), но дисперсии неоднородны (для критерия Ливена $p \leq 0{,}05$), то можно использовать результаты ***подхода Уэлча*** (Вэлча), которые пакет выдаёт чуть ниже: *Welch F test …* **Внимание!** В подходе Уэлча получаются дробные степени свободы: их уменьшение — плата за нарушение требований нормальности: чем больше отклонение, тем больше плата. В данном случае вместо df = 33 мы получаем только df = 14,73. Результат дисперсионного анализа с подходом Уэлча мы бы записали так:

$$F_{(3;\ 14{,}73)} = 8{,}21;\ P = 0{,}002;$$

2) если распределение ошибки значимо отклоняется от нормального, то следует:

а) попробовать нормализовать данные перед анализом с помощью преобразований. Одно из лучших преобразований — преобразование Бокса — Кокса. В пакете PAST путь: Transform — Box-Cox. **ВАЖНО!** Преобразование нужно применить

ко всему набору данных. Часто преобразования устраняют также и неоднородность дисперсий;

б) использовать непараметрический ранговый аналог однофакторного ДА — критерий Краскела — Уоллиса (мы рассмотрим его позже);

в) использовать рандомизационный вариант дисперсионного анализа (нет в пакете PAST).

В нашем случае требования модели дисперсионного анализа были соблюдены и можно использовать стандартную таблицу результатов дисперсионного анализа с закладки One – way ANOVA. Её нужно правильно оформить и вставить в квалификационную работу.

④ **Оформление результатов дисперсионного анализа**.

Если дисперсионных анализов в работе не очень много (до 5–7), можно все таблицы результатов привести в основной части работы. Если же таких анализов много, то в тексте приводятся только таблицы средних значений и/или графики, а результаты дисперсионного анализа описываются кратко: только $F_{(df_1;\, df_2)} = \ldots$, $P = \ldots$

Таблица 1 – Результаты дисперсионного анализа размеров заячьего клеща

Источник изменчивости	Сумма квадратов *SS*	Степени свободы *df*	Средний квадрат *MS*	*F*-критерий	Оценка значимости *P*
Между группами	1 807,73	3	602,576	5,26	0,004
Внутри групп (ошибка)	3 778,00	33	114,485	–	–
Общая	5 585,73	36	–	–	–

К сведению. На примере этой таблицы удобно рассмотреть принцип дисперсионного анализа. Вспомним формулу дисперсии: мы говорили о том, что дисперсия — это сумма квадратов отклонений от среднего значения, делённая на число степеней свободы:

$$s^2 = \frac{SS}{df} = \frac{\sum (x_i - \bar{x})^2}{n - 1}.$$

Если мы рассчитаем сумму квадратов отклонений *SS* во всём наборе данных (37 значений) без учёта групп, то получим 5 585,73. Число степеней свободы *df* будет 37 – 1 = 36. Эти значения представлены в таблице в последней строке «Общая». Если сейчас мы разделим *SS* на *df*, то получившийся средний квадрат *MS* будет самой обычной дисперсией. Идея дисперсионного анализа заключается в разложении этой дисперсии на 2 ча-

сти (для более сложных анализов таких частей будет больше): дисперсию межгрупповых различий и дисперсию внутригрупповых различий. Если сейчас в формуле дисперсии вместо x_i подставить групповые средние $\bar{x}_k$, то *SS* будет 1 807,73, а $df = k - 1$, где k — число групп, то есть 4 – 1 = 3. Деля этот *SS* на *df*, получим средний квадрат *MS*, который будет дисперсией для эффекта между группами ($MS_{\text{между}}$). Аналогично рассчитывается *SS* внутри групп; при этом суммируются квадраты отклонений исходных данных от их групповых средних, а $df = n - k$, то есть 37 – 4 = 33. Деля этот *SS* на *df*, получим средний квадрат *MS*, который будет дисперсией для эффекта внутри групп ($MS_{\text{внутри}}$).

Таким образом, в ходе анализа общая сумма квадратов разбивается на две части (5 585,73 = 1 807,73 + 3 778), а общее число степеней свободы — также на соответствующие две части (36 = 3 + 33). Средние квадраты, которые рассчитываются как $MS = SS / df$, представляют собой дисперсии, и теперь мы можем проверить, больше ли дисперсия между группами по сравнению с дисперсией внутри групп. Ранее, проверяя равенство дисперсий, мы делили одну дисперсию на другую и получали значение статистики *F*-критерия Снедекора — Фишера. Аналогично мы поступаем и теперь: делим $MS_{\text{между}}$ на $MS_{\text{внутри}}$. Если значение *F*-критерия будет равно 1, значит вся изменчивость признака объясняется исключительно внутригрупповой изменчивостью, а межгрупповая изменчивость отсутствует. Если же межгрупповая изменчивость статистически значимо больше внутригрупповой, значит для рассматриваемого явления неслучаен компонент изменчивости (дисперсии), обусловливающий межгрупповые различия, или — иначе говоря — группы различаются статистически значимо. Поскольку эффект «между группами» оценивается относительно эффекта «внутри групп», последний выступает в анализе в качестве ошибки. Чем меньше эта ошибка, тем более слабые различия между группами мы сможем обнаружить.

Таким образом, в ходе дисперсионного анализа мы работаем исключительно с дисперсиями — отсюда и название метода, — однако в результате можем делать вывод о различиях средних значений. В нашем случае различия между средними размерами личинок с разных хозяев были высоко статистически значимыми. Далее в зависимости от задачи исследования приступают либо к множественным апостериорным сравнениям средних, либо к расчёту компонентов дисперсии в процентах от общей.

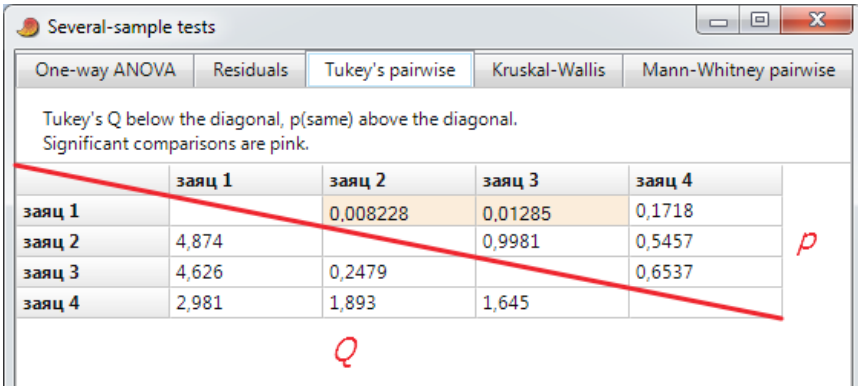
Вывод по разделу. В ходе однофакторного дисперсионного анализа были обнаружены высоко статистически значимые различия в средних размерах щитков личинок, собранных с разных хозяев: $F_{(3;\,33)} = 5{,}26$, $P = 0{,}004$.

Каков биологический смысл этого заключения? **Задание**: предположите несколько гипотез, объясняющих почему клещи различных хозяев различаются больше, чем клещи любого одного хозяина. Эти различия могут быть обусловлены разным воздействием отдельного хозяина на клеща или генетическими различиями между клещами. Клещи одного хозяина могут быть сибсами — потомками одной пары родителей — и в этом случае

различия между выборками разных хозяев представляют собой межсемейные, то есть генетические различия. Исходя из биологии рассматриваемого организма эта возможность кажется наиболее резонным объяснением.

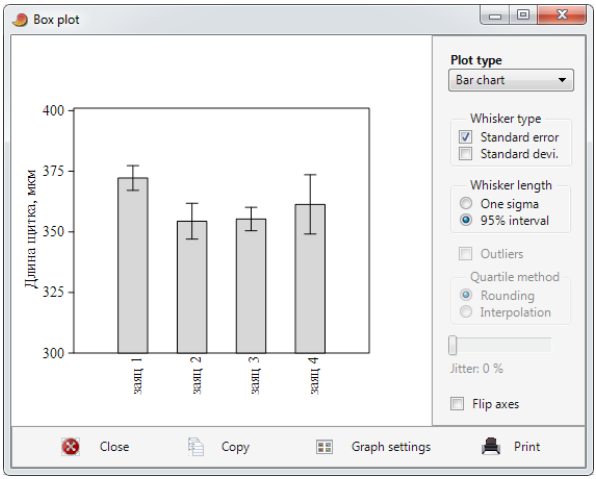
⑤ Множественные апостериорные сравнения для модели I. Поскольку омнибусный критерий обнаружил статистически значимые различия между размерами личинок клещей с разных хозяев, далее может быть полезным оценить, за счёт каких групп эти различия проявились. **Наш пример относится к модели II, поэтому здесь это неважно и показано в дидактических целях**, но в случае эксперимента с контролем и опытом — важно всегда.

Закладка Tukey's pairwise (Попарные сравнения групп методом Тьюки). В нижней треугольной матрице — сами значения критерия Тьюки Q (обычно их не приводят в работе), в верхней треугольной матрице — соответствующие значения p; цветом пакет выделяет ячейки для пар со статистически значимыми различиями ($p \leq 0{,}05$).



Вывод по разделу. Апостериорные сравнения методом Тьюки показали, что статистически значимые различия в дисперсионном анализе связаны с различиями средних размеров личинок 1-го и 2-го зайцев (p=0,008) и 1-го и 3-го зайцев (p=0,013).

⑥ График в работу для модели I.
6.1. Данные выделяются.
6.2. Путь: Plot — Barchart/Boxplot.



Изменяем длину усов Whisker length на 95% interval. В «Graph settings» подбираем значение Y start таким образом, чтобы средние значения оказались приблизительно в центре графика и/или чтобы различия по 95% ДИ были хорошо видны (не забываем подтверждать вводимые значения клавишей «Enter»). Удобно начать с какой-нибудь круглой цифры; в нашем случае это число 300. Можно было бы установить Y start в 340, а Y end в 380 — различия стали бы ещё отчётливее. Подбираем количество делений оси *y* так, чтобы цифр было немного и они были кратны 5 или 10. **Внимание!** Если при удачном подборе интервалов пропадает максимальное значение шкалы *y* — схитрите, выставив Y end равным не 400, а 401. Окончательно доработать график можно в векторных редакторах типа TpX.

По графику мы видим, что не перекрываются ДИ 1-го и 2-го зайцев и 1-го и 3-го зайцев, то есть визуальный анализ совпал с результатами сравнений методом Тьюки.

⑦ Расчёт компонентов дисперсии для модели II.

Как уже указывалось выше, рассмотренный пример относится к модели II, в которой нас интересует разложение изменчивости на составляющие части — ***компоненты дисперсии*** (*components of variance*). Принципы и формулы такой процедуры хорошо описаны в учебнике Монтгомери [7]. Отметим здесь только, что для расчётов необходимо знание математических ожидаемых

средних квадратов, которые, будучи вычисленными как дисперсии, тем не менее могут быть сложными составными выражениями, включающими интересующие дисперсии лишь в качестве членов. После вычисления таких дисперсий их сумма принимается за 100 % и для членов модели ДА рассчитывается соответствующая доля в этой сумме.

Возвращаемся на закладку $\overbrace{\text{One – way ANOVA}}$ и смотрим значения в Components of variance. Для нашего примера компонент дисперсии, привносимый межгрупповой изменчивостью Var(group) = 54,1781, а внутригрупповой — Var(error) = 114,485. Таким образом, суммарно 100 % изменчивости признака составляет 54,1781 + 114,485 = 168,6631. Доля межгрупповых различий составляет в этой сумме 54,1781 / 168,6631 = 0,321, или 32,1 %. Эта величина называется ***внутриклассовым коэффициентом корреляции*** *(intraclass correlation coefficient, ICC).* Он показывает корреляцию между объектами внутри группы относительно различий между группами. В нашем примере интерпретация компонентов дисперсии будет такой: размер щитка личинок заячьего клеща на 32,1 % обусловлен генетически, тогда как доля средовой изменчивости составляет 100 % – 32,1 % = 67,9 %.

⑧ График в работу для модели II.

Компоненты дисперсии логично представить круговой диаграммой, состоящей в случае однофакторного дисперсионного анализа только из двух секторов — долей внутригрупповой и межгрупповой дисперсии. Для этого:

8.1. Внесём в произвольном столбце таблицы PAST колонку значений 32,1 и 67,9 и выделим их.

8.2. Путь: Plot — Pie.

Поле доработки в TpX полученный рисунок может быть таким:

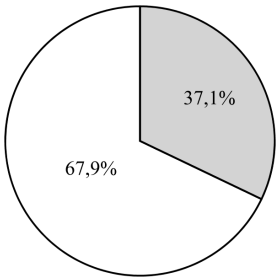


Рис. 8.1. Доля генетически обусловленной изменчивости (серая зона) размеров щитка личинок заячьего клеща в общей изменчивости

⑨ **Оформление в квалификационной работе**.

9.1. Статистическая часть раздела «Материалы и методы».

Сравнения нескольких групп по количественным показателям с приблизительно нормальным распределением проводили в ходе однофакторного дисперсионного анализа. Проверку требований метода осуществляли с помощью критериев: Ливена — для оценки однородности дисперсий и Шапиро — Уилка — для оценки нормальности распределения ошибки. Множественные апостериорные сравнения средних в рамках дисперсионного комплекса проводили методом Тьюки. Эффекты считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к различиям. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

9.2. Раздел «Результаты и обсуждение».

Если анализов в работе мало (1–3), можно привести результаты проверки требований модели ДА. Даются таблицы результатов ДА, графики и их описание.

9.3. Раздел «Выводы» (варианты).

9.3.1. Для модели I. В ходе однофакторного дисперсионного анализа были обнаружены высоко статистически значимые различия в средних размерах щитков личинок, собранных с разных хозяев: критерий Снедекора — Фишера $F_{(3;\,33)} = 5{,}26$; $P = 0{,}004$. Апостериорные сравнения методом Тьюки показали, что они были обусловлены преимущественно различиями личинок первого и второго зайца ($P = 0{,}008$) и первого и третьего зайца ($P = 0{,}013$).

9.3.2. Для модели II. В ходе однофакторного дисперсионного анализа были обнаружены высоко статистически значимые различия в средних размерах щитков личинок, собранных с разных хозяев: критерий Снедекора — Фишера $F_{(3;\,33)} = 5{,}26$; $P = 0{,}004$. Компоненты дисперсии для внутригрупповой и межгрупповой дисперсии составили соответственно 37,1 и 67,9 %. Они могут интерпретироваться как доли генетически обусловленной и прочей изменчивости размеров признака.

II. Количественные признаки с ненормальным распределением и порядковые признаки

В качестве омнибусного критерия используется ***H-критерий Краскела — Уоллиса*** (*Kruskal-Wallis test*). Это непараметрический ранговый критерий, который может рассматриваться как обобщение критерия Манна — Уитни на случай нескольких групп и как прямой ранговый аналог однофакторного ДА. Также он может быть получен как частный случай ридит-анализа, то есть иметь вероятностную интерпретацию межгрупповых различий. Критерий достаточно мощный: его асимптотическая эффективность равна 95 %, то есть на больших выборках он только на 5 % уступает в мощности дисперсионному анализу, однако менее требователен к данным. Распределение статистики критерия близко к теоретическому распределению хи-квадрат, поэтому наряду или вместо статистики H в пакетах может указываться статистика χ^2.

В пакете PAST

① Данные для разных групп вбиваются в соседние столбцы и выделяются. Или лучше использовать способ внесения данных в один столбец, а во второй поместить метку принадлежности к группе (см. выше дисперсионный анализ).

② Путь: Univariate — ANOVA etc. (several samples) — Several-sample tests (ANOVA, Kruskal-Wallis). Закладка Kruskal – Wallis.

Из результатов выписываем значение статистики H-критерия с *поправкой на связанные значения* (одинаковые значения в разных группах): H_c (*tie corrected*) и *p.* Поскольку статистика этого критерия аппроксимируется распределением хи-квадрат, рассчитаем число степеней свободы как для критерия хи-квадрат:

$$df = k - 1,$$

где k — число групп.

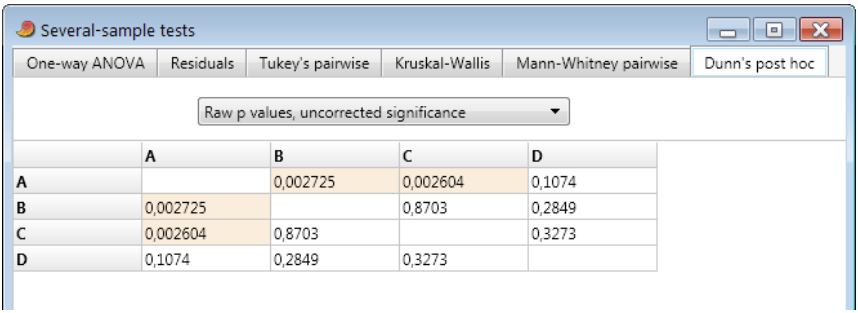
В нашем случае $df = 4 - 1 = 3$.

Вывод по разделу. В ходе сравнения групп методом Краскела — Уоллиса обнаружены статистически значимые различия в средних размерах щитков личинок, собранных с разных хозяев: $H_{(3)} = 11{,}5$; $P = 0{,}009$.

③ Множественные апостериорные сравнения. Для ранговых множественных апостериорных сравнений (*post-hoc comparisons*) используются довольно редкие в пакетах ***методы Стила — Двасса*** (*Steel-Dwass' test* — ранговый аналог метода Тьюки), ***Данна*** (*Dunn's test*), ***Неменьи*** (*Nemenyi test*) и другие, более современные. Также для попарных сравнений возможно использовать метод Манна — Уитни с поправкой Бонферрони на множественность сравнений, что менее предпочтительно даже при использовании последовательных (*sequential*) техник.

Закладка Dunn's post hoc.

Начиная с версии 3.14, в пакете PAST реализован ***метод Данна*** (*Dunn's test*). Этот метод специально разрабатывался для множественных сравнений и не нуждается в поправках, поэтому оставляем «Raw p-values, uncorrected significance». (Для попарных сравнений по Манну — Уитни следовало бы выбрать «Raw p-values, sequential Bonferroni significance»).



	A	B	C	D
A		0,002725	0,002604	0,1074
B	0,002725		0,8703	0,2849
C	0,002604	0,8703		0,3273
D	0,1074	0,2849	0,3273	

Видно, что статистически значимые различия связаны с различиями средних размеров личинок 1-го и 2-го зайцев и 1-го и 3-го зайцев.

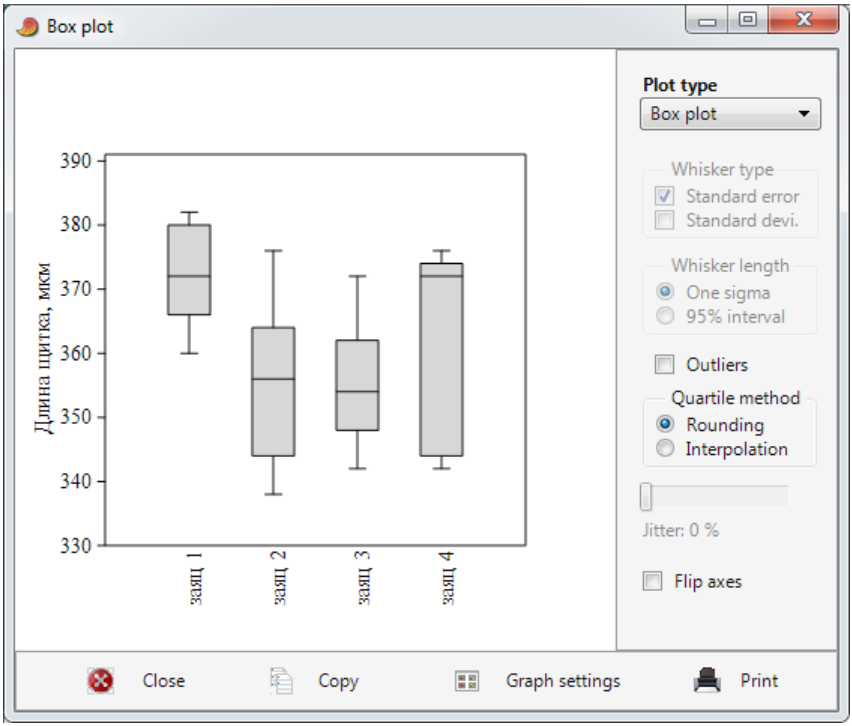
Вывод по разделу. Апостериорные сравнения методом Данна показали, что группы различаются статистически значимо. Межгрупповые различия обусловили преимущественно различия между размерами личинок 1-го и 2-го зайца (P = 0,003) и 1-го и 3-го зайца (P = 0,003).

④ График в работу.

4.1. Данные выделяются.

4.2. Путь: Plot — Barchart/Boxplot.

4.3. Изменяем тип графика Plot type на Box plot и получаем *коробчатую диаграмму*, которую можно доработать в «Graph settings» (шрифт, интервалы на осях и т. д.) и вставлять в работу.



Коробчатый график хорошо показывает особенности распределения показателя в группах, но не позволяет визуально оценить статистическую значимость различий. Поэтому в публикациях такой график часто снабжают дорисованными в графических редакторах скобками с указанием значения *p*. Для этого удобно использовать простые векторные редакторы (например, бесплатный редактор TpX), хотя можно и растровый типа Paint.

Для доработки графика в TpX нужно сохранить максимально приближенный к идеалу график в формате *.svg (Export — Save as…), открыть в TpX и пририсовать скобки со значениями *p*. Иногда вместо значений *p* рисуют звёздочки: * для $p \leq 0{,}05$, ** для $p \leq 0{,}01$, *** для $p \leq 0{,}001$, но это хуже и архаичнее точных значений.

⑤ **Оформление в квалификационной работе**.

5.1. Статистическая часть раздела «Материалы и методы».

Сравнения нескольких групп по количественным показателям с ненормальным распределением проводили с помощью критерия Краскела — Уоллиса. Для множественных апостериорных сравнений использовали критерий Данна. Различия считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к различиям. Расчёты и графические построения выполнены в пакетах PAST (v. 3.19; Hammer et al., 2001) и TpX (Дать ссылку на источник).

5.2. Раздел «Результаты и обсуждение».

В работу даются таблицы описательной статистики и/или графики, а также делается их описание с выделением наиболее существенных моментов. **Внимание!** Раньше мы получали описательную статистику по группам, располагая их в соседних столбцах. На этом занятии мы научились использовать для обозначения групп отдельный столбец с меткой. Описательную статистику по группам мы можем получить и при таком их определении. **Задание**: получите описательную статистику для данных по размерам личинок клещей новым способом.

> **К сведению**. Мы уже говорили, что исходные данные удобно хранить в листах электронных таблиц типа Excel. По умолчанию Excel создаёт в файле три листа. Рационально «Лист 1» переименовать в «Данные», «Лист 2» — в «Коды», а «Лист 3» — в «Описательная статистика». В пакете PAST под таблицей результатов следует нажать на [Copy] и скопировать данные в буфер, а в Excel вставить их из буфера в лист «Описательная статистика» и сохранить. Таким образом, в одном файле будут сохранены и сами данные, и описательная статистика к ним, которая может далее пригодиться при переоформлении работы или написании статьи. Также можно создать дополнительные листы для результатов сравнений, поиска связей и т. д.

Доработанный в редакторе TpX график выглядит так, как показано на рис. 8.2

5.3. Раздел «Выводы».

Обнаружены высоко статистически значимые различия в средних размерах щитков личинок, собранных с разных хозяев: критерий Краскела — Уоллиса $H_{(3)} = 11{,}5$, $P = 0{,}009$. Множественные апостериорные сравнения методом Данна показали, что они были обусловлены преимущественно различиями личинок

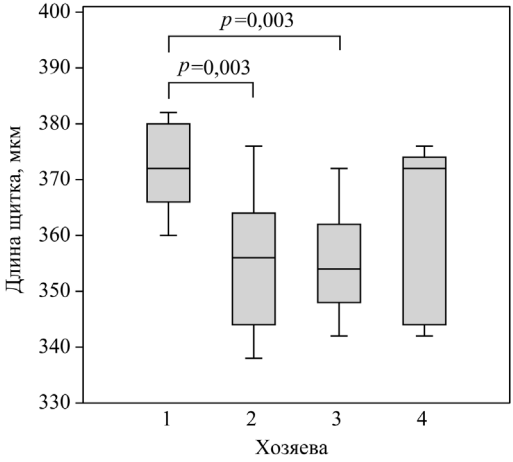


Рис. 8.2. Различия размеров щитка личинок клещей, собранных с разных хозяев

первого и второго зайцев ($P = 0{,}003$) и первого и третьего зайцев ($P = 0{,}003$).

ЛАБОРАТОРНАЯ РАБОТА № 9

Сравнение трёх и более выборок по качественным показателям

Тема 8. Выборочные сравнения для случая трёх и более групп и одного действующего фактора.

Количество часов: 2.

Цель: Овладеть методами анализа различий между выборками по качественным показателям в ходе анализа таблиц сопряжённости. Научиться находить и интерпретировать стандартизованные остатки. Работа на ПК, решение задач.

При анализе качественных признаков часто таблицы сопряжённости (ТС) получаются больше, чем те таблицы 2×2, которые мы научились анализировать на лабораторном занятии № 6. Либо для двух групп число категорий оказывается больше двух (например, цвет венчика цветка: белый, сиреневый, фиолетовый), либо при наличии двух признаков число групп больше двух (например, наличие седины у людей пяти возрастных категорий), либо и признаков, и групп больше двух. Во всех трёх случаях говорят о ***таблицах сопряжённости r×c*** (от английского *r* — *rows* — ряды, строки и *c* — *columns* — колонки, столбцы).

Рассмотрим варианты анализа таких таблиц для случая независимых выборок и зависимых.

I. Независимые выборки

В случае независимых выборок в ячейках таблицы представлены данные, относящиеся к разным объектам исследования (образцы, животные, люди и т. д.).

Если один из входов таблицы можно упорядочить (например: мало, средне, много), то две группы можно — и даже правильнее — сравнить критерием Манна — Уитни, а несколько — критерием Краскела — Уоллиса. Для этого категории «мало» присваивается ранг 1, «средне» — 2, «много» — 3. Пример см. в лабораторной работе № 6 (с. 98). Именно таким образом сравнивает упорядоченные категории пакет StatXact от компании Cytel.

Если категории упорядочить нельзя, то есть если данные представлены номинальной шкалой, анализ проводят в два этапа:

Этап 1. Омнибусный критерий, который проверяет согласие наблюдаемых и ожидаемых частот для всех ячеек таблицы. Здесь используются те же критерии согласия или современные рандомизационные критерии, которые мы рассмотрели для таблиц 2×2 (см. лабораторную работу № 6). Если нулевая гипотеза H_0 об отсутствии различий с ожидаемыми частотами принимается ($p > 0{,}10$), то констатируем отсутствие межгрупповых различий. Если H_0 отклоняется ($p \leq 0{,}05$), то далее:

Этап 2. Вместо апостериорных сравнений для таблиц сопряжённости проводят выявление ячеек, давших наибольший и неслучайный вклад в отклонение нулевой гипотезы. Это делается с помощью расчёта ***отклонений Фримана — Тьюки*** (*Freeman-Tukey deviation*, FT_{dev}) или ***согласованных стандартизованных остатков*** (*Adjusted residuals, AR*), называемых также ***остатками Хабермана***.

Если требуется, то на заключительном этапе анализа рассчитываются показатели величины эффекта — относительные риски или отношения шансов. При этом может потребоваться свёртка большой ТС в таблицу 2×2 путём объединения менее важных категорий.

П р и м е р. Среди 282 членов актёрской ассоциации был проведён социологический опрос. При этом отмечался пол и цвет волос респондента. Получены следующие абсолютные частоты (количество человек):

	Цвет волос			
	Чёрный	Коричневый	Светлый	Рыжий
Мужчины	32	43	16	3
Женщины	55	65	64	4

Задание: оценить различия между мужчинами и женщинами по соотношению обладателей волос разного цвета. Если различия есть, то установить, в чём они заключаются и каковы их возможные причины (биологические, социальные, иные)?

> **Комментарий**. Вопрос можно переформулировать и для задачи сравнения нескольких групп: различаются ли обладатели волос разного цвета соотношением полов?

В пакете PAST

① Дать названия строчкам и колонкам: как в таблице с данными. Ввести 8 значений данных в соседние ячейки и выделить.

② Путь: Univariate — Contingency table.

③ Выписываем значение критерия хи-квадрат (Chi^2), степени свободы (degrees of freedom), *p*. Если в таблице есть значения 5 и менее (наш случай) — выписываем *p*, вычисленное рандомизационной процедурой Монте-Карло. При этом число перестановок Permutation N можно увеличить до 99 999 или даже 999 999 и нажать [Recompute]. Указанием на достаточность числа перестановок является неизменное число в третьем знаке после запятой для *p* при нескольких последовательных нажатиях [Recompute].

Вывод промежуточный: мужчины и женщины статистически значимо различались соотношением обладателей волос разного цвета: критерий хи-квадрат Пирсона $\chi^2_{(3)} = 9{,}19$; $p = 0{,}026$.

Таким образом, различия мы обнаружили, но пока непонятно, в чём именно они заключались. Для того чтобы разобраться в ситуации, нужно рассчитать относительные частоты (в процентах), а также выявить ячейки, давшие неслучайный вклад в статистику критерия. Но начать полезно с графика.

④ График. Путь: Plot — mosaic plot. Программа сообщает об ошибке: слишком много колонок. Поэтому развернём таблицу иначе: транспонируем матрицу данных. Путь: Edit — Rearrange — Transpose. В полученной таблице выделяем данные и опять: Plot — mosaic plot. Можно раздвинуть блоки сильнее (Spacing = 4–5) и добавить на график проценты каждой категории от общего числа наблюдений: ◉ Percentages.

В ***мозаичном графике*** площадь плитки пропорциональна частоте (рис. 9.1). Из него видно, что в выборке было почти в 2 раза больше женщин, чем мужчин. Наиболее сильные различия между полами наблюдались по светлому цвету волос: женщин-блондинок было заметно больше. График можно доработать в редакторе TpX (хотя он более полезен в качестве средства эксплораторного анализа).

⑤ Расчёт относительных частот. Поскольку мы сравниваем мужчин и женщин, относительные частоты нужно рассчитывать для каждого пола отдельно, а не как на мозаичном графике —

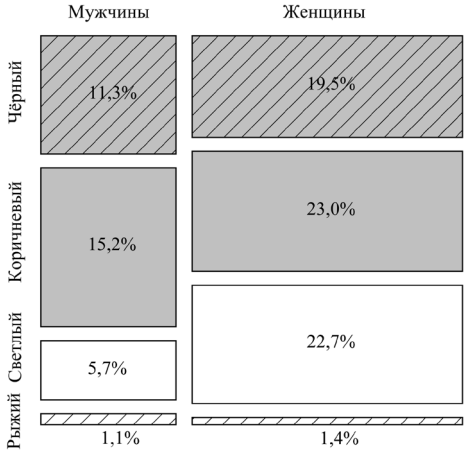


Рис. 9.1. Соотношение мужчин и женщин с разным цветом волос в актёрской ассоциации

от общего числа. Например, для ячейки 11 ($r = 1$, $c = 1$ — мужчины с чёрными волосами) имеем: 32 / (32 + 43 + 16 + 3) = 32 / 94 = 0,340, или 34,0 %. Итоговая таблица в процентах:

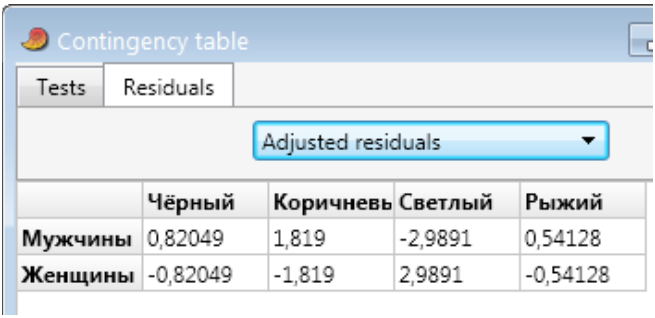
	Цвет волос				
	Чёрный	Коричневый	Светлый	Рыжий	Всего
Мужчины	34,0	45,7	17,0	3,2	100
Женщины	29,3	34,6	34,0	2,1	100

Видно, что наиболее сильные различия наблюдаются по доле обладателей светлых волос: среди мужчин таких было 17,0 %, в то время как среди женщин — 34,0 %.

⑥ Расчёт согласованных стандартизованных остатков. Закладка Residuals, в окне выбираем согласованные остатки — Adjusted residuals (рис. на с. 137).

Знак остатков указывает на направление отклонения. Например, мужчин с чёрным цветом волос было несколько больше ($AR = +0{,}82$), а женщин — несколько меньше ($AR = -0{,}82$), чем ожидалось в соответствии с нулевой гипотезой. К сожалению, пока пакет PAST не рассчитывает значимость остатков.

⑦ Статистическая значимость остатков. Стандартизованные



	Чёрный	Коричневь	Светлый	Рыжий
Мужчины	0,82049	1,819	-2,9891	0,54128
Женщины	-0,82049	-1,819	2,9891	-0,54128

остатки Хабермана распределены нормально, а значит, значения, равные или большие 1,96, статистически значимы на 5%-ном уровне значимости ($p \leq 0{,}05$). В нашей таблице таких ячеек две, обе — для светлого цвета волос. Для значений между 1,64 и 1,96 p-значение также не слишком мало ($p \leq 0{,}10$), что можно рассматривать как тенденцию к различиям — на такие ячейки полезно обращать внимание: возможно, с увеличением объёмов выборки они также окажутся значимыми. Таких значений в таблице тоже 2 — для коричневого цвета волос.

Более точно p-значения можно рассчитать в Excel:

7.1. Введите в ячейку A1 нужное значение остатка: –2,9891.

7.2. В ячейку A2 нужно поместить формулу

=2*(1-НОРМСТРАСП(ABS(A1)))

7.3. Выписать результат: 0,02798, или округлённо $p = 0{,}003$.

7.4. Теперь, изменяя значения в ячейке A1, можем получить p-значения и для других интересующих ячеек. Например, для мужчин-брюнетов имеем $AR = 0{,}820$; $p = 0{,}412$ (различия незначимы), для мужчин-шатенов $AR = 1{,}819$; $p = 0{,}069$ (тенденция к различиям) и т. д.

⑧ Интерпретация: статистическая значимость гендерных различий по соотношению обладателей волос разного цвета (см. п. 3) была обусловлена преимущественно различиями между блондинами и шатенами: среди женщин было существенно больше блондинок и несколько меньше шатенок, у мужчин ситуация была обратной. Если это представляется важным, можно вычислить относительный риск или отношения шансов для интересующих эффектов.

Вопрос: каковы возможные причины обнаруженных различий?

Мы не являемся специалистами по генетике человека и не знаем механизмов, по которым наследуется и проявляется у потомков окраска волос. Весь наш опыт базируется на сугубо личных наблюдениях в кругу семьи, в семьях родственников, друзей и знакомых. Однако это не означает, что мы не должны пытаться объяснить обнаруженное явление. Возможно, для трактовки каких-то явлений нам будет достаточно имеющихся знаний и здравого смысла. Давайте рассуждать: могут ли различия между полами иметь биологическую природу? Если это так, то каковы возможные механизмы? Почему блондинок больше, чем блондинов?

Здесь возможны два варианта: 1) сцепленное с полом наследование окрасок волос: если предположить, что за окраску волос отвечает несколько генов и какие-то гены этой системы расположены на половых хромосомах, то наблюдаемая картина возможна; 2) селективные преимущества определённых генотипов окрасок или прочно сцепленных с ними генов, проявляющиеся различной выживаемостью организмов разного пола на ранних этапах онтогенеза (например, мужские половые гормоны так взаимодействуют с продуктами генов окрасок или сцепленными с ними генами, что вызывают повышенную гибель мальчиков-блондинов до рождения). Окраска волос человека является одной из главных фенотипических черт, и очень маловероятно, что мы ничего бы не читали и не слышали о таких биологических механизмах: скорее всего, мы бы знали об этом ещё со школы. Поэтому вернёмся к данным примера и посмотрим ещё раз, что у нас была за выборка и может ли она отражать различия между мужчинами и женщинами без явного смещения оценок.

Поскольку объектами исследования были члены актёрской ассоциации, полученные данные корректно распространять в первую очередь на актёров. Мы знаем, что успешная актёрская карьера обусловлена удачными ролями в кинофильмах и/или спектаклях, а на эти роли претендентов назначает режиссёр. Таким образом, соотношение актёров с разным цветом волос может отражать выбор режиссёрами определённых типажей для персонажей фильмов или постановок. В таком случае преобладание в выборке блондинок и шатенов связано, скорее всего, с текущим запросом режиссёров и зрителей на определённые типажи. Следовательно, наиболее правдоподобным объяснением обнаруженных различий являются социальные, а не биологические причины.

⑨ **Оформление в квалификационной работе (вариант)**.

9.1. Статистическая часть раздела «Материалы и методы».

Сравнение независимых выборок по качественным номинальным показателям проводили в ходе анализа таблиц сопряжённости с помощью критерия хи-квадрат Пирсона. Для слабонасыщенных таблиц (имелись ячейки со значениями $f_{ij} \leq 5$), оценку статистической значимости проводили рандомизационной техникой Монте-

Карло (n = 99 999). Для выявления ячеек таблицы, давших неслучайный вклад в статистику критерия, рассчитывали согласованные стандартизованные остатки Хабермана. Различия считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к различиям. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

9.2. Раздел «Результаты и обсуждение».

Даются таблицы с абсолютными (в штуках, единицах) и относительными (в процентах) частотами. Последние желательно снабдить 95 % ДИ, вычисленными по Джеффрису (Уилсону, Агрести — Коулу или Клопперу — Пирсону, см. лабораторную работу № 2). Также приводятся результаты статистического сравнения. Можно сделать столбчатые диаграммы с 95% ДИ.

9.3. Раздел «Выводы».

Между мужчинами и женщинами — членами актёрской ассоциации — обнаружены статистически значимые различия по соотношению обладателей волос разного цвета: критерий хи-квадрат Пирсона $\chi^2_{(3)} = 9{,}19$; $P = 0{,}026$. Эти различия заключались преимущественно в преобладании блондинок над блондинами (стандартизованный остаток $AR = 2{,}99$; $P = 0{,}003$) и шатенов над шатенками ($AR = 1{,}819$; $P = 0{,}069$) и объяснялись, вероятно, социальными факторами, действующими в профессиональной актёрской среде.

II. Зависимые выборки

При зависимых выборках в ячейках таблицы два или более раз фигурируют одни и те же объекты исследования (образцы, животные, люди и т. д.).

Если категориальные данные можно упорядочить (например: мало, средне, много), то две группы можно сравнить парным критерием Уилкоксона (лабораторная работа № 7), а несколько — критерием Фридмана (лабораторная работа № 10). Для этого категории «мало» присваивается ранг 1, «средне» — 2, «много» — 3, аналогично тому, как было показано в лабораторной работе № 6 (с. 98). Если категории упорядочить нельзя, то есть если данные представлены номинальной шкалой, анализ обычно проводят с использованием ***критерия симметрии Боукера*** (*Bowker's symmetry test*), который является обобщением критерия

Макнемара на случай нескольких зависимых выборок, и может называться в статпакетах критерием Макнемара — Боукера или некорректно — просто критерием Макнемара. Несколько реже применяют ***критерии краевой однородности*** (*marginal homogeneity tests*) ***Стюарта — Максвелла*** (*Stuart-Maxwell test*) или ***Бхапкара*** (*Bhapkar's test*). Статистика всех трёх критериев аппроксимируется распределением хи-квадрат, то есть их числовые значения близки и на практике все они обычно приводят к одинаковым выводам. Более предпочтительной альтернативой этим критериям является ***точный биномиальный критерий*** (*Binomial exact test*).

Функциональное ограничение всех четырёх методов заключается в том, что по нескольким категориям сравниваются только две зависимые выборки; более сложные ситуации моделируются с использованием ***обобщённых линейных моделей*** (*Generalized Linear Models, GLM*).

 **П р и м е р**. Катаракта — заболевание, при котором нарушается прозрачность хрусталика, что приводит к снижению зрения (вплоть до слепоты), а также к повышению риска травматизма и депрессии. Разные типы катаракты имеют разную этиологию, а в таком случае должно наблюдаться соответствие между типами катаракты, развивающейся в левом и правом глазу больного. В ходе небольшого (n = 95) исследования у пациентов глазной клиники, имеющих катаракту обоих глаз, регистрировался её тип в левом и правом глазу. Получены следующее данные:

Левый глаз	Правый глаз			Всего
	Ядерная	Кортикальная	Субкапсулярная	
Ядерная	18	11	6	35
Кортикальная	3	15	7	25
Субкапсулярная	10	9	16	35
Всего	31	35	29	95

Задание: определить, отличаются ли левый и правый глаза по частотам развития катаракты трёх типов.

> **Комментарий**. Поскольку оба глаза образуют пару, принадлежащую одному индивиду, выборки являются зависимыми. Обратите внимание на следующие моменты в таблице. Если бы между катарактами правого и левого глаз

было идеальное соответствие, то все значения были бы сосредоточены в трёх ячейках на главной диагонали таблицы, то есть на пересечении строк и столбцов: ядерная — ядерная, кортикальная — кортикальная и субкапсулярная — субкапсулярная. Если же соответствие неполное, но отклонения от него для правого и левого глаз одинаковы, то, во-первых, значения над диагональю (11, 6, 7) должны являться зеркальным отражением значений под диагональю (3, 10, 9), а во-вторых, краевые частоты таблицы в столбце «Всего» (35, 25, 35) и в строке «Всего» (34, 35, 29) должны быть одинаковыми. Критерий Боукера оценивает нарушение симметрии наддиагональной и поддиагональной частей таблицы, а критерии краевой однородности Стюарта — Максвелла и Бхапкара оценивают различия в краевых частотах.

① **Расчёт критерия Боукера.** В пакете PAST необходимые критерии отсутствуют, однако ручной расчёт очень прост и не требует вычисления ожидаемых частот. Алгоритм действий следующий:

1.1. Находим диагональ таблицы, значения в ячейках которой указывают на сходство зависимых выборок. Они не помогают нам выявить различия между выборками, а потому не участвуют в расчётах: зачеркнём диагональю значения 18, 15, 16.

18	11	6
3	15	7
10	9	16

1.2. Находим пары значений, симметричные относительно диагонали, и подставляем их в формулу критерия Боукера:

$$\chi^2 = \sum \frac{\left(f_{ij} - f_{ji}\right)^2}{f_{ij} + f_{ji}}.$$

$$\chi^2 = \frac{(11-3)^2}{11+3} + \frac{(6-10)^2}{6+10} + \frac{(7-9)^2}{7+9} = \underline{4{,}5714} + 1{,}0000 + 0{,}2500 = 5{,}8214.$$

В ходе расчёта критерия удобно попутно обращать внимание и на значения членов критерия. Подчеркнём слагаемое, давшее максимальный вклад в статистику критерия: 4,5714. Значение статистики округлим до сотых: 5,82.

1.3. Рассчитываем степени свободы как число слагаемых в критерии Боукера или по формуле $df = i(i - 1) / 2$, где i — число категорий:

$$df = 3 \times (3 - 1) / 2 = 3 \times 1 = 3.$$

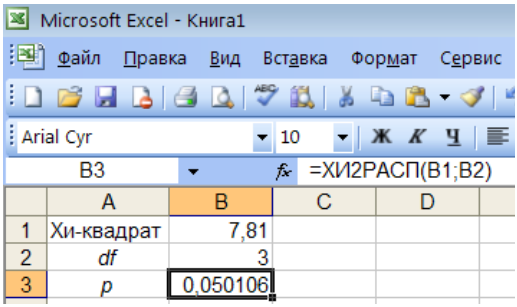
② Оценка статистической значимости.

Полученное значение χ^2 при нужном числе степеней свободы сравнивается с табличным [1. С. 134]. Если оно превосходит табличное, значит различия статистически значимы. В нашем случае $5{,}82 < 7{,}81$ и $p > 0{,}05$, следовательно, различия незначимы:

Уровень значимости α (двусторонний)	χ^2 критическое
	5,82
0,05	7,81
0,01	11,35
0,001	16,27

В пакете Excel

Более точно оценку p можно рассчитать в электронной таблице Excel. Для этого создадим небольшой расчётный блок. В столбец A поместим названия, в столбец B — значения. Ячейку B3 сделаем расчётной и поместим в неё статистическую формулу для расчёта вероятности p по значениям величины распределения и степени свободы:



Изменяя значения статистики хи-квадрат на табличные (7,81, 11,35 и 16,27), убеждаемся в том, что наш блок считает правильно. Затем подставляем значение 5,82 и получаем $p = 0{,}121$. Окончательно имеем

$$\chi^2_{(3)} = 5{,}82;\ p = 0{,}121.$$

③ Интерпретация.

Мы видели, что наиболее сильные различия наблюдались для пары ядерной и кортикальной катаракты: 4,5714 / 5,8214 = 0,785, или 78,5 % всех различий между правым и левым глазом. Если бы различия были статистически значимы, то мы бы считали, что при ядерной катаракте в левом глазу в правом чаще развивается кортикальная катаракта: отношение шансов $OR = 11 / 3 = 3{,}67$. Однако поскольку в нашем случае различия не были статистически значимыми, констатируем отсутствие каких бы то ни было различий между правым и левым глазом в развитии катаракты трёх типов.

④ **Оформление в квалификационной работе (вариант)**.

4.1. Статистическая часть раздела «Материалы и методы».

Сравнение двух зависимых выборок групп по нескольким качественным номинальным показателям проводили в ходе анализа таблиц частот с помощью критерия симметрии Боукера. В качестве показателя величины эффекта рассчитывали отношения шансов *OR*. Различия считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к различиям.

4.2. Раздел «Результаты и обсуждение».

Даются таблицы с абсолютными (в штуках, единицах) и относительными (в процентах) частотами. Последние желательно снабдить 95% ДИ, вычисленными по Джеффрису, Вилсону, Агрести — Коулу или Клопперу — Пирсону (см. лабораторную работу № 2). Поскольку в данном случае в качестве объёма выборки используется число пар значений, это число и следует использовать при расчёте ДИ; в нашем примере $n = 95$. Можно сделать столбчатые диаграммы с 95% ДИ, но нужно подумать, как их разместить и сгруппировать. Или можно дать мозаичный график, который в случае критериев симметрии будет весьма информативен.

4.3. Раздел «Выводы».

У пациентов глазной клиники с катарактой обоих глаз не обнаружено различий в частотах развития ядерной, кортикальной и субкапсулярной катаракты в правом и левом глазу: критерий симметрии Боукера: $\chi^2_{(3)} = 5{,}82$; $P = 0{,}121$.

ЛАБОРАТОРНАЯ РАБОТА № 10

Сложные модели дисперсионного анализа

Тема 9. Выборочные сравнения для трёх типов данных в случае нескольких действующих факторов.

Количество часов: 2.

Цель: Научиться различать модели с фиксированными факторами, случайными факторами и смешанные модели. Овладеть методами двухфакторного дисперсионного анализа (в том числе с единственным наблюдением на ячейку), анализа повторных измерений и критерием Фридмана. Работа на ПК.

I. Двухфакторный дисперсионный анализ

В ходе лабораторной работы № 8 мы познакомились с однофакторным дисперсионным анализом (ДА) — методом, позволяющим исследовать влияние одного контролируемого (модель I) или случайного (модель II) фактора. Однако в исследовательской практике типичны ситуации, когда требуется контролировать или учитывать влияние сразу нескольких факторов. Например, при сравнительной оценке влияния на урожайность нескольких видов удобрений (фактор 1) необходимо учесть также тип почвы (фактор 2). Или при сравнении разных способов лечения заболевания (фактор 1) необходимо учесть пол пациента (фактор 2) и его возраст (фактор 3). Такие задачи решаются в ходе ***двухфакторного*** (*Two-way ANOVA*) или ***многофакторного*** (*Factorial ANOVA*) дисперсионного анализа.

В простейшем случае все факторы являются фиксированными, а дисперсионный комплекс является *равномерным*, то есть содержит одинаковое число наблюдений в каждой ячейке. В более сложном случае какие-то из факторов могут быть случайными, а в многофакторных комплексах это имеет принципиальное значение для правильного соотнесения средних квадратов эффектов и оценки их статистической значимости. Кроме того, в сложных случаях возможно сочетание зависимых и независимых выборок. Например, при оценке эффективности лечения в двух группах пациентов (фактор 1), каждый из пациентов обследовался на нескольких сроках лечения (фактор 2). В этом случае группы пациентов являются независимыми, а одни и те же па-

циенты внутри своей группы на разных сроках — зависимыми. Для того чтобы разобраться в этом вопросе, рекомендуем прочитать гл. 7 в учебнике Д. К. Монтгомери «Планирование эксперимента и анализ данных» [7].

Все методы ДА являются параметрическими и требуют нормального распределения признака в популяции для каждого сочетания градаций. На практике это требование проверяется анализом остатков модели дисперсионного анализа: они должны быть распределены нормально со средним, равным нулю. Вторым требованием ДА является однородность дисперсий в ячейках комплекса. Если эти два требования не выполняются, на практике обычно используют преобразования исходных данных. Такие преобразования часто позволяют решить одновременно обе проблемы. В качестве альтернативы можно использовать непараметрические ранговые аналоги дисперсионного анализа, однако чем сложнее вариант анализа, тем меньше шансов найти полностью подходящую непараметрическую процедуру.

Если в сложном ДА анализе обнаруживаются статистически значимые эффекты, то далее проводят множественные запланированные или незапланированные сравнения (для фиксированных факторов, см. лабораторную работу № 8) или проводят разложение дисперсии на компоненты (для случайных факторов).

Принципиально новым эффектом в двухфакторных и более сложных ДА является ***взаимодействие факторов*** (*interaction of factors*) (рис. 10.1).

Двухфакторный и более сложные схемы имеют самое важное приложение в экспериментальной практике. Так, если мы изучаем влияние на признак двух факторов А и В, то наиболее информативной будет такая схема исследования, когда объекты будут разделены по блокам, образованным всеми возможными комбинациями уровней факторов. Так, при двух градациях фактора А (например, контроль — опыт или мужчины — женщины) и четырёх градациях фактора В (например, типы почв или возрастные группы) экспериментальные единицы следует разделить по ячейкам такой таблицы:

		Фактор В			
		1	2	3	4
Фактор А	1	A_1B_1	A_1B_2	A_1B_3	A_1B_4
	2	A_2B_1	A_2B_2	A_2B_3	A_2B_4

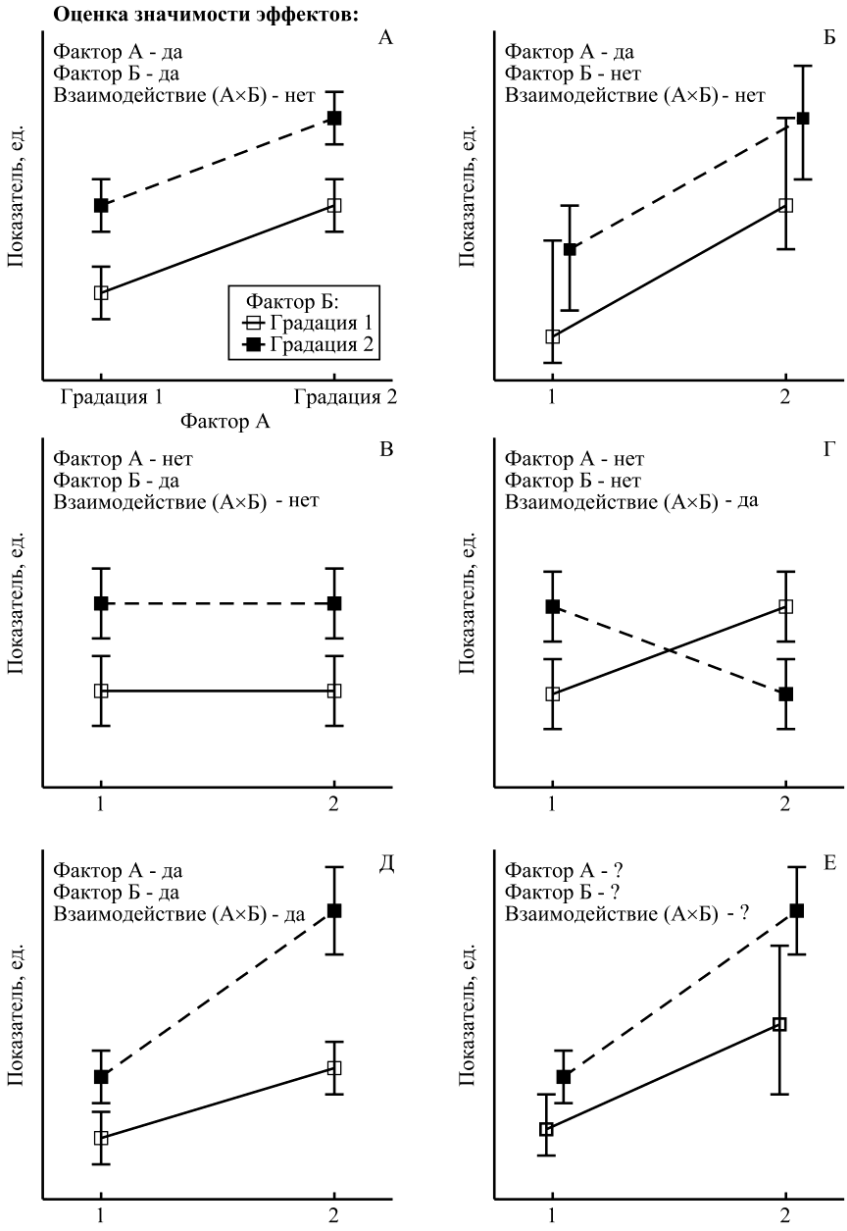


Рис. 10.1. Варианты взаимодействия факторов на графике профилей по результатам двухфакторного дисперсионного анализа

Такой эксперимент будет полноблочным, а если в каждом блоке будет одинаковое число объектов — ещё и равномерным. Именно такая схема является оптимальной для анализа данных, однако может быть невыгодна по экономическим соображениям. Поэтому в экспериментах вместо полных рандомизированных блоков могут использоваться менее ресурсоёмкие схемы, например, латинские и греко-латинские квадраты (см. теоретический материал).

Если каждый блок эксперимента будет содержать только одно значение, то в качестве ошибки для оценки действия факторов будет использоваться взаимодействие факторов, значимость которого оценить будет нельзя. Однако, если в каждом блоке будет хотя бы по два наблюдения, у нас появятся данные для вычленения изменчивости (дисперсии) внутри блоков, а следовательно, появится возможность статистически оценить не только главные эффекты A и B, но также их взаимодействие AB. Если такое взаимодействие будет значимым, значит эффекты факторов A и B неаддитивны: в каких-то их сочетаниях отклик изучаемой системы оказывается сильнее и/или слабее, чем можно было бы предполагать на основе аддитивной модели. Предположим, например, что тип почвы B_3 в среднем позволяет получить в 1,5 раза больший урожай по сравнению с остальными типами, а применение удобрения A_2 — ещё в среднем на 20 % (то есть в 1,2 раза) увеличивает урожайность по сравнению с контролем A_1. Таким образом, можно было бы ожидать, что сочетание A_2B_3 повысит урожайность в $1{,}5 \times 1{,}2 = 1{,}8$ раза. Однако при взаимодействии факторов может оказаться, что именно данное сочетание увеличивает урожайность в 3 раза, либо, напротив, удобрение оказывается неэффективным для данного типа почвы и его эффект не проявляется, «работает» только тип почвы и прирост будет 1,5 раза.

Пример. Для изучения влияния температуры на развитие дрозофил был поставлен небольшой эксперимент. В пять пробирок с питательной средой были посажены пять пар дрозофил. После откладки самками яиц из каждой пробирки были отобраны 120 яиц и помещены в шесть пробирок по 20 яиц в каждой. Две пробирки от каждой самки инкубировались при температуре 20 °C, две — при 25 °C, две — при 30 °C. По окончании эксперимента

подсчитывали количество живых мух и рассчитывали количество неразвившихся яиц и погибших личинок. Данные по погибшим особям представлены в таблице.

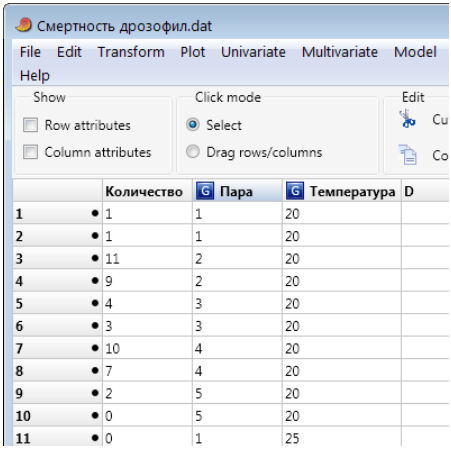
Пара (фактор А)	Температура (фактор Б)					
	20 °C		25 °C		30 °C	
	Пробирка 1	Пробирка 2	Пробирка 1	Пробирка 2	Пробирка 1	Пробирка 2
1	1	1	0	4	0	1
2	11	9	5	5	10	14
3	4	3	3	2	1	1
4	10	7	8	6	5	7
5	2	0	2	0	2	4

Задание: определить, как влияют температура и генетические различия на выживаемость потомства мух.

> **Комментарий**. Как видно из таблицы, данный эксперимент был спланирован так, чтобы в ячейке дисперсионного комплекса (выделена рамкой в таблице) оказалось более одного наблюдения. Такое дублирование обеспечило повторности для выражения неконтролируемой в эксперименте изменчивости.

В пакете PAST

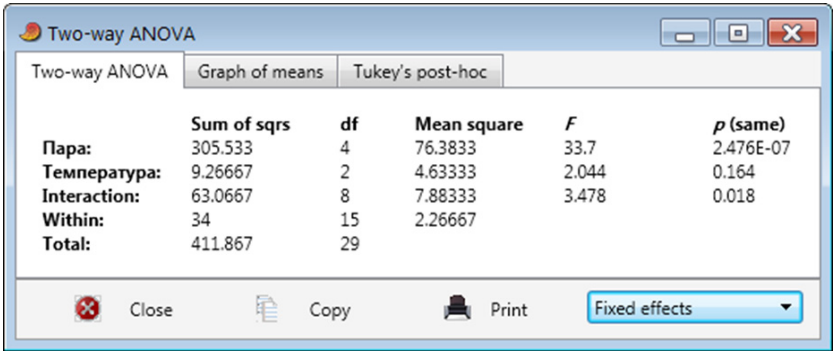
① Внесение данных в пакет. Все данные вносятся в один столбец «Количество», а в два дополнительных столбца «Пара» и «Температура» вносятся метки принадлежности к градациям данных факторов. Такие метки не будут обрабатываться как цифровые значения, а потому их можно задать также буквами или словами. Для задания столбцов с метками как факторов, не забудьте



	Количество	Пара	Температура	D
1	1	1	20	
2	1	1	20	
3	11	2	20	
4	9	2	20	
5	4	3	20	
6	3	3	20	
7	10	4	20	
8	7	4	20	
9	2	5	20	
10	0	5	20	
11	0	1	25	

в ☐ Column attributes в строке Type изменить прочерк на Group. После этого рядом с названием колонки появится синий значок G, сигнализирующий о том, что данная колонка содержит метки *группирующей переменной*. Далее галочку ☑ Column attributes можно снять, сохранить файл и выделить все три колонки.

② Путь: Univariate — ANOVA etc. (several samples) — Two-way ANOVA.



	Sum of sqrs	df	Mean square	*F*	*p* (same)
Пара:	305.533	4	76.3833	33.7	2.476E-07
Температура:	9.26667	2	4.63333	2.044	0.164
Interaction:	63.0667	8	7.88333	3.478	0.018
Within:	34	15	2.26667		
Total:	411.867	29			

По умолчанию открывается форма на закладке Two-way ANOVA:

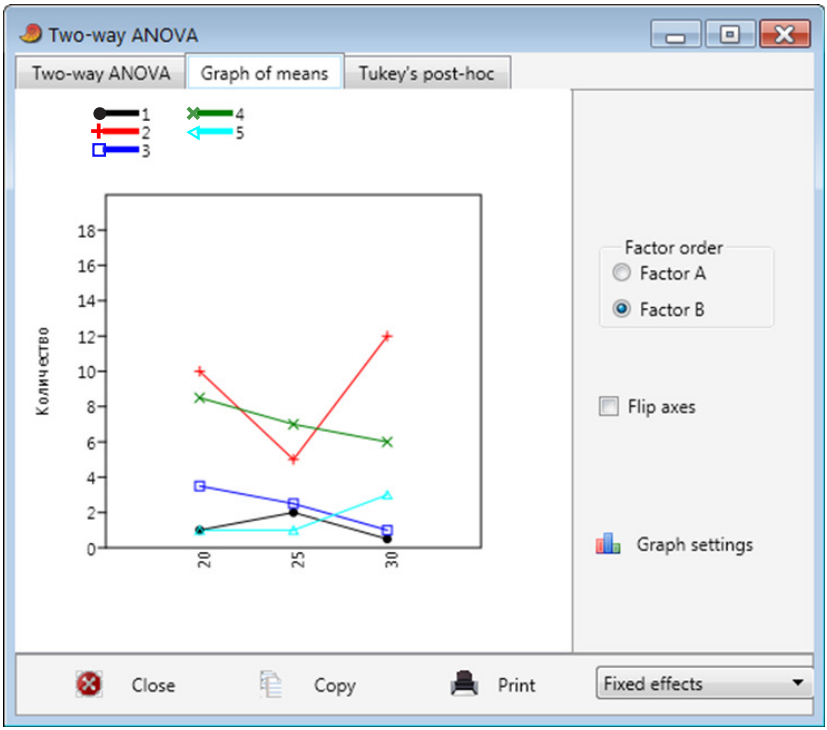
> **Комментарий**. Обратите внимание на автоматически выбранную опцию [Fixed effects], указывающую на модель I ДА с фиксированными эффектами. Также можно выбрать модель II со случайными эффектами (Ranom effects) и её иерархический вариант (Nested effect), но в текущей версии пакета (3.19) нельзя выбрать смешанную модель (Mixed effects model), сочетающую фиксированные и случайные факторы.

Как видно из таблицы результатов, высоко статистически значимым оказался эффект пары, то есть в разных родительских парах дрозофил выживаемость потомков существенно различалась. Влияния температуры в данном эксперименте обнаружено не было, однако взаимодействие факторов «Пара × Температура» было значимым. Пока это можно трактовать так: в среднем температура не влияет, но для каких-то пар её эффект неслучаен.

Данные из этой таблицы необходимо перенести в таблицу результатов дисперсионного анализа (см. п. 5). Такие таблицы принято вставлять в отчёты, публикации и квалификационные работы, если их не слишком много (например, не более пяти). Если дисперсионных анализов и таблиц результатов в работе много, их можно вынести в отдельное приложение на 3–5 страниц

«Результаты дисперсионных анализов». Если же анализов в работе десятки, приводить таблицы целиком не следует, можно ограничиться краткой формой подтверждения статистической значимости: $F_{(df_1;\, df_2)} = \ldots; P = \ldots$

③ Закладка Graph of means (График средних). Устанавливаем радиометку ◉ в тот фактор, который делает рисунок более удобным для восприятия. В нашем случае взаимодействия факторов удобнее обнаружить, если линией профиля будет родительская пара дрозофил:



Из этого рисунка видно, что профили пяти пар мух отличались. Особенно выделились пары 2-я и 4-я: для них была характерна не только повышенная смертность потомков, но и отчётливо проявилось влияние температуры. Именно ситуация в парах 2-й и 4-й преимущественно и обеспечила статистическую значимость взаимодействия.

④ На последней закладке Tukey's post-hoc можно посмотреть результаты апостериорных сравнений для статистически значимых эффектов.

⑤ **Оформление в квалификационной работе (вариант)**.

5.1. Статистическая часть раздела «Материалы и методы».

Сравнения нескольких групп по количественным показателям с приблизительно нормальным распределением проводили в ходе однофакторного дисперсионного анализа. Проверку требований метода осуществляли с помощью критериев Ливена — для оценки однородности дисперсий и Шапиро — Уилка — для оценки нормальности распределения ошибки. Множественные апостериорные сравнения средних в рамках дисперсионного комплекса проводили методом Тьюки. Эффекты считали статистически значимыми при $p \leq 0{,}05$, незначимыми — при $p > 0{,}10$, в промежуточных случаях ($0{,}05 < p \leq 0{,}10$) обсуждали тенденции к различиям. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

5.2. Раздел «Результаты и обсуждение».

Обычно таблице результатов ДА предшествует таблица средних значений со стандартным отклонением или с 95% ДИ…

Результаты анализа представлены в табл. 1.

Таблица 1 — Результаты дисперсионного анализа эксперимента по выживаемости потомков дрозофил

Источник изменчивости	Сумма квадратов *SS*	Степени свободы *df*	Средний квадрат *MS*	F-критерий	Оценка значимости *P*
Пара	305,53	4	76,38	33,70	**<<0,001**
Температура	9,27	2	4,63	2,04	0,164
Взаимодействие факторов	63,07	8	7,88	3,48	**0,018**
Ошибка	34,00	15	2,27	–	–
Общая	411,87	29	–	–	–

Примечание. Жирным шрифтом выделены статистически значимые эффекты.

Далее проводится обсуждение результатов. Для наглядного подтверждения результатов приводятся графики профилей со средними значениями, которые желательно снабдить 95% ДИ (невозможно сделать средствами PAST 3.19).

5.3. Раздел «Выводы».

В ходе двухфакторного дисперсионного анализа было установлено, что изменения температуры в диапазоне 20–30 °C не оказывали влияния на личиночную смертность дрозофил: критерий Снедекора — Фишера $F_{(2;\ 15)} = 2{,}04$; $p = 0{,}164$. Более существенное влияние демонстрировали генетические факторы и их взаимодействие с эффектом температуры: соответственно, $F_{(4;\ 15)} = 33{,}70$; $p << 0{,}001$ и $F_{(8;\ 15)} = 3{,}48$; $p = 0{,}018$.

II. Анализ повторных измерений

Достаточно распространённым на практике является такой экспериментальный план, когда одни и те же объекты измеряются не однократно, а в течение какого-либо отрезка времени. В экологических работах это могут быть, например, численности видов на нескольких пробных площадках, оцениваемые в течение года с интервалом в месяц. В медицинских исследованиях — показатели состояния пациентов, оцениваемые либо до и после лечения, либо, например, до операции и спустя 1, 3, 7, 14 суток после операции. При анализе таких данных необходимо учесть, что объекты на разных сроках представляют собой зависимые выборки, то есть речь идёт о ***повторных измерениях*** (*repeated measurements*) одних и тех же объектов. В целом дизайны подобных исследований могут быть весьма сложными, а их грамотный анализ может потребовать искушённости именно в данной области биостатистики. Для знакомства с этой областью отсылаем читателя к теоретическому материалу, в котором полезно начать с учебника Монтгомери [7].

Для краткого знакомства с анализом повторных измерений рассмотрим непараметрический аналог дисперсионного анализа повторных измерений — ***критерий Фридмана*** (*Friedman test*). Этот ранговый критерий интересен тем, что может быть использован для решения сразу четырёх типов задач: 1) классического анализа повторных измерений; 2) двухфакторного дисперсионного анализа *с единственным наблюдением на ячейку комплекса*; 3) двухфакторного дисперсионного анализа с дизайном рандомизированных блоков; 4) для оценки степени согласия (коррелированности, ***конкордации***) нескольких показателей или экспертов, что применяется в практике *экспертных оценок*.

П р и м е р. На станции озера в течение трёх летних дней, с интервалом в два дня, измерялась температура воды (в градусах по Цельсию) на пяти глубинах. Получены данные:

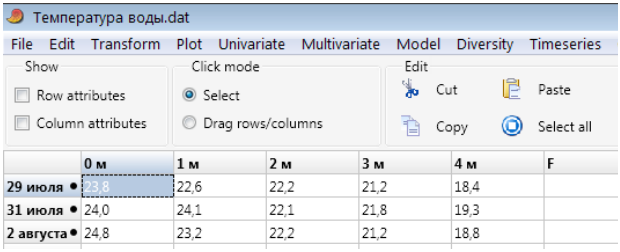
Глубина, м (a = 5)	Дни (b = 3)		
	29 июля	31 июля	2 августа
0	23,8	24,0	24,8
1	22,6	24,1	23,2
2	22,2	22,1	22,2
3	21,2	21,8	21,2
4	18,4	19,3	18,8

Задание. Оценить статистическую значимость различий глубин по средней температуре воды. Определить степень согласованности (конкордации) изменения температуры по глубинам для разных дней.

В пакете PAST

① Внести данные и сохранить файл «Температура_воды.dat». Выделить область данных.

② Использованный способ организации данных является привычным для пользователей: повторные наблюдения для одних и тех же объектов располагаются последовательно в столбцах. Однако для правильного задания расчёта необходимо определиться, что в исследовании является более важным: различия между объектами или между сроками. В нашем случае интерес могут представлять оба значения, однако мы хотим сделать акцент именно на температурной стратификации толщи воды, то есть оценить, различаются ли разные глубины средней температурой воды, и использовать для этого в качестве повторностей разные дни наблюдения. Поэтому матрицу данных следует *транспонировать* так, чтобы сравниваемы группы оказались в колонках: Путь: Edit — Rearrange — Transpose:



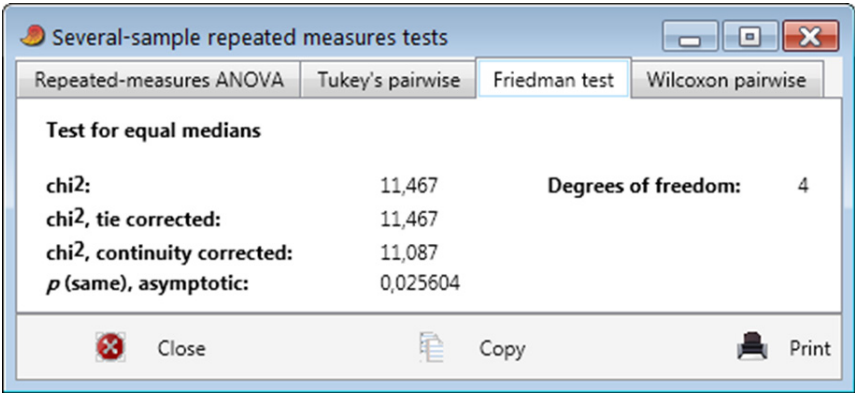
При транспонировании пакет может добавлять дополнительные пустые строки и колонки; на это можно не обращать внимания.

Комментарий 1. Как уже отмечалось выше, критерий Фридмана — ранговый метод. Для нашего примера ранжирование групп (глубин) даёт следующую таблицу рангов:

5	4	3	2	1
4	5	3	2	1
5	4	3	2	1

Замена температур рангами позволяет более отчётливо увидеть закономерность: минимальные температуры наблюдаются на глубине 4 м, максимальные — на поверхности. В отсутствие изменчивости будет наблюдаться полное согласие рангов, как для глубин 2–4 м в нашем примере. Однако в результате изменчивости идеальное согласие нарушается, а критерий Фридмана позволяет оценить вероятность того, насколько такое нарушение может быть случайным. Если различия между глубинами окажутся статистически значимыми, мы можем рассчитать средние ранги и использовать их в качестве обобщающих значений вместо средних значений.

③ Выделяем данные. Путь: Univariate — ANOVA etc. (several samples) — Several sample repeated measure test. Закладка Friedman test:



④ Из результатов анализа видно, что статистка критерия Фридмана имеет распределение хи-квадрат с числом степеней свободы $df = n_{групп} - 1$.

Выписываем хи-квадрат с поправкой на связанные значения (*tie corrected*), степени свободы, *p*-значение.

⑤ **Вывод 1**. Разные глубины статистически значимо различались температурой воды: критерий Фридмана $\chi^2_{(4)} = 11{,}47$; $P = 0{,}026$.

⑥ Оценка степени *согласованности* (*конкордации*) изменения температуры по глубинам для разных дней.

Для любых двух дней мы можем оценить степень сходства (корреляции) распределения температуры по глубинам. С несколькими мерами оценки таких парных связей мы познакомимся в лабораторной работе № 11. Однако в нашем случае дней было три и интерес может представлять оценка степени сходства по всем трём дням. Для таких случаев рассчитывается ***коэффициент конкордации Кендалла*** (*Kendall's concordation*). Он изменяется от 0 (полное отсутствие согласия) до 1 (идеальное согласие) и по силе может быть интерпретирован как меры корреляции (см. теоретический материал и лабораторную работу № 11). Судя по таблице рангов, которую мы рассматривали на этапе 2, в нашем случае можно ожидать весьма высокой конкордации — близкой к 1.

В пакете PAST (версия 3.19) коэффициент конкордации Кендалла отсутствует, однако его можно легко рассчитать из величины статистики хи-квадрат Фридмана:

$$W = \frac{\chi^2}{b(a-1)},$$

где a — число групп, b — число повторностей для группы.

В нашем случае $a = 5$, $b = 3$ и $\chi^2 = 11{,}467$. Тогда

$$W = \frac{11{,}467}{3(5-1)} = \frac{11{,}467}{12} = 0{,}96$$ (достаточно округлить до сотых).

Поскольку полученное значение очень близко к 1, мы можем констатировать очень высокое сходство распределения температур по глубинам в разные дни. **ВАЖНО!** Если цель исследования заключается не в оценке различий, а в оценке именно сходства, конкордации, то величину критерия Фридмана можно не приводить, а использовать его только для расчёта и оценки статистической значимости конкордации Кендалла.

⑦ **Вывод 2**. Обнаружена высокая степень согласованности изменения температур по глубинам озера в разные дни: коэффициент конкордации Кендалла $W = 0{,}96$; $P = 0{,}026$.

⑧ **Оформление в квалификационной работе** проводится как обычно, однако необходимо акцентировать описание либо для задачи сравнения групп, либо для задачи поиска конкордации. Например:

а) для оценки различий глубин по средней температуре воды использовали анализ Фридмана;

б) для оценки согласованности изменения температур по глубинам озера в разные дни по результатам анализа Фридмана рассчитывали коэффициент конкордации Кендалла.

> **К сведению**. Конкордация Кендалла очень удобна для оценки согласия экспертов по какому-либо вопросу, а потому широко применяется в ***анализе экспертных оценок***. Предположим, что нам нужно выбрать лучшую схему лечения заболевания из пяти используемых во врачебной практике. Для этого можно спланировать и провести соответствующий эксперимент. Однако можно поступить проще: найти нескольких заслуживающих доверия экспертов, способных ранжировать все пять схем. Каждый метод эксперты оценивают, например, по 5-балльной шкале. Если в оценках экспертов наблюдается статистически значимое согласие, то в качестве интегрального показателя эффективности схемы лечения можно использовать ***средние ранги*** экспертных оценок и выбрать лучшую схему лечения.
>
> Также оценка конкордации находит широкое применение в экологических исследованиях, когда роль повторностей или экспертов отводится разным признакам. Например, если в градиенте техногенной нагрузки признаки ведут себя согласованно (согласованно изменяется уровень флуктуирующей асимметрии разных признаков, согласованно изменяется численность разных видов организмов и т. п.), значит факт техногенного воздействия можно считать доказанным, а также можно вычислить интегральный показатель импактного воздействия на территории — средний ранг для разных территорий.

 **Домашнее задание**. Проведите анализ Фридмана для исходного файла «Температура_воды.dat», то есть до транспонирования матрицы данных. Различаются ли три изученных дня средней температурой? Рассчитайте коэффициент конкордации Кендалла по результатам проведённого анализа и интерпретируйте его результаты. Оформите выводы.

ЛАБОРАТОРНАЯ РАБОТА № 11

Анализ связей между показателями. Графическое представление связей

Тема 10. Анализ связей. Корреляция и ассоциация.
Количество часов: 2.
Цель: Освоить основные методы корреляционного анализа: корреляцию Пирсона, непараметрическую корреляцию Спирмена и Кендалла, меры ассоциации для таблиц сопряжённости. Научиться строить диаграммы рассеяния с доверительным эллипсом. Работа на ПК.

Анализ связей между показателями — крайне важный этап научного исследования, особенно в его начальной части. Рассуждения о природе обнаруженных связей позволяют выделять из них возможные *причинно-следственные связи*, которые далее можно проверять экспериментально и пытаться использовать на практике. Традиционно связь между двумя и более количественными и порядковыми показателями называют ***корреляцией*** (*correlation*), а связь между качественными номинальными показателями — ***ассоциацией*** (*association*).

Приступая к поиску связей, важно помнить, что статистически значимая корреляция или ассоциация необязательно подразумевает отношения между показателями по типу «причина — следствие». Для пары связанных показателей X_1 и X_2, возможно:

1) X_1 является причиной X_2;
2) X_2 является причиной X_1;
3) X_1 и X_2 являются следствиями одной причины N;
4) связь обусловлена более сложными механизмами с вовлечением большого числа показателей.

В силу неизвестных отношений между X_1 и X_2 следует избегать употребления термина «взаимосвязь», подразумевающего взаимную обусловленность показателей — конкретный и далеко не очевидный тип связи, нуждающийся в доказательстве. Грамотнее использовать термины «связь», «ассоциация», «корреляция».

Любая корреляция и ассоциация может и должна быть охарактеризована следующими тремя характеристиками.

1) *направление связи*: прямая (с увеличением X_1 увеличивается X_2) или обратная (с увеличением X_1 уменьшается X_2). На *прямую*

связь указывает положительное значение коэффициента корреляции, на *обратную связь* — отрицательное значение. Также используются термины «*положительная корреляция*» и «*отрицательная корреляция*»;

2) *сила связи*; количественно характеризуется величиной ***коэффициента корреляции (ассоциации)***. Коэффициенты корреляции изменяются от –1 (идеальная обратная связь) через 0 (отсутствие связи) до +1 (идеальная прямая связь). Коэффициенты ассоциации изменяются от 0 (отсутствие связи) до +1 (максимально возможная связь). В биологии и медицине считается, что, если абсолютное значение коэффициента находится в интервале

(0; 0,3] — связь слабая,

(0,3; 0,7] — связь средней силы,

(0,7; 1] — связь сильная;

3) *статистическая значимость* — самостоятельная характеристика, зависящая как от силы связи, так и от объёма выборки. На очень маленькой выборке достаточно сильная связь может оказаться незначимой ($P > 0{,}05$), а на очень большой выборке можно обнаружить значимой ($P < 0{,}05$) даже очень слабую связь.

Наиболее частой и простой практической задачей является установление связи между двумя показателями — *парной связи*. Однако если показателей в исследовании несколько, бывает важно установить степень их совместной корреляции — ***конкордации*** (*concordation*). Это можно сделать путём расчёта коэффициента конкордации Кендалла (*Kendall's concordation*) — непараметрического рангового показателя, который был рассмотрен в лабораторной работе № 10. Также при наличии большого числа показателей информация о парных связях может быть использована далее для выявления целых групп связанных показателей, в основе совместного варьирования которых могут лежать одни и те же процессы. С такими *многомерными методами* мы познакомимся ближе к концу курса (см. лабораторную работу № 16).

I. Количественные признаки с нормальным распределением

Установление связи проводится в ходе ***линейного корреляционного анализа*** по Пирсону (*correlation analysis*, *Pearson's correlation*). Когда говорят о корреляции и не уточняют деталей — подразумевается именно данная техника. В качестве по-

казателя направления и силы связи используется ***коэффициент корреляции Пирсона***, который обозначается буквой *r*. Квадрат этой величины называется ***коэффициентом детерминации*** (*coefficient of determination*), который обозначается R^2. Он изменяется от 0 до 1 (или от 0 до 100 %) и показывает долю общей дисперсии в суммарной дисперсии данных.

Метод является параметрическим. В основе техники линейной корреляции лежит предположение о том, что данные извлечены из *двумерного нормального распределения*. Поэтому перед использованием этого статистического метода необходимо убедиться, по крайней мере, в симметрии и унимодальности распределений каждого из признаков. Для унимодальных, но асимметрично распределённых признаков следует использовать нормализующие преобразования данных или переходить к ранговым корреляциям (см. далее).

П р и м е р. Проанализируем данные по содержанию билирубина в образцах венозной и капиллярной крови 16 пациентов из лабораторной работы № 7, где они использовались в качестве примера для сравнения двух зависимых выборок. Напомним, что было обнаружено статистически значимо меньшее содержание билирубина в капиллярной крови по сравнению с венозной. Однако возможно, что между этими двумя показателями существует связь. Если она будет достаточно сильная, то её можно положить в основу модели, позволяющей пересчитывать содержание билирубина в капиллярной крови на его содержание в венозной крови.

Данные: содержание общего билирубина (в мкмоль/л) в сыворотке крови 16 пациентов. Файл: Билирубин.dat.

Задание. Рассчитать коэффициент корреляции Пирсона между показателями. Сделать вывод о наличии, направлении, силе и статистической значимости связи. Построить график.

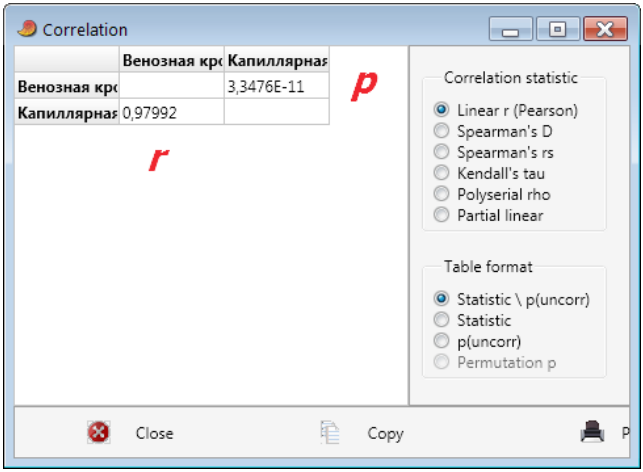
В пакете PAST

① Открыть файл «Билирубин.dat» и выделить область данных.

② Путь: Univariate — Correlation. Радиометка в ◉ Linear r (Pearson).

В таблице результатов искомая статистика расположена в нижнем левом углу, а *p*-значение — в верхнем правом. Подсказка дана в области Table format.

Мы видим, что корреляция очень сильная, поскольку r близок к 1. Также она высоко статистически значима. Напомним, что *экспоненциальная форма* записи 3,3476E-11 означает $p = 3{,}3476 \times 10^{-11}$, что намного меньше даже 0,001 ($p << 0{,}001$).



③ Выписываем r и p; округляем: r — до сотых, p — до тысячных. Результат: $r = 0{,}98$; $p << 0{,}001$.

④ **Вывод**. Обнаружена сильная положительная высоко статистически значимая связь между концентрацией билирубина общего в венозной и капиллярной крови: коэффициент корреляции Пирсона $r = 0{,}98$; $P << 0{,}001$.

⑤ Построение графика — ***диаграммы рассеяния*** (*scattergram*). Путь: Plot — XY graph. В случае использования линейной корреляции Пирсона мы имеем право обвести облако точек *95%-ным* ***доверительным эллипсом***: ☑ 95 % ellipses. Чем у́же эллипс — тем сильнее связь. Полученный рисунок можно доработать и вставлять в статью или квалификационную работу.

Задание. Посмотрите внимательно на полученный рисунок. Что в нём выглядит неестественно?

Доверительный эллипс уходит в область отрицательных значений, тогда как концентрация не может быть отрицательной. Это значит, что модель двумерного нормального распределения, лежащая в основе корреляции Пирсона, не вполне подходит к нашим данным. Постройте гистограммы распределения признаков

(можно сразу при выделенных двух колонках — разные признаки будут отражены разным цветом). Далее прологарифмируйте исходные данные и посмотрите, что изменилось на гистограммах распределений. Заново рассчитайте корреляцию на логарифмах данных и постройте график с доверительным эллипсом. **Вопрос**: Помогло ли логарифмическое преобразование шкал устранить проблему асимметрии распределений и отрицательных значений внутри доверительного эллипса?

II. Количественные признаки с ненормальным распределением и порядковые признаки

Для количественных признаков с ненормальным распределением можно использовать корреляцию Пирсона после нормализующих преобразований (логарифмирование, преобразование Бокса — Кокса, угловые преобразования для долей и т. д. — см. теоретический материал). Этот подход мы только что опробовали и для логарифмического преобразования нашли его эффективным, хотя и не в полной мере.

Чаще всего в случае сомнений относительно нормальности распределений показателей и/или линейности связи используют ранговые корреляции: ***корреляцию Спирмена*** r_S, ρ (греческая буква «ро») или ***корреляцию Кендалла*** τ (греческая буква «тау»). Поскольку многие биологические признаки имеют ненормальное распределение, а связи в биологических системах чаще нелинейны, использование непараметрических ранговых корреляций нередко оказывается более удобным или даже предпочтительным (по крайней мере — на стадии разведочного анализа данных).

Корреляция Спирмена — это прямой ранговый аналог корреляции Пирсона. Её формулу можно вывести из формулы корреляции Пирсона, если от параметрической статистики перейти к порядковой. Корреляция Кендалла также основана на рангах, но выведена из других теоретических построений о природе связи как таковой и имеет вероятностную интерпретацию (см. теоретический материал). Коэффициенты корреляций Спирмена и Кендалла изменяются от –1 до 1, но на практике всегда $\tau < r_S$. В силу ряда причин у обеих статистических техник есть свои приверженцы.

В пакете PAST

① Открыть файл «Билирубин.dat» и выделить область данных. Проще отменить последнюю операцию логарифмирования и восстановить исходные данные: Edit — Undo.

② Путь: Univariate — Correlation. Радиометка в ◉ Spearman's rs, затем — в ◉ Kendall's tau.

③ Округление: r_S или τ — до сотых, p — до тысячных.

④ **Вывод**. Обнаружена положительная высоко статистически значимая связь между концентрацией билирубина общего в венозной и капиллярной крови: коэффициент корреляции Спирмена $r_S = 0{,}94$; $P << 0{,}001$. (Если использовалась корреляция Кендалла — аналогично. Результаты по обеим корреляциям приводить не следует, нужно выбрать какую-то одну).

⑤ Построение графика — диаграммы рассеяния. Путь такой же: Plot — XY graph. Но обводить облако точек доверительным эллипсом в случае ранговых корреляций мы не имеем права. График следует доработать (см. рис. 11.1) и можно вставлять в работу.

Задание. Посмотрите внимательно на диаграмму рассеяния. Какой из вариантов *корреляционного квартета Анскомба* более всего подходит к нашей ситуации и почему? Для выполнения задания ознакомьтесь с теоретическим материалом.

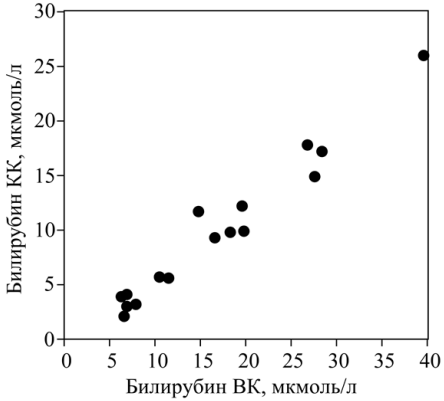


Рис. 11.1. Содержание билирубина общего в сыворотке венозной (ВК) и капиллярной (КК) крови

Домашнее задание. Помимо обычной диаграммы рассеяния в случае ненормально распределённых показателей их кор-

реляцию можно изобразить на специальном типе графика — ***мешочной диаграмме*** (*bagplot*), сочетающем диаграмму рассеяния и двумерный коробчатый график. Воспользуйтесь онлайновым калькулятором по адресу https://www.wessa.net/rwasp_bagplot.wasp и постройте такой график для данных по билирубину в капиллярной и венозной крови. Десятичную запятую на точку можно не менять (программа изменит автоматически), показывать потенциальные выбросы не нужно (Show outliers — FALSE), сделайте график квадратным 500×500. Полученный рисунок можно сохранить в формате *.png и доработать в растровом графическом редакторе. Распечатайте рисунок и вклейте в тетрадь в конце этого занятия. Русские названия осей можете написать вручную. Однако можно также сохранить его в формате *.ps (PostScript), сконвертировать в формат *.svg с помощью любого онлайнового конвертера (например: https://convertio.co/ru/vector-converter) и доработать в TpX: разгруппировать (Modify — Ungroup), изменить толщину линий, цвета и т. д. (рис. 11.2).

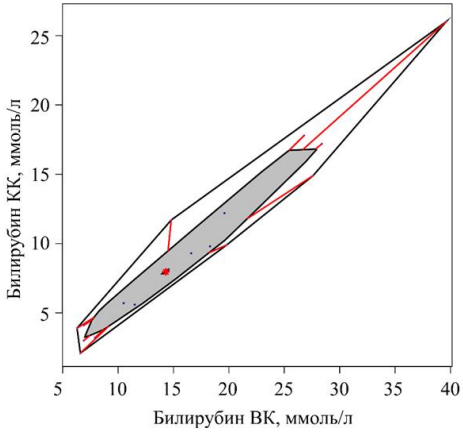


Рис. 11.2. Содержание билирубина общего в сыворотке венозной (ВК) и капиллярной (КК) крови

III. Качественные номинальные признаки

Для качественных номинальных признаков оценка силы связи (ассоциации) признаков проводится по таблицам сопряжённо-

сти (ТС). Ранее мы уже использовали ТС для анализа различий в примере с уровнем холестерина и заболеваниями сердца. В этом нет противоречий: по ТС можно найти как различия между частотами, так и ассоциацию признаков. Что выбрать — зависит от формулировки задачи. В лабораторной работе № 6 мы на первом этапе рассчитывали критерий хи-квадрат, а на втором этапе — для оценки силы различий — вычисляли специфические меры различий: относительный риск и отношение шансов. Для оценки ассоциации на первом этапе мы также рассчитываем критерии типа хи-квадрат, а на втором этапе — специфические меры, но только уже ассоциации.

Мер для оценки ассоциации качественных признаков предложено много. Большинство из них напрямую задействуют в вычислениях значение критерия хи-квадрат Пирсона (см. теоретический материал). Наиболее часто используются на практике: ***коэффициент сопряжённости Пирсона*** (в том числе — в модификации Сакоды), а также ***коэффициенты ассоциации Крамера*** или ***Чупрова***; последние в случае таблиц 2×2 дают одинаковое значение. Все эти коэффициенты изменяются от 0 (отсутствие связи) до +1 (максимально возможная связь). Если направление связи возможно установить, то о нём судят по частотам ТС (см. ниже).

П р и м е р. Проанализируем уже знакомые данные по холестерину и заболеваниям сердца:

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Сформулируем задачу не в терминах поиска различий, а в терминах поиска связи. **Задание**: определить, существует ли связь между уровнем холестерина и заболеваниями ССС?

В пакете PAST

① Вносим четыре значения данных в соседние ячейки и выделяем.

② Путь: Univariate — Contingency table. (Текущая версия пакета не может вычислить точный критерий Фишера, о чём сообщает в окне предупреждения). Смотрим раздел Other statistics (Другие статистики).

③ Выписываем коэффициент ассоциации Крамера (*Cramer's V*) или коэффициент сопряжённости Пирсона (*Contingency C*). Обе меры приводить не следует, нужно выбрать какую-то одну. Видно, что обе меры близки и с точностью до десятых равны 0,15. Эти значения лежат в интервале (0; 0,3], что соответствует «слабой» по силе ассоциации.

> **К сведению**. Выше мы отметили, что многие меры ассоциации могут быть переписаны формулами со значением статистики хи-квадрат. В рассмотренном примере $\chi^2 = 31{,}082$. Для расчётов также понадобится значение общей суммы: $n = 1329$.
>
> ***Коэффициент ассоциации Крамера***
>
> $$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}},$$
>
> где min $(r - 1, c - 1)$ — минимальное из двух значений: числа рядов или числа колонок таблицы за вычетом единицы. Для таблицы 2×2 это всегда 1.
>
> $$V = \sqrt{\frac{31{,}082}{1329 \cdot 1}} = 0{,}152929753 \text{ , или } 0{,}153.$$
>
> ***Коэффициент сопряжённости Пирсона***
>
> $$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{31{,}082}{1329 + 31{,}082}} = 0{,}151172199, \text{ или } 0{,}152.$$
>
> Результаты расчётов по формулам совпадают с результатом программы PAST.

④ Направление связи в данном случае определить возможно, так как в основе первого входа ТС (Уровень холестерина) лежало разбиение на категории количественного признака. Поскольку с увеличением уровня холестерина доля больных также увеличивалась, определяем направление связи как «прямое».

> **К сведению**. Когда оба входа ТС представлены не упорядоченными, а номинальными категориями, установить направление связи невозможно: оно не имеет смысла. Например, ассоциация цвета и формы объекта.

Вопрос: можно ли установить направление связи для ассоциации наличия седины волос с тремя категориями возраста (молодой, средний, пожилой)? Седины волос с двумя категориями стресса на работе (слабый и высокий)?

⑤ Округляем коэффициенты с точностью до сотых или ты-

сячных. Значение *p* берём из результатов критерия хи-квадрат, а если в таблице были значения ≤5 — берём *p*, вычисленное рандомизационным критерием Монте-Карло.

⑥ **Вывод**. Обнаружена слабая, но высоко статистически значимая прямая связь между уровнем холестерина в сыворотке и заболеваниями сердечно-сосудистой системы: коэффициент ассоциации Крамера $V = 0{,}15$; $P << 0{,}001$.

⑦ Графики. Графическое представление парных ассоциаций не является распространённой практикой. Для этой цели можно использовать мозаичные графики и ***диаграммы Венна*** (*Venn's diagram*).

⑧ **Оформление в квалификационной работе (вариант)**.

8.1. Раздел «Материалы и методы».

Для поиска связей между количественными показателями использовали ранговый корреляционный анализ по Спирмену, а между качественными показателями — расчёт коэффициента ассоциации Крамера. В последнем случае оценку статистической значимости проводили в ходе анализа таблиц сопряжённости с использованием рандомизационной процедуры Монте-Карло ($n = 99\,999$). Связи считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденции к связям. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

8.2. Раздел «Результаты и обсуждение».

Даются таблицы с коэффициентами корреляции и/или ассоциации и соответствующими *p*-значениями. Для количественных признаков можно привести диаграммы рассеяния. Обсуждение связей строится на основе трёх критериев: значимости, направлении и силе связи. По возможности в обсуждение встраиваются блоки, объясняющие наблюдаемые связи. Обычно наибольший интерес представляют те связи, в основе которых могут лежать причинно-следственные отношения; но не обязательно. Бывает, что, напротив, наибольший интерес представляет отсутствие ожидаемой связи между показателями или нелинейная связь (видна на диаграмме рассеяния), коэффициент корреляции для которой может быть близким к нулю и др. Все эти моменты следует обсудить.

8.3. Раздел «Выводы». См. примеры из разделов I, II и III данной лабораторной работы.

ЛАБОРАТОРНАЯ РАБОТА № 12

Анализ зависимостей. Линейная регрессия

Тема 11. Анализ зависимостей. Линейная регрессия.
Количество часов: 2.
Цель: Освоить правило выбора и использование трёх методов линейного регрессионного анализа: обычная регрессия (OLS), главные оси (MA) и сокращённые главные оси (RMA). Работа на ПК.

Анализ зависимостей проводится в ходе ***регрессионного анализа*** (*regression analysis*). Этот раздел прикладной статистики очень хорошо разработан, поскольку регрессионные модели позволяют: 1) обобщить и представить зависимость между показателями в виде функции; 2) использовать её для анализа явления и 3) для прогноза. Поэтому существует большое число вариантов регрессионного анализа: линейный и нелинейный, парный и множественный, для количественных показателей и качественных и т. д. (см. теоретическую часть).

В отличие от корреляции, где все переменные обрабатываются как равноценные (X_1, X_2, … X_n — матрица **X**), в регрессии выделяется два блока данных:

1) матрица **X** — ***независимые переменные*** (*independent variables*), которые называются ***регрессорами,*** или ***предикторами***;

2) матрица **Y** — ***зависимые переменные*** (*dependent variables*), которые называются переменными ***отклика*** и изменяются в зависимости от значений предикторов.

Деление показателей на эти две группы определяется дизайном и/или целями исследования. На этом занятии мы познакомимся с простыми регрессионными техниками, включающими только один предиктор X и один отклик Y.

I. Количественные признаки с нормальным распределением

Рассмотрим кратко основные понятия ***линейной регрессии*** (рис. 12.1).

Линейная регрессия вида $y = ax + b$ имеет два параметра: 1) ***коэффициент регрессии*** a (*slope* — «наклон»), равный тангенсу угла наклона линии к оси x, и 2) ***свободный член*** (*intersept* —

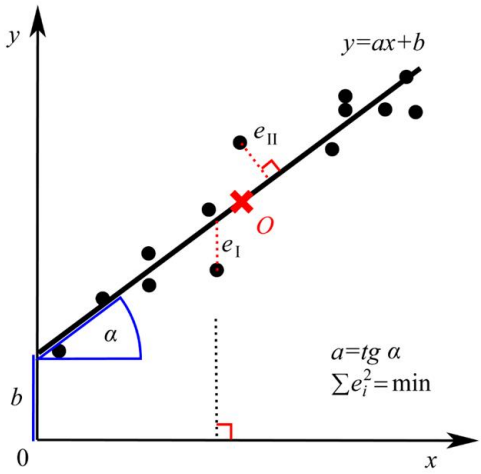


Рис. 12.1. Основные элементы линейной регрессии

«отрезок») b, равный значению функции в точке $x = 0$ и графически представляющий собой отрезок, отсекаемый линией регрессии на оси y. Центр системы O — среднее значение предиктора x и отклика y: $O(\bar{x}, \bar{y})$; он строго фиксирован. А вот линию через этот центр можно провести по-разному, что определяется выбором той или иной модели.

Расстояния e_i называются ***остатками*** (*residuals*); они являются отклонениями реальных значений от предсказанных построенной функцией. Если ***подгонка модели*** (*model fitting*) регрессии к данным проводится таким образом, что минимизируется сумма квадратов таких остатков, то говорят, что зависимость построена ***методом наименьших квадратов*** (*least squares method*).

Подавляющее большинство регрессионных техник — параметрические, поэтому нормализующие преобразования данных здесь весьма распространены. Эти техники подразумевают, что отклик нормально распределён на всех уровнях регрессора. Если это так, то остатки будут распределены нормально, со средним, равным нулю.

Для корректного линейного регрессионного анализа нужно правильно выбрать модель регрессии:

1) ***модель I регрессии*** — классическая регрессия, где независимая переменная X задаётся или контролируется исследовате-

лем, а переменная Y изменяется в зависимости от X. Примеры: изменение биологического показателя в зависимости от концентрации, дозы, возраста и т. п. В данной модели считается, что ошибка регрессора X столь мала, что ей можно пренебречь. Поэтому остатки e_i рассчитываются как перпендикуляр в проекции на ось x (e_I на рис 12.1). В простейшем случае применяется ***обычная линейная регрессия методом наименьших квадратов*** (*Ordinary Least Squares regression, OLS regression*);

2) ***модель II регрессии*** — регрессия, где обе переменные берутся из популяции, а потому содержат ошибки, свойственные выборкам. Часто (но не обязательно) это тот случай, когда правильнее говорить о корреляции, но по каким-то причинам связь нужно изобразить не облаком точек, а одной линией. Выбор предиктора и отклика здесь условен и определяется задачей исследователя. Примеры: зависимость массы тела от роста, численности гетеротрофов от численности автотрофов и т. п.

Поскольку для такого варианта регрессии и предиктор, и отклик имеют выборочные ошибки, остатки e_i рассчитываются как перпендикуляр к линии регрессии (e_{II} на рис 12.1).

В случае если оба показателя модели II измерены в одних единицах, то используют ***регрессию главных осей*** (*Major axes, MA*), а если в разных — ***регрессию стандартных главных осей*** (*Reduced Major Axes, RMA*). В последнем случае линия регрессии проходит по наибольшему диаметру корреляционного эллипса.

П р и м е р. Проанализируем данные по содержанию билирубина в образцах венозной и капиллярной крови 16 пациентов из лабораторных работ № 7 и 11, где они использовались в качестве примера для сравнения двух зависимых выборок и для поиска связи. Поскольку связь между концентрациями общего билирубина (ОБ) в венозной и капиллярной крови была очень сильная и статистически значимая ($r = 0{,}98$; $p << 0{,}001$), имеет смысл построить модель, позволяющую пересчитывать содержание ОБ в капиллярной крови на его содержание в венозной крови. Напомним, что анализ капиллярной крови из пальца является более удобной для пациента процедурой, в то время как существующие представления о значениях этого показателя в норме и при различных патологиях получены на данных по венозной крови.

Вопрос: какую регрессионную технику корректно использовать в данном случае?

Данные: содержание общего билирубина (в мкмоль/л) в сыворотке крови 16 пациентов. Файл: Билирубин.dat.

Задание. Получить формулу линейной зависимости концентрации ОБ в венозной крови от значения этого показателя в капиллярной крови. Изобразить зависимость графически.

В пакете PAST

① Открыть файл «Билирубин.dat» и выделить область данных.

② Реорганизовать данные так, чтобы столбец «Венозная кровь» находился после столбца «Капиллярная кровь». Это необходимо для того, чтобы регрессором выступил ОБ в капиллярной крови, а полученное уравнение позволяло прогнозировать ОБ в венозной крови — как будто кровь пациента была взята из вены.

③ Путь: Model — Linear — Bivariate. Радиометка в ◉ MA (*Major axes*), то есть используем регрессию главных осей.

④ Выписываем значения двух ***параметров регрессии*:**

Slope — *коэффициент регрессии*: $a = 1{,}4856$;

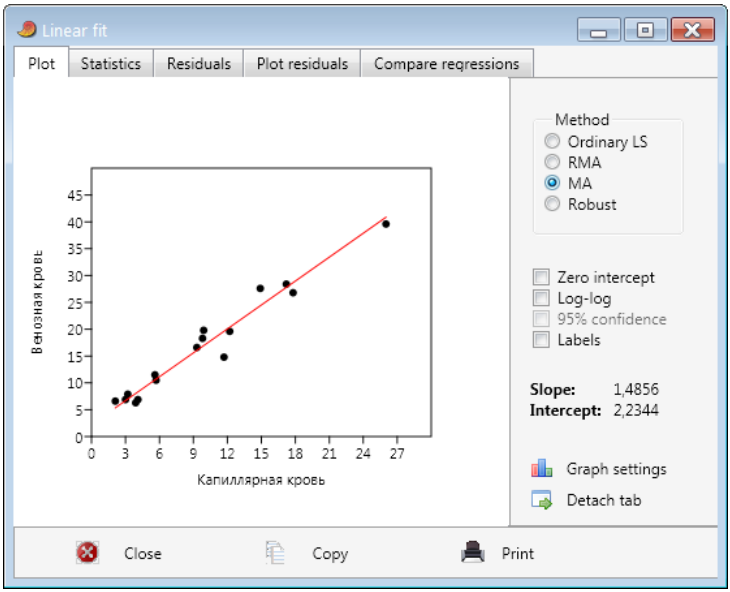
Intercept — *свободный член*: $b = 2{,}2344$.

Оформляем уравнение регрессии в виде $y = ax + b$:

$$y = 1{,}4856x + 2{,}2344$$

Задание. Поместите радиометку в ◉ RMA (*Reduced Major Axes*) и в ◉ Ordinary LS (*Ordinary Least Squares*). Как изменилось уравнение регрессии? Для самого распространённого на практике варианта (OLS) доступна опция построения ***95%-ных доверительных границ для регрессии*** — отобразите их: ☑ 95% confidence. Обратите внимание, что ошибка регрессии минимальна в центре системы (среднее значение показателя по оси X) и увеличивается к периферии. Пакет PAST рассчитывает также вариант робастной регрессии (◉ Robust), который рекомендуется применять, когда есть подозрения на имеющиеся в данных выбросы.

⑤ Верните ◉ MA, закладка Statistics. Здесь приведена более детальная статистка для построенной зависимости. К сожалению, пока пакет PAST не выдаёт статистику оценки значимости собственно регрессии, поэтому мы рассчитаем её немного позже (в п. 7) с использованием расчётной электронной таблицы Excel. На закладке



Statistics мы видим статистику по корреляции, что не вполне корректно, а также 95% ДИ для параметров регрессии, вычисленные бутстрепом. Последние можно использовать для косвенной оценки значимости каждого из параметров регрессии и регрессии в целом:

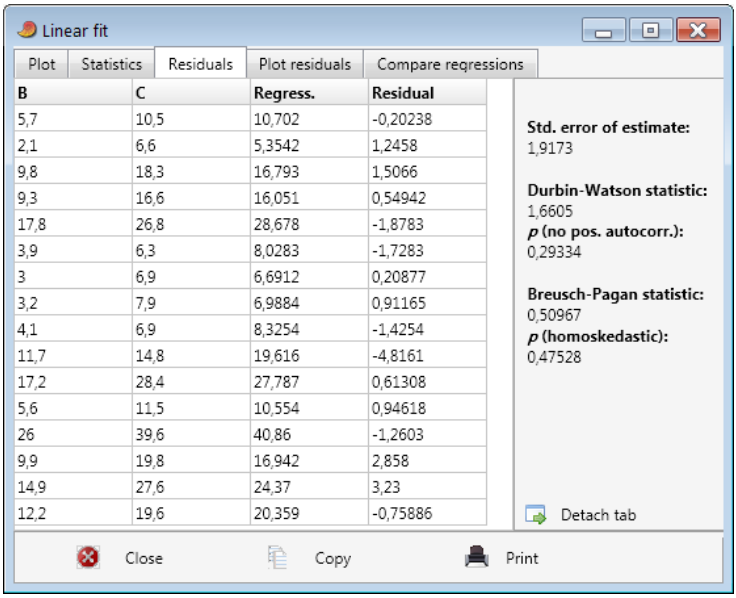
– если 95% ДИ для коэффициента регрессии (Slope *a*) не содержит 1 — регрессия статистически значима ($p \leq 0{,}05$), то есть линия регрессии не параллельна оси *X*, а имеет наклон. Знак этого параметра указывает на направление зависимости: «+» — зависимость прямая (наш случай), «–» — зависимость обратная;

– если 95% ДИ для свободного члена (Intercept *b*) не содержит 0 — он статистически значим, то есть линия выходит не из начала координат (0; 0).

⑥ Закладки Residuals и Plot residuals — анализ остатков. Несмотря на значимость линейной регрессии, такая её форма может не быть оптимальной для данных. Поэтому полезно проанализировать остатки: они должны быть случайно и нормально распределены относительно линии регрессии, со средним, равным 0. ► **Остаток** — отклонение реального значения *Y* от предсказанного для данной точки регрессией. Например, для значения $x_1 = 5{,}7$ было $y_1 = 10{,}5$, а вычисленное по уравнению регрессии значение составило:

$\hat{y}_1 = 1{,}4856 \times 5{,}7 + 2{,}2344 = 10{,}70232$. Соответствующий остаток $r_1 = y_1 - \hat{y}_1 = 10{,}5 - 10{,}70232 = -0{,}20232$.

Задание. Рассчитайте самостоятельно остаток r_2.



6.1. ***Критерий Дарбина — Уотсона*** (*Durbin-Watson test*) проверяет случайность распределения остатков. Если $p \leq 0{,}05$, значит существует ***автокорреляция***: каждое последующее значение остатка зависит от предыдущего. На графике ◉ Scatter plot в Plot residuals автокорреляция будет видна в виде тренда или циклических колебаний относительно нулевой отметки. Автокорреляция может быть устранена подбором более адекватной данным нелинейной зависимости. В нашем случае автокорреляция статистически незначима: критерий Дарбина — Уотсона $DW = 1{,}66$; $p = 0{,}293$.

6.2. ***Критерий Бройша — Пагана*** (*Breusch-Pagan test*) проверяет однородность дисперсии остатков (*homoscedasticity*). Он имеет распределение хи-квадрат с числом степеней свободы $df = k_{\text{параметров модели}} - 1$. Если $p \leq 0{,}05$, значит дисперсии остатков неоднородны (модель гетероскедастична). На графике Scatterplot в Plot residuals это будет выглядеть как расходящийся или сходящийся клин точек относительно нулевой отметки. Неоднородность

может быть устранена преобразованием данных. В нашем случае остатки регрессии были однородны: критерий Бройша — Пагана $\chi^2_{(1)} = 0{,}51$; $p = 0{,}475$.

6.3. На графике ◉ Histogram можно посмотреть гистограмму распределения остатков. Распределение должно быть симметричным и унимодальным. Если распределение отчётливо асимметричное, следует попробовать преобразовать данные перед анализом. В нашем случае, возможно, небольшая асимметрия присутствует, однако малый объём выборки не позволяет исследовать этот вопрос детальнее.

⑦ Оценка статистической значимости регрессии. Как мы уже отметили в п. 5, пакет PAST версии 3.19 не позволяет рассчитать статистическую значимость для всей регрессионной зависимости целиком. Поэтому воспользуемся готовой расчётной таблицей Excel «Регрессия_оценка значимости в ANOVA.xlsx», которая устранит этот недостаток. Скачать её можно по ссылке: https://yadi.sk/d/g50i73pt3J6pAa, в папке «Программы».

В пакете Excel

① Открыть файл «Регрессия_оценка значимости в ANOVA.xlsx» и посмотреть примечание, наведя указатель мыши на ячейку со знаком вопроса «?».

② В пакете PAST копируем в буфер данные из первых двух столбцов, а в электронной таблице вставляем их в столбцы X и Y.

③ Становимся в ячейку E4 и в верхней строке изменяем формулу на требующуюся. Для нашей регрессии главных осей уравнение, полученное в п. 4, было: $y = 1{,}4856x + 2{,}2344$. Значит, в ячейке с формулой нужно ввести

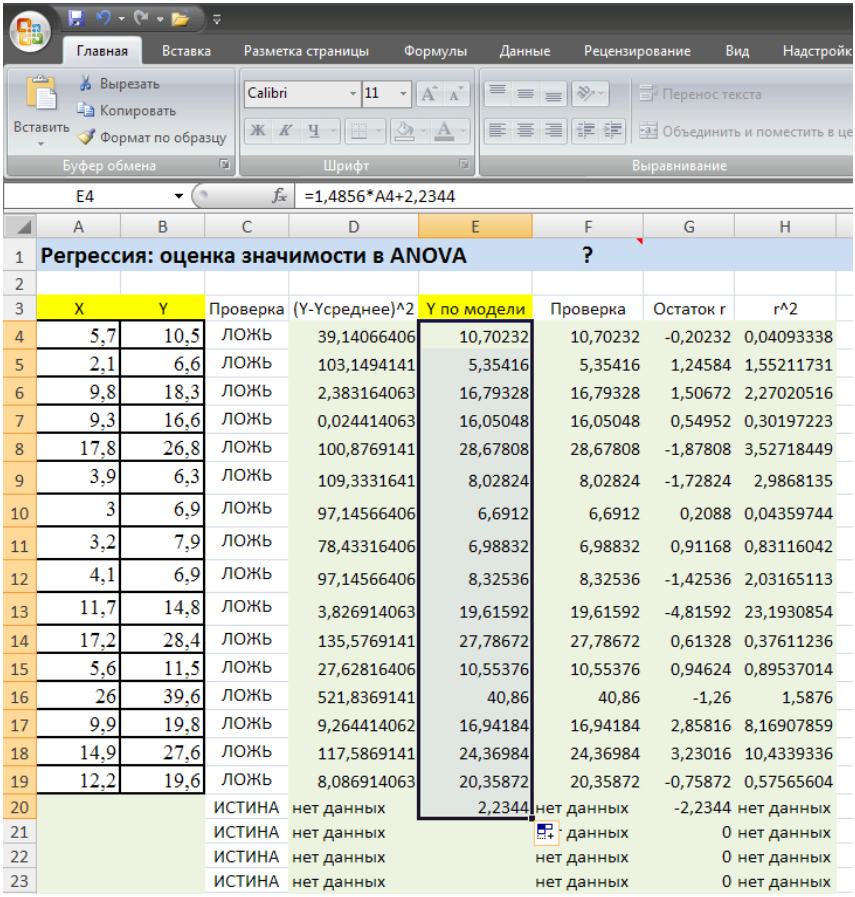
```
=1,4856*A4+2,2344
```

④ Копируем эту ячейку протяжкой в остальные ячейки колонки «Y по модели». Таблица работает с выборками до 500 наблюдений, но во все 500 ячеек копировать необязательно, главное чтобы ячеек с введённой формулой было не меньше чем строк с данными (больше — можно) (см. скриншот на с. 174).

⑤ Копируем таблицу результатов дисперсионного анализа и вставляем в протокол анализа или в работу.

Таблица результатов дисперсионного анализа

Источник изменчивости	*SS*	*df*	*MS*	*F*	*p*
Регрессия	1392,6229	1	1392,623	331,48403	3,82711E-11
Остатки	58,8164712	14	4,201177		
Общая	1451,43938	15			



	A	B	C	D	E	F	G	H
1	Регрессия: оценка значимости в ANOVA					?		
2								
3	X	Y	Проверка	(Y-Yсреднее)^2	Y по модели	Проверка	Остаток r	r^2
4	5,7	10,5	ЛОЖЬ	39,14066406	10,70232	10,70232	-0,20232	0,04093338
5	2,1	6,6	ЛОЖЬ	103,1494141	5,35416	5,35416	1,24584	1,55211731
6	9,8	18,3	ЛОЖЬ	2,383164063	16,79328	16,79328	1,50672	2,27020516
7	9,3	16,6	ЛОЖЬ	0,024414063	16,05048	16,05048	0,54952	0,30197223
8	17,8	26,8	ЛОЖЬ	100,8769141	28,67808	28,67808	-1,87808	3,52718449
9	3,9	6,3	ЛОЖЬ	109,3331641	8,02824	8,02824	-1,72824	2,9868135
10	3	6,9	ЛОЖЬ	97,14566406	6,6912	6,6912	0,2088	0,04359744
11	3,2	7,9	ЛОЖЬ	78,43316406	6,98832	6,98832	0,91168	0,83116042
12	4,1	6,9	ЛОЖЬ	97,14566406	8,32536	8,32536	-1,42536	2,03165113
13	11,7	14,8	ЛОЖЬ	3,826914063	19,61592	19,61592	-4,81592	23,1930854
14	17,2	28,4	ЛОЖЬ	135,5769141	27,78672	27,78672	0,61328	0,37611236
15	5,6	11,5	ЛОЖЬ	27,62816406	10,55376	10,55376	0,94624	0,89537014
16	26	39,6	ЛОЖЬ	521,8369141	40,86	40,86	-1,26	1,5876
17	9,9	19,8	ЛОЖЬ	9,264414062	16,94184	16,94184	2,85816	8,16907859
18	14,9	27,6	ЛОЖЬ	117,5869141	24,36984	24,36984	3,23016	10,4339336
19	12,2	19,6	ЛОЖЬ	8,086914063	20,35872	20,35872	-0,75872	0,57565604
20			ИСТИНА	нет данных	2,2344	нет данных	-2,2344	нет данных
21			ИСТИНА	нет данных		данных	0	нет данных
22			ИСТИНА	нет данных		нет данных	0	нет данных
23			ИСТИНА	нет данных		нет данных	0	нет данных

⑥ Из таблицы выписываем значение *F*-критерия и степени свободы для него ($df_{\text{регрессия}}$; $df_{\text{остатки}}$), а также оценку статистической значимости *p*. Округляем. Оформляем результат: $F_{(1;\,14)} = 331{,}5$; $p << 0{,}001$.

К сведению. Анализ использованной расчётной таблицы Excel — очень полезная практика. Основная сложность создания подобных таблиц заключается в том, что объём выборки заранее неизвестен, а расчёт должен выполняться корректно во всём диапазоне, для которого она создаётся. Это можно запрограммировать разными способами. В данном случае были использованы: 1) проверка ячейки на пустоту; 2) замена цифровых данных для пустых ячеек текстовой переменной «нет данных» и 3) то обстоятельство, что использованные в расчёте функции Excel не обрабатывают ячейки с текстом. Из статистических особенностей следует отметить, что в качестве суммы квадратов отклонений общей регрессии используется квадрат отклонений от среднего: $SS_{\text{общая}} = (y_i - \bar{y})^2$. Это означает, что значимость регрессии мы оцениваем относительно линии, проходящей параллельно оси *X*, на уровне среднего значения *Y*, то есть относительно равенства коэффициента линейной регрессии единице.

⑦ Оформление в квалификационной работе (вариант).

7.1. Статистическая часть раздела «Материалы и методы».

Поиск зависимости проводили в ходе линейного регрессионного анализа методом главных осей (Major axes regression). Проверку полученной модели на автокорреляцию и однородность дисперсии остатков проводили, соответственно, с помощью критериев Дарбина — Уотсона и Бройша — Пагана, а оценку статистической значимости — в ходе дисперсионного анализа. Эффекты считали статистически значимыми при $P \leq 0{,}05$, незначимыми — при $P > 0{,}10$, в промежуточных случаях ($0{,}05 < P \leq 0{,}10$) обсуждали тенденцию к различиям. Расчёты и графические построения выполнены в пакетах PAST (v. 3.19; Hammer et al., 2001) и Excel (Microsoft Office 2007).

7.2. Раздел «Результаты и обсуждение».

Поскольку данные о концентрации билирубина в капиллярной и венозной крови были получены от одних и тех же пациентов, вместо обычной линейной регрессии использовали регрессию главных осей (рис. 12.2). Полученная зависимость была высоко статистически значимой: $F_{(1;\,14)} = 331{,}5$; $p << 0{,}001$. Коэффициент детерминации R^2 составил 0,960, что указывает на близость полученной зависимости к функциональной и возможность её использования для прогноза…

Если в работе регрессионная зависимость одна и/или очень важна, имеет смысл дать результаты по ней подробнее. Например, можно привести статистику критериев Дарбина — Уотсона и Бройша — Пагана, а параметры уравнения дать с 95% ДИ. Далее обычно стоит обсудить форму зависимости и возможно-

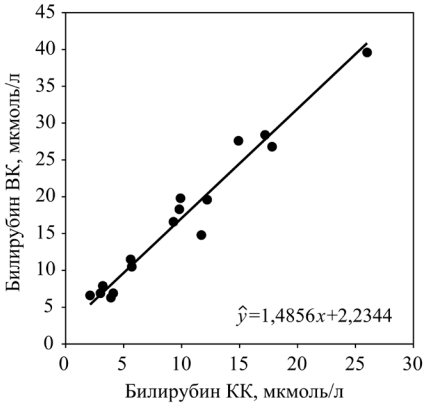


Рис. 12.2. Зависимость концентрации билирубина общего в венозной крови от его концентрации в капиллярной крови

сти её практического применения. Для нашего примера линейность не вызывает сомнений, а практическое использование заключается в возможности обоснованного прогноза содержания ОБ в венозной крови по более удобному для пациента анализу крови из пальца.

7.3. Раздел «Выводы». С использованием регрессии главных осей получена высоко статистически значимая ($F_{(1;\,14)} = 331{,}5$; $p << 0{,}001$) и надёжная ($R^2 = 0{,}960$) модель для расчёта в мкмоль/л содержания билирубина общего в венозной крови (y) по его содержанию в крови капиллярной (x): $y = 1{,}4856x + 2{,}2344$.

> **Комментарий**. На протяжении трёх лабораторных занятий мы работали с данными по содержанию общего билирубина в капиллярной (КК) и венозной (ВК) крови одних и тех же пациентов. Мы установили, что данный показатель был статистически значимо и существенно (на 41,7 %) меньше в КК, а значит, использовать кровь из пальца для целей диагностики нельзя. Тем не менее далее мы выяснили, что значения в КК и ВК сильно коррелируют, а на этом занятии получили формулу для пересчёта концентрации билирубина из данных по КК на ВК. **Вопрос**: как вы думаете, могут ли медицинские лаборатории использовать подобные, полученные по результатам собственных исследований формулы? Конечно, не могут, поскольку лаборатории должны строго придерживаться тех методик и оборудования, которые указаны в области аккредитации. Тем не менее проводить подобные исследования и публиковать их результаты необходимо для формирования информационной среды, благоприятствующей актуализации существующих протоколов медицинской диагностики.

ЛАБОРАТОРНАЯ РАБОТА № 13

Анализ зависимостей. Нелинейная регрессия

Тема 12. Анализ зависимостей. Нелинейная регрессия.
Количество часов: 2.
Цель: Познакомиться с разными вариантами нелинейного регрессионного анализа: специальные виды регрессий, логистическая регрессия, полиномиальная регрессия, сглаживание сплайнами. Работа на ПК.

Одной из важнейших практических задач является прогнозирование поведения системы (биологической, экономической, социальной), решение которой традиционно базируется на регрессионных техниках. Поскольку большинство зависимостей в природе и социуме имеют нелинейный характер, методы ***нелинейной регрессии*** (*nonlinear regression*) очень хорошо разработаны и крайне разнообразны. Чтобы ориентироваться в этом многообразии, полезно сначала отнести задачу к одной из типичных ситуаций:

1) тип нелинейной зависимости приблизительно известен;
2) тип неизвестен, но форма отчётливо видна графически и не слишком сложна;
3) форма зависимости неизвестна и сложна либо данные сильно зашумлены.

На этом занятии мы бегло познакомимся с несколькими типами нелинейных зависимостей, относящихся ко всем трём ситуациям. **ВАЖНО:** читая научные статьи в своей предметной области, обращайте внимание на используемые авторами регрессионные техники, модели и статистические пакеты.

I. Тип нелинейной зависимости приблизительно известен

В эту очень большую категорию попадают самые разные нелинейные зависимости, для которых на основе теоретического анализа явления построены модели в форме дифференциальных уравнений. Решение таких уравнений приводит к специальным зависимостям, параметры которых имеют смысл с точки зрения лежащей в их основе теории. Примеры: уравнения роста (***Берталанфи, Гомперца, аллометрическая кривая, S-образные***

логистические кривые), уравнения ферментативной кинетики (***Михаэлиса — Ментен, Хилла, Ленгмюра — Хиншельвуда*** и др.), зависимости типа ***«доза-эффект»***, зависимости с дихотомическим откликом (***бинарная логистическая регрессия, пробит-регрессия, cloglog-регрессия***). Подгонка таких зависимостей может использоваться для решения задач: 1) оценки статистической значимости зависимости; 2) оценки параметров; 3) расчётов по модели и 4) прогноза.

Раньше подгонку таких моделей и оценку параметров проводили с помощью специально подобранных *линеаризующих* конкретную зависимость преобразований, а далее рассчитывали параметры обычной линейной регрессии методом наименьших квадратов (МНК) и ретрансформировали полученные значения в исходную шкалу с помощью обратного преобразования. В настоящее время подгонка сложных нелинейных моделей осуществляется на компьютерах, обычно МНК по *итерационным алгоритмам* напрямую. В зависимости от характера распределения ошибки зависимости более корректным может быть как старый подход с линеаризацией-ретрансформацией, так и современный итерационный (см. теоретический материал). Следует быть готовым к тому, что для сложных зависимостей разные пакеты могут выдавать несколько отличающиеся результаты, что не является ошибкой пакета, а связано обычно с использованием разработчиками разных алгоритмов поиска решения. **ВАЖНО**: всегда следует указывать в разделе «Материалы и методы» используемый статпакет и его версию.

В целом в рассматриваемой ситуации решение задачи сводится лишь к отнесению зависимости к одному из известных типов и поиску программы для анализа. В статистических пакетах имеется разный и не всегда полный набор известных нелинейных зависимостей, поэтому следует искать пакеты или онлайновые ресурсы, позволяющие провести нужный анализ максимально эффективно (удобство ввода данных, подробный отчёт по анализу с параметрами, их ошибками или ДИ, статистиками качества подгонки модели, остатками модели, а также с хорошими графиками, желательно с 95%-ными доверительными границами).

П р и м е р 1. **Зависимая переменная — количественный показатель**. Переваривание пищи — сложный процесс,

протекающий при участии большого числа ферментов. Аминокислота триптофан (L-триптофан), входящая в состав белков всех известных живых организмов, образуется в результате гидролиза промежуточного продукта — карбобензоксиглицил-L-триптофана. Этот процесс протекает при участии фермента карбоксипептидазы, вырабатываемого поджелудочной железой:

$$\text{карбобензоксиглицил-L-триптофан} + H_2O \xrightarrow{\text{карбоксипептидаза}} \text{L-триптофан} + \text{карбобензоксиглицин}$$

В экспериментальных условиях (25 °C, pH = 7,5) была измерена скорость этой реакции в зависимости от концентрации субстрата. Получены данные:

Концентрация субстрата, ммоль	Скорость, моль/с
2,5	0,024
5	0,036
10	0,053
15	0,060
20	0,064

Задание. Подогнать зависимость скорости реакции ферментативного гидролиза субстрата от его концентрации уравнением Михаэлиса — Ментен. Определить и интерпретировать параметры зависимости, построить график и оценить статистическую значимость регрессии.

В пакете PAST

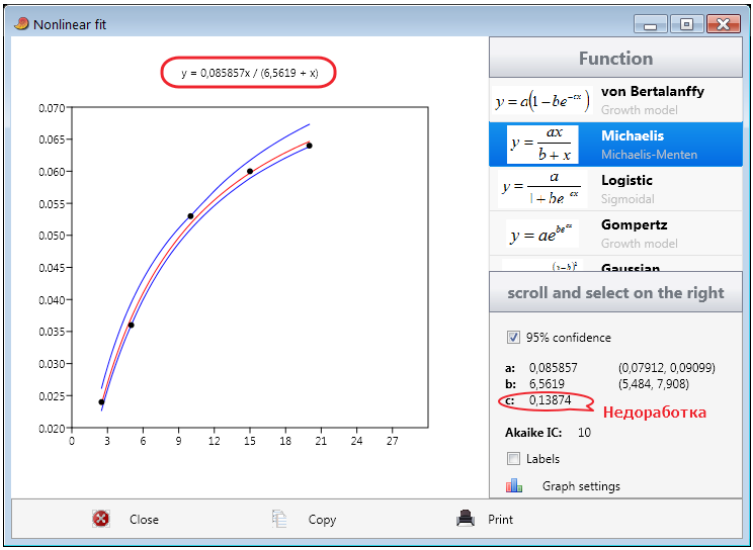
① Создать из данных файл «Триптофан.dat» и выделить область данных.

② Путь: Model — Nonlinear fit (нелинейная подгонка).

③ Вращая колёсико мыши, перемещаемся по имеющимся в пакете функциям (посмотрите все):

Function	Функция
Linear (with slope, intercept)	Линейная (с коэф. регрессии и своб. членом)
Quadratic (2nd order polynomial)	Квадратическая (полином второй степени)
Pover (Allomrtric equation)	Степенная (аллометрическое уравнение)
Exponential (Increase or decay)	Экспоненциальная (рост или распад)
von Bertalanffy (Growth model)	Берталанфи (модель роста)
Michaelis (Michaelis-Menten)	Михаэлиса — Ментен
Logistic (Sigmoidal)	Логистическая (сигмоидная, S-образная)
Gompertz (Growth model)	Гомперца (модель роста)
Gaussian (Normal distribution)	Гауссиана (нормальное распределение)

④ Возвращаемся на Michaelis (Michaelis-Menten) и выписываем параметры: *a* и *b*. Уравнение зависимости с этими параметрами изображается над графиком.



Оформляем уравнение зависимости в виде $\hat{y} = \dfrac{0{,}085857x}{6{,}5619 + x}$.

В текущей версии PAST (3.19) в данном разделе есть недоработки. Во-первых, параметр *c* отсутствует в формуле, но отображается в результатах от предыдущей функции — просто игнорируйте его. Более существенным недостатком является отсутствие какой-либо статистики, отражающей качество подгонки модели. Поэтому по результатам данного модуля мы пока не можем судить о статистической значимости регрессии и рассчитаем её самостоятельно в п. 6.

⑤ Интерпретация параметров зависимости и расчёты по параметрам. **ВАЖНО**: для каждой нелинейной регрессии параметры интерпретируются по-разному, поэтому необходимо знать теорию, лежащую в основе данной конкретной модели. Для нашего примера параметр *a* соответствует максимальной скорости реакции V_{max}, а параметр *b*, который называется константой Михаэлиса K_M, является функцией от трёх констант, определяющих скорости: 1) образования фермент-субстратного комплекса; 2) его дис-

социации; 3) его превращения в фермент и продукт. Как видно из графика, константа Михаэлиса соответствует концентрации субстрата, при которой достигается половина скорости V_{max}:

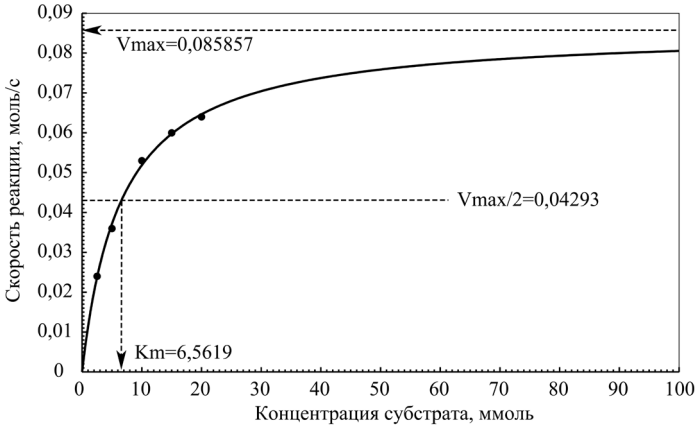


Рис. 13.1. Графическая интерпретация параметров уравнения Михаэлиса — Ментен

⑥ Расчёт статистической значимости можно провести, воспользовавшись уже знакомой по предыдущему занятию расчётной таблицей Excel «Регрессия_оценка значимости в ANOVA.xlsx».

В пакете Excel

6.1. Открыть файл «Регрессия_оценка значимости в ANOVA.xlsx» и посмотреть примечание, наведя указатель мыши на ячейку со знаком вопроса «?».

6.2. В пакете PAST копируем в буфер данные из первых двух столбцов, а в электронной таблице вставляем их в столбцы X и Y.

6.3. Становимся в ячейку E4 и в верхней строке изменяем формулу на требующуюся. Для нашей регрессии Михаэлиса — Ментен вносим

```
=(0,085857*A4)/(6,5619+A4)
```

6.4. Копируем эту ячейку протяжкой в последующие пять или более ячеек.

6.5. Копируем таблицу результатов дисперсионного анализа и вставляем в протокол анализа или в работу:

Таблица результатов дисперсионного анализа

Источник изменчивости	*SS*	*df*	*MS*	*F*	*p*
Регрессия	0,00113999	1	0,00114	1064,9967	6,32386E-05
Остатки	3,2112E-06	3	1,07E-06		
Общая	0,0011432	4			

6.6. Из таблицы выписываем значение *F*-критерия и степени свободы для него ($df_{\text{регрессия}}$; $df_{\text{остатки}}$), а также оценку статистической значимости *p.* Округляем. Оформляем результат: $F_{(1;\,3)} = 1\,065{,}0$; $p << 0{,}001$.

⑦ **Оформление в квалификационной работе (вариант)**.

7.1. Статистическая часть раздела «Материалы и методы».

Для оценки скорости реакции ферментативного гидролиза субстрата от его концентрации использовали нелинейную регрессию Михаэлиса — Ментен. Оценку статистической значимости полученной зависимости проводили в ходе дисперсионного анализа и считали её статистически значимой при $P \leq 0{,}05$. Расчёты и графические построения выполнены в пакетах PAST (v. 3.19; Hammer et al., 2001) и Excel (Microsoft Office 2007).

7.2. Раздел «Результаты и обсуждение».

Помещаем график полученной зависимости. Результаты обсуждаются в соответствии с целями исследования, которые могут заключаться как в простой констатации наличия зависимости, так и иметь какую-либо свою специфику.

7.3. Раздел «Выводы».

Установлено, что зависимость скорости реакции ферментативного гидролиза карбобензоксиглицил-L-триптофана от его концентрации была высоко статистически значима ($F_{(1;\,3)} = 1\,065{,}0$; $P << 0{,}001$) и описывалась уравнением Михаэлиса — Ментен с параметрами: $V_{max} = 0{,}086$ (95% ДИ: от 0,079 до 0,091) и $K_M = 6{,}562$ (95% ДИ: от 5,48 до 7,91).

П р и м е р 2. Зависимая переменная — качественный дихотомический показатель. Рак предстательной железы (простаты) — одно из наиболее распространённых злокачественных новообразований у мужчин: он занимает 2–3-е место среди других онкологических заболеваний в мире и является наиболее час-

той причиной смерти пожилых мужчин. Для его диагностики широко применяется простатический специфический антиген (ПСА) — опухолевый маркёр, определение которого проводится в сыворотке крови в двух формах — свободном и связанном с α-1-антихимотрипсином — которые в сумме составляют показатель «ПСА общий». Высокий уровень «ПСА общий» и низкий уровень отношения «ПСА свободный/общий» являются основанием подозревать наличие рака простаты.

Данные: у 130 мужчин, обратившихся с жалобами к урологу, были определены концентрации в ПСА сыворотке. В ходе последующих диагностических процедур у 63 пациентов диагноз «рак простаты» подтвердился, а у 67 была диагностирована доброкачественная опухоль — аденома простаты. В файле «ПСА.xls» приведены данные по концентрации ПСА свободного, связанного и общего (в нг/мл), отношению ПСА свободный/общий (в долях единицы) и наличию (1) или отсутствию (0) рака предстательной железы.

Задание. Установить зависимость наличия рака простаты от показателей ПСА. Оценить статистическую значимость зависимости и построить график бинарной логистической регрессии. Определить пороговое значение показателя для отнесения пациента к группе онкологического риска.

 **В пакете Excel** открыть файл «ПСА.xls», выделить область данных и скопировать в буфер обмена.

 **В пакете PAST**

① Подготовить поля для вставки данных с заголовком, вставить данные из буфера и сохранить как «ПСА.dat».

② Выделить кликом мыши колонки «ПСА свободный/общий» и «Рак», удерживая клавишу [Ctrl].

③ Путь: Model — Generalized Linear Model (обобщённая линейная модель).

④ В качестве распределения (Distribution) для зависимой переменной выбираем биномиальное — [Binomial], а в качестве связующей функции (Link function) — логит — [Logit].

⑤ На закладке Plot получаем график, который дорабатываем в векторном редакторе.

⑥ На закладке Statistics выписываем параметры логистической

регрессии (*logistic regression*): коэффициент регрессии *a* и свободный член (константа) *b*. Записываем уравнение регрессии двумя способами:

1) в форме логитов, которые понадобятся далее для расчётов:

$$\text{Логит}\,(y) = 1{,}3663 - 4{,}9182x$$

2) в форме функции вида $y = f(x)$, которая изображена на графике:

$$y = \frac{1}{1 + e^{-(1{,}3663 - 4{,}9182x)}}$$

⑦ Оценка статистической значимости проводится по критерию отношения правдоподобия (*G*-критерию), который имеет теоретическое распределение хи-квадрат с одной степенью свободы: $G_{(1)} = 32{,}44$; $p << 0{,}001$ (зависимость высоко статистически значима).

⑧ Расчёты по модели в силу её относительной сложности распишем подробнее.

8.1. Нахождение ***порогового значения*** (*cutoff value*) для отнесения пациента к группе риска по раку простаты.

Для бинарного исхода ($y = 0$ или $y = 1$) пороговое значение будет соответствовать вероятности $P = 0{,}5$: значение $P < 0{,}5$ мы будем расценивать как 0, значение $P > 0{,}5$ — как 1. Таким образом, нам нужно рассчитать такое значение регрессора *X*, при котором $P = 0{,}5$. Решим соответствующее уравнение, подставив в него полученные значения свободного члена (константа) и коэффициента регрессии:

$$\text{Логит}\;\;(P) = \ln\left(\frac{P}{1-P}\right) = \text{Константа} + aX$$

$$\ln\left(\frac{0{,}5}{1-0{,}5}\right) = 1{,}3663 - 4{,}9182X;$$

$0 = 1{,}3663 - 4{,}9182X$;

$4{,}9182X = 1{,}3663$;

$4{,}9182X = 1{,}3663 / 4{,}9182 = 0{,}2778049 \approx 0{,}278$.

Таким образом, пороговое значение ПСА своб./общий составляет 0,278, или 27,8 %. В нашем случае: чем меньше значение, тем

выше риск (см. график), поэтому всех пациентов со значением отношения ПСА своб./общий менее 0,278 следует причислить к группе риска по наличию рака простаты.

8.2. Нахождение вероятности отнесения конкретного пациента к группе риска.

Согласно одному из распространённых определений, ► **доказательная медицина** (*evidence based medicine*) — это добросовестное, точное и осмысленное использование лучших результатов клинических исследований для выбора лечения конкретного больного. Поэтому на основании полученной в исследовании информации важно уметь рассчитать величину риска для конкретного пациента.

Предположим, что у очередного обратившегося пациента отношение ПСА свободный/общий по результатам лабораторного исследования составило 0,480, или 48,0 %. Рассчитаем вероятность наличия у него рака простаты:

$$\boxed{\text{Логит}\,(P) = \text{Константа} + aX}$$

Логит $(P) = 1{,}3663 - 4{,}9182 \times 0{,}480 = -0{,}994436$;

$$P = \left(\frac{1}{1 + e^{-\text{Логит}(P)}}\right)$$

$$P = \left(\frac{1}{1 + e^{-(-0{,}994436)}}\right) = \left(\frac{1}{1 + e^{0{,}994436}}\right) = 0{,}2700368 \approx 0{,}270, \text{ или } 27{,}0\ \%.$$

Таким образом, для данного конкретного пациента вероятность наличия рака простаты составляет 0,270, или 27,0 %. Это меньше 0,5, или 50 %, поэтому по данному диагностическому критерию нет оснований причислить пациента к группе риска.

⑨ **Оформление в квалификационной работе (вариант)**.

9.1. Статистическая часть раздела «Материалы и методы».

Для поиска зависимости качественного дихотомического отклика от количественного показателя и прогноза риска использовалась модель бинарной логистической регрессии. Расчёты и графические построения выполнены в пакетах PAST (v. 3.18; Hammer et al., 2001) и TpX (version 1.5; https://sourceforge.net/projects/tpx/).

9.2. Раздел «Результаты и обсуждение».

В данный раздел квалификационной работы можно поместить график полученной зависимости с указанием вычисленного

порогового значения для отнесения пациента к группе риска (рис. 13.2).

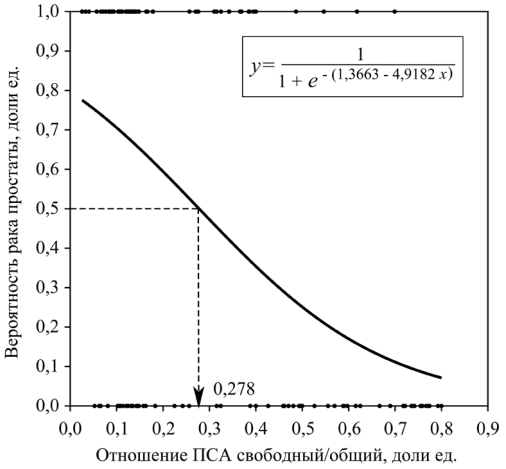


Рис. 13.2. Кривая логистической регрессии для расчёта вероятности рака простаты по значению онкомаркёра. Пунктир — пороговое значение (пояснения в тексте)

В публикациях такие графики редки по двум причинам: 1) форма зависимости для данного типа регрессии тривиальна, а то, насколько хорошо она подходит к данным, из рисунка не совсем понятно: нужно иметь опыт, чтобы оценить крутизну S-образного перехода; 2) часто в подобные модели включают не один показатель, а сразу несколько с получением модели ***множественной логистической регрессии*** (*multiple logistic regression*), которую нельзя изобразить графически.

Намного важнее графика привести сами параметры регрессии (константу и коэффициента регрессии). Если показателей или анализов много — их можно свести в таблицу. Поскольку качество диагностики для отнесения пациента к группе риска зависит от порогового значения, обсуждение строится обычно вокруг именно этого числа, которое в разных работах или условиях варьирует в каких-то пределах.

9.3. Раздел «Выводы».

С использованием логистической регрессии установлена высоко статистически значимая зависимость вероятности *y* наличия рака

предстательной железы от отношения ПСА свободный/общий (критерий отношения правдоподобия $G_{(1)} = 32{,}44$; $P << 0{,}001$), которая описывалась уравнением: $y = 1 / (1 + e^{-(1{,}3663-4{,}9182x)})$. Согласно последнему пациента следует причислить к группе риска при значении отношения ПСА свободный/общий менее 0,278.

9.4. Приложение.

В квалификационной работе полезно создать приложение «Расчёты по моделям для прогноза риска», куда поместить 1–2 примера с расчётами, аналогичными п. 8.2. Это придаст работе бесспорную практическую значимость.

II. Тип нелинейной зависимости неизвестен, но форма отчётливо видна на графике и не слишком сложна

На графике такая зависимость выглядит как *монотонное* увеличение или уменьшение показателя либо как *немонотонная* функция (есть участки с увеличением и уменьшением значений) с одним-двумя изгибами. Такие зависимости либо не изучены теоретически, либо имеют исключительно утилитарное значение. Пример: зависимость отклика прибора от концентрации вещества в широком диапазоне концентраций. Такая зависимость, в принципе, может быть смоделирована, но не особо интересна теоретически, поскольку используется лишь в качестве полезного *интерполятора* для расчёта концентраций в конкретных образцах по отклику прибора. Производителям прибора проще написать рекомендации для разбиения зависимости на небольшие линейные участки отклика, чем моделировать сложную нелинейную зависимость.

Подгонка подобных зависимостей может использоваться для решения задач: 1) оценки статистической значимости зависимости, 2) интерполяции значений на участке функции, 3) осторожного прогноза. Параметры модели не имеют самостоятельной ценности, поскольку за ними не стоит никакой теории.

В данной ситуации задача обычно решается подгонкой неизвестной монотонной зависимости МНК наиболее подходящей нелинейной функцией из числа *основных элементарных функций*:

1) ***степенная*** (*power function*): $y = x^a$;

2) ***показательная***, или ***экспоненциальная*** (*exponential function*): $y = a^x$;

3) ***обратная*** (*reciprocal function*): $y = 1/x$.

Также может использоваться ***полиномиальная функция*** (*polynomial function*): $y = ax^n + bx^{(n-1)} + \ldots nx + k$. Для немонотонных кривых с одним изгибом используют полином второй степени (ветвь параболы: $y = ax^2 + bx + c$), с двумя — третьей степени, с тремя — четвёртой. Полиномиальная регрессия более высоких порядков обычно не используется. Для немонотонных периодических зависимостей в простейшем случае применяются ***синусоиды*** (*sinusoidal function*).

В пакете PAST часть из этих зависимостей можно найти в уже рассмотренном модуле по пути Model — Nonlinear fit (нелинейная подгонка). Полиномиальную регрессию можно провести в модуле по пути: Model — Polynomial fit. Представить сложную нелинейную зависимость в виде накладывающихся друг на друга синусоид можно по пути: Model — Sum-of-sinusoids.

В этих модулях о качестве подгонки данных моделью можно судить по трём характеристикам: *p*-значению (чем меньше, тем лучше), коэффициенту детерминации R^2 (отображается программой как R^2; чем ближе к 1, тем лучше) и ***информационному критерию Акаике*** *AIC* (отображается программой как Akaike *IC*, чем меньше, тем модель лучше). В отличие от *p* и R^2, которые всегда улучшаются с увеличением сложности модели, критерий Акаике использует штраф за увеличение числа параметров модели. Этот штраф позволяет выбрать достаточно общую и, вероятно, более подходящую модель к подобным зависимостям вообще, а не к конкретному частному набору данных.

При необходимости оценки качества модели в ходе дисперсионного анализа можно воспользоваться уравнением полученной зависимости, значениями остатков модели (Residuals) и файлом «Регрессия_оценка значимости в ANOVA.xlsx» аналогично тому, как это было сделано в п. 6 примера с регрессией Михаэлиса — Ментен.

III. Форма зависимости неизвестна и сложна либо данные сильно зашумлены

На графике такая зависимость выглядит как сложная функция с большим числом изгибов либо в облаке точек зависимость плохо видна в силу большой изменчивости показателей. Примеры: возрастная динамика биохимического показателя, изменение

общего микробного числа воды на участке реки. Такие зависимости пока неинтересны для теоретического обобщения или не разрабатываются теоретически в силу сложности процессов и большого числа влияющих на отклик факторов. Задача в этом случае сводится не столько к оценке статистической значимости, сколько, во-первых, к поиску и/или визуализации наиболее общих закономерностей, которые далее позволят выдвинуть какие-либо гипотезы о возможных причинах наблюдаемого нелинейного поведения системы, и, во-вторых, к очень осторожному прогнозу, скорее качественному, чем количественному.

Визуализация подобных сложных зависимостей обычно проводится методами, отличными от метода наименьших квадратов (МНК). Это может быть простое сглаживание методом ***скользящих средних***, ***сглаживание сплайнами***, ***локально взвешенные регрессии*** (*Local regressions, Locally scatterplot smoothing — LOESS, Locally weighted scatterplot smoothing — LOWESS*), ***обобщённые аддитивные модели*** (*Generalized Additive Model, GAM*). При этом речь обычно идёт не о подгонке модели, а о ***сглаживании*** *(smoothing)* или ***фильтрации*** (*filtering data*) данных.

Пример 3. Рассмотрим фильтрацию данных на примере с ПСА. Теоретически понятно, что свободный и связанный ПСА могут коррелировать, и нам интересно изобразить эту связь линией. Установим самый общий характер этой связи двумя способами.

В пакете PAST

① Открыть файл «ПСА.dat» и выделить столбцы с ПСАсвоб. и ПСАсвяз.

② Путь: Model — Smoothing spline (сглаживание сплайнами).

③ Мы видим, что данные расположены на рисунке крайне неоднородно, что указывает на асимметрию их распределения. Убедитесь в этом, посмотрев гистограммы распределения. Прологарифмируйте обе колонки значений (Transform — Log) и снова выполните сглаживание: Model — Smoothing spline. Логарифмы концентраций выглядят более однородными и естественными; продолжим работу с ними.

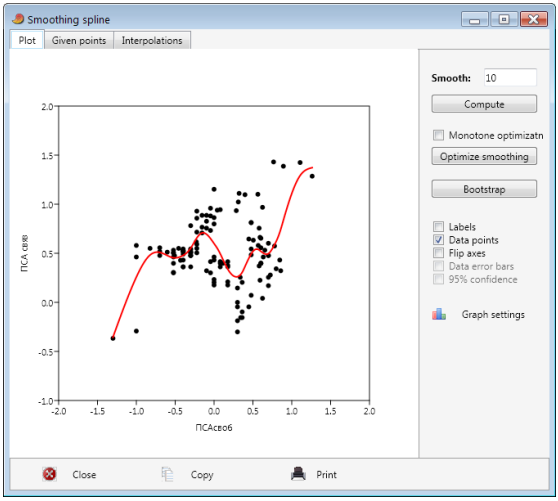
④ Введите в поле Smooth число 1 и нажмите [Compute]. Увеличивайте степень сглаживания, увеличивая число в окне Smooth

на 1 и нажимая [Compute]. Дойдите до числа 5, затем введите 11, затем 15.

> **Комментарий**. ► **Сплайн** (*spline*) — функция, область определения которой разбита на несколько отрезков, каждый из которых подгоняется неким полиномом (например, квадратным или кубическим). Сглаживание сплайнами можно представить следующей механической моделью: мы продеваем через точки с данными полоску из упругого материала так, чтобы она оказалась как можно ближе к данным, в идеале — прошла бы через все точки. Увеличивая жёсткость материала полоски, мы будем получать всё более близкую к линии форму и всё более общую форму интерполируемой сплайном зависимости.

На какой степени сглаживания нужно остановиться, чтобы получить оптимальную по сложности и степени обобщения зависимость, решает исследователь исходя из знаний в предметной области. Также можно воспользоваться кнопкой [Optimize smoothing], чтобы получить «оптимальное» значение, выдаваемое по результатам ***перекрёстной проверки*** (кроссвалидации, *cross-validation*). Про эту современную ресэмплинг-технику см. теоретический материал.

⑤ Нажмите [Optimize smoothing] и получите итоговое изображение. Судя по нему, исследуемая зависимость — прямая: чем больше один показатель, тем больше другой. Однако вряд ли она имеет столь сложный характер; возможно, на этом графике совмещено две или более зависимостей на разных уровнях. Закройте изображение.

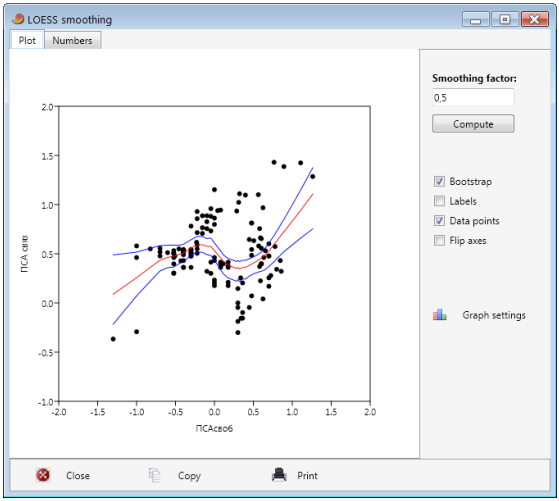


⑥ Файл «ПСА.dat» открыт, столбцы с логарифмами ПСАсвоб. и ПСАсвяз. выделены. Путь: Model — LOESS smoothing (локально взвешенное сглаживание).

⑦ Увеличивайте на 0,1 ***параметр сглаживания*** в окне Smoothing factor и нажимайте [Compute]. Дойдите до 0,6. Введите число 1, затем 0,5. Вы видели, что при значении 1 зависимость практически исчезает (линия регрессии параллельна оси *x*), что противоречит наблюдаемой нами картине. Вероятно, наиболее близкое к оптимальному сглаживанию значение — около 0,4–0,6.

> **Комментарий**. Сглаживание локальной регрессией проводится по следующему алгоритму. Для имеющегося числа точек *n* и выбранного пользователем параметра сглаживания *q* программа подгоняет *nq* точек в окрестностях каждой точки прямой линией со взвешиванием: чем дальше от неё точки, тем меньше их вес.

⑧ Поставьте галочку в ☑ Bootstrap — получим 95%-ные доверительные границы для регрессии, вычисленные методом бутстрепа. Как видно из рисунка, они содержат намного меньше 95 % имеющихся значений, что указывает на иной характер рассматриваемой зависимости:



Тем не менее мы познакомились с двумя техниками, которые могут оказаться полезными в других ситуациях. Выбор примера с ПСА был продиктован исключительно экономией данных для обучения.

ЛАБОРАТОРНАЯ РАБОТА № 14

Специфические задачи в биологических исследованиях на примере анализа выживаемости и оценки диагностической эффективности тест-системы

Тема 13. Некоторые специфические задачи в биологических исследованиях.
Количество часов: 2.
Цель: Познакомиться с анализом цензурированных данных на примере анализа выживаемости. Познакомиться с анализом чувствительности и специфичности диагностических тестов и принципом построения ROC-кривых. Работа на ПК.

I. Понятие об анализе выживаемости

► **Анализ выживаемости** (*survival analysis*) — собирательное название для группы статистических методов анализа данных о состоянии объектов во времени при наличии цензурированных наблюдений. Эти методы фокусируются на том, как долго объекты исследования сохраняются («выживают») в данном состоянии до периода отказа (смерть, рецидив, поломка и т. д.). Традиционно анализ выживаемости (АВ) использовался в демографии — для изучения ожидаемой продолжительности жизни — и в медицине — для изучения продолжительности заболеваний и последствий лечения, в том числе при учёте влияния различных сопутствующих факторов (*ковариат*). В настоящее время идеи и техники АВ используются также в технометрии — для анализа срока службы технических изделий, в социологии — для анализа трудовой занятости и др. Главной особенностью данных в АВ является наличие ***цензурированных наблюдений*** (*censored observations*) — наблюдений, выполненных не до окончания периода исследования явления, а только до какого-то срока, причём этот срок для разных объектов может быть различным. Например, известно, что пациент прожил после операции пять лет, но потом выбыл из поля зрения врачей, поскольку сменил место жительства и/или клинику и его текущий статус неизвестен (1). Либо он погиб в результате несчастного случая (2). Либо после операции прошло пять лет и пациент до сих пор жив (3), а реше-

ние о переходе на новый метод лечения необходимо принимать уже сейчас. Во всех трёх случаях имеет место ***цензурирование*** (*censoring*), то есть мы не можем приписать данному пациенту в качестве продолжительности жизни после операции пять лет, поскольку: 1) он, возможно, жив до сих пор; 2) возможно, был бы жив, не случись несчастного случая; 3) определённо жив. Однако мы должны каким-то образом учесть информацию о том, что после операции он прожил *по меньшей мере* пять лет. Строго говоря, такие неполные данные представляют собой только *цензурированные справа* данные, или *данные типа «более чем». Цензурированные слева* данные типа «менее чем» появляются вследствие ограниченной разрешающей способности аналитических методов, когда показатели у части выборки находятся ниже предела обнаружения (см. теоретический материал). Анализ таких данных мы в ходе курса не рассматриваем, но отметим, что один из способов работы с ними — работа методами AB, развёрнутыми в обратную сторону.

Обычные показатели описательной статистики, такие как доли (излеченных или умерших субъектов) и простые средние значения (время дожития), плохо подходят для анализа данных о выживаемости; та же проблема имеется и в отношении задачи сравнения выборок. Поэтому были разработаны специальные методы AB, которые включают в себя анализ таблиц смертности, множительную ***оценку Каплана — Мейера*** (*Kaplan-Meier estimator*), ***регрессии Кокса*** (*Cox regression*) — ***модель пропорциональных рисков*** или ***модель с зависящими от времени ковариатами***, ***логранговый критерий*** (*logrank test*) и др.

П р и м е р. Используем известный набор данных «Leukemia dataset II» (Freireich et al., 1963), входящий во многие статистические пакеты. Меркаптопурин — антиметаболит, относящийся к группе аналогов пурина. Поступая в организм и активируясь в печени, он ингибирует синтез ДНК. Такое цитостатическое действие приводит к противоопухолевому эффекту, а потому данный препарат используется в химиотерапии, особенно при лечении различных лейкозов. 42 ребёнка, пролеченные преднизоном от острого лейкоза и находящиеся в стадии ремиссии, были разделены на две равные группы. Первая группа (группа сравнения) получала плацебо, вторая (основная

группа) — химиотерапию 6-меркаптопурином. Исследование продолжалось 35 недель. Регистрировался рецидив лейкоза (код состояния — 1) либо его отсутствие (код состояния — 0).

Данные:

Плацебо		6-меркаптопурин	
Время, нед.	Состояние	Время, нед.	Состояние
1	1	6	0
1	1	6	1
2	1	6	1
2	1	6	1
3	1	7	1
4	1	9	0
4	1	10	0
5	1	10	1
5	1	11	0
8	1	13	1
8	1	16	1
8	1	17	0
8	1	19	0
11	1	20	0
11	1	22	1
12	1	23	1
12	1	25	0
15	1	32	0
17	1	32	0
22	1	34	0
23	1	35	0

Задача. Используя методы анализа выживаемости, оценить эффективность поддерживающей терапии с помощью меркаптопурина, оценить статистическую значимость различий в продолжительности жизни пациентов, принимавших плацебо и меркаптопурин.

> **Комментарий**. Как видно из представленных данных, в группе сравнения у всех пациентов произошёл рецидив заболевания: код состояния только 1. В группе получавших цитостатик у части пациентов рецидива не произо-

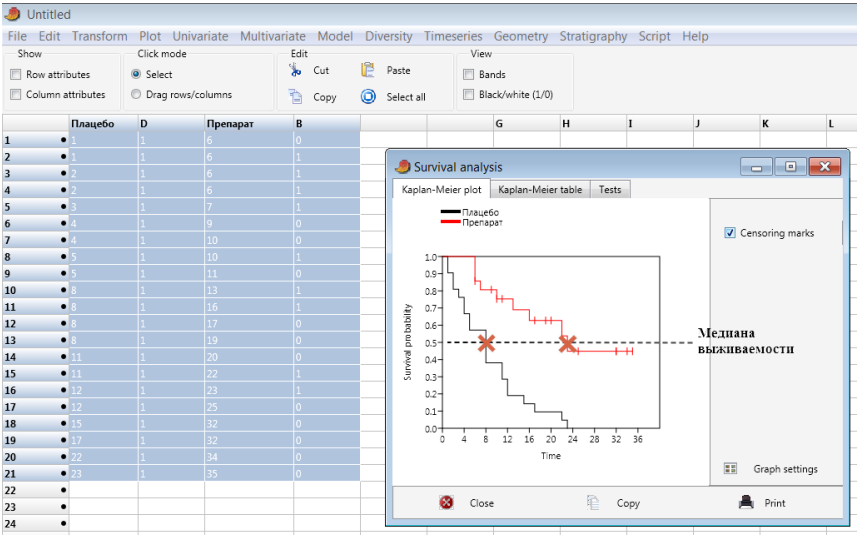
шло. Если бы все пациенты были прослежены в течение всех 35 недель, то мы имели бы полные данные, которые можно было свести в таблицу 2×2 и обработать критериями типа хи-квадрат с входами: «Группа» (плацебо или препарат) и «Рецидив лейкоза за 35 недель» (есть или нет). Однако мы видим, что состояние пациентов было прослежено в течение разного времени: кто-то наблюдался всего 6 недель, а кто-то все 35. Поэтому мы не можем просто объединить всех пациентов с кодом состояния 0 в одну группу. Возможно, что у пациента № 1 из основной группы также возник рецидив, например на 7 неделе, однако мы этого не знаем, нам известно только то, что в течение по меньшей мере 6 недель рецидива не было. Таким образом, мы имеем неполные данные, и код состояния «0» маркирует именно цензурированное наблюдение.

В пакете PAST

① Внести данные в четыре столбца: 1) время для первой группы; 2) код состояния для первой группы; 3) время для второй группы; 4) код состояния для второй группы. Дать названия колонкам. Сохранить файл. Выделить область данных.

② Путь: Univariate — Survival analysis. На закладке Kaplan – Meier plot ставим галочку в ☑ Censoring marks для отображения цензурированных наблюдений.

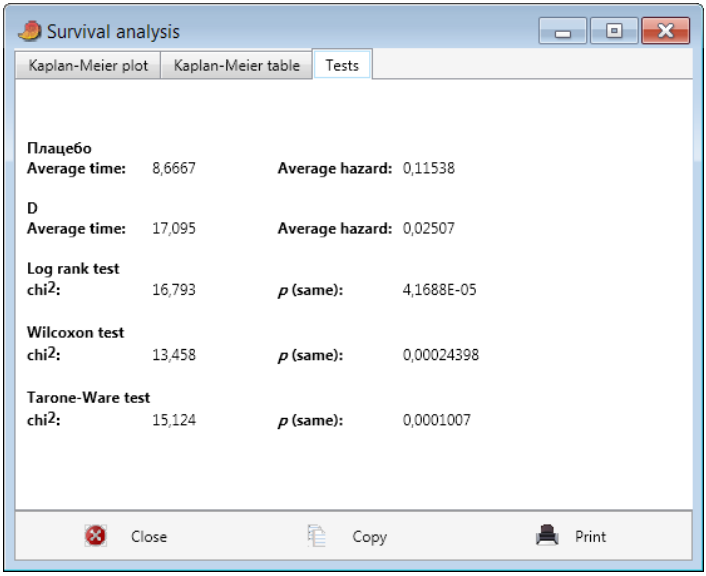
③ Полученный график называется ***кривой выживания*** (*survival plot*). В подавляющем большинстве случаев их получают именно методом Каплана — Мейера, однако в более углублённых



вариантах AB их можно приблизить (смоделировать) определёнными регрессионными зависимостями. По оси *X* здесь отложено время, а по оси *Y* — вероятность, или ***функция выживаемости***. График можно доработать в Graph settings и вставлять в отчёт.

Если провести линию через вероятность 0,5, то она пересечёт кривые выживаемости в их медианном значении: до этого срока дожила половина объектов. К сожалению, текущая версия пакета PAST (3.19) не выдаёт привычный для медиков показатель — ***медиану выживаемости***, а также квартили выживаемости.

④ Переходим на закладку Tests и смотрим, какую статистику пакет выдаёт.



Average time — ***среднее время отказа***, рассчитанное с учётом цензурирования. В нашем случае отказом была не смерть пациента, а рецидив лейкоза. Мы видим, что в группе «Плацебо» среднее время наступления рецидива составило 8,7 недели, а в группе «Препарат» (пакет неверно указывает её название как D) — 17,1 недели.

Average hazard — ***среднее значение интенсивности отказов***. Оно рассчитывается как отношение числа отказов к сумме отрезков времени до наступления отказа или цензурирования.

Критерии:

Logrank test (*Mantel-Cox test*) — ***логранговый критерий***, или логарифмический ранговый критерий, или критерий Мантела — Кокса. Непараметрический критерий, используемый для сравнения двух кривых выживаемости. Данный критерий является наиболее мощным в случае, если выживаемость подчиняется *экспоненциальному распределению* или *распределению Вейбулла*, а также в ситуациях со случайным, но равным в группах цензурированием.

Wilcoxon test (*Gehan-Breslow-Wilcoxon test*) — ***критерий Гехана — Бреслоу — Уилкоксона*** (варианты написания: критерий Гехана, критерий Гехана — Бреслоу, критерий Гехана — Уилкоксона). Является обобщением критерия Уилкоксона на случай цензурированных данных. Он является более мощным в случае, когда выживаемость подчиняется логарифмически нормальному распределению, но может сильно терять в мощности в случае большой доли цензурированных наблюдений.

Tarone-Ware tests — ***критерий Тарона — Вэра***. Проявляет высокую мощность в большинстве ситуаций, но необязательно самую высокую в различных частных случаях.

Все три критерия очень близки в плане вычислительной техники, и все аппроксимируются распределением хи-квадрат с одной степенью свободы. Различия касаются весов, приписываемых отказам на разных сроках. Критерий Гехана придаёт больший вес отказам на ранних сроках, логранговый критерий придаёт всем отказам равный вес, критерий Тарона — Вэра занимает промежуточное положение.

В нашем случае одна группа не содержала цензурированных наблюдений, поэтому логранговый критерий лучше не использовать, а из двух его модификаций в условиях неизвестного распределения выживаемости логично отдать предпочтение критерию Тарона — Вэра. Выписываем его значение и соответствующее значение p: $\chi^2_{(1)} = 15{,}12$; $p \ll 0{,}001$. Различия высоко статистически значимы.

⑤ **Оформление в квалификационной работе (вариант)**.

5.1. Раздел «Материал и методы».

Оценку различий в выживаемости пациентов двух групп проводили с использованием техники Каплана — Мейера и критерия Тарона — Вэра. Расчёты и графические построения выполнены в пакете PAST (v. 3.19; Hammer et al., 2001).

5.2. Раздел «Результаты и обсуждение».

Приводятся таблицы с показателями выживаемости в группах. В нашем случае (пакет PAST, v. 3.19) это только среднее время наступления рецидива заболевания. Также приводится график (рис. 14.1). Всё обсуждается с привлечением литературных данных.

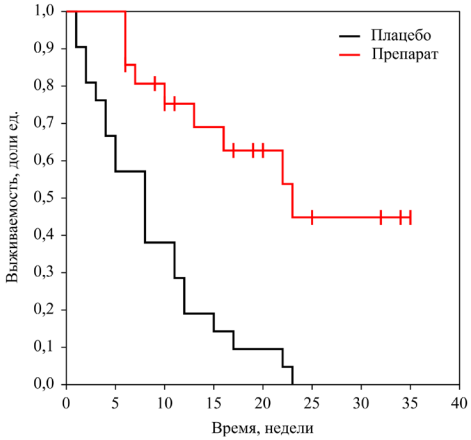


Рис. 14.1. Кривые выживаемости Капалана — Мейера для групп пациентов, получавших плацебо либо поддерживающую терапию 6-меркаптопурином

5.3. Раздел «Выводы». Поддерживающая терапия препаратом 6-меркаптопурин высоко статистически значимо увеличивала продолжительность периода ремиссии: критерий Тарона — Вэра $\chi^2_{(1)} = 15{,}12$; $P \ll 0{,}001$. В группе пациентов, получавших плацебо, среднее время наступления рецидива острого лейкоза составило 8,7 недели, тогда как в группе получавших препарат, — 17,1 недели.

II. Анализ чувствительности и специфичности диагностических тестов, ROC-анализ

В медицине и некоторых других областях науки большое значение имеет правильная диагностика наступления качественных альтернативных событий, например: имеется заболевание или нет, является вещество канцерогеном или нет. При этом в качестве исходных данных используются различные количествен-

ные, порядковые и качественные показатели, полученные в ходе каких-либо диагностических процедур. В ходе лабораторной работы № 13 мы уже узнали один из способов моделирования дихотомического отклика — с помощью логистической регрессии. На этом занятии мы познакомимся с принципиально другим подходом к диагностике и оценке её качества — через анализ чувствительности и специфичности.

► **Чувствительность** (*sensitivity*) — доля объектов из общего числа носителей признака, верно классифицированных как несущих признак (*true positive rate, TPR*). В медицине это доля верно классифицированных больных с данным диагнозом. ► **Специфичность** (*specificity*) — доля объектов из общего числа неносителей признака, верно классифицированных как не имеющих признака (*true negative rate, TNR*). В медицине это доля верно классифицированных здоровых или лиц с другим диагнозом. В результате ошибок диагностики часть объектов неверно классифицируются как имеющие признак, когда его в действительности нет (*false positive rate, FPR*), или как не имеющие его, когда признак в действительности есть (*false negative rate, FNR*). Мы говорим «в действительности», поскольку подразумеваем, что имеется некая практически безошибочная процедура диагностики, которая в медицине традиционно называлась ***«золотым стандартом»*** (в настоящее время рекомендуется использовать термин ***«критериальный стандарт»*** (*criterion standard*)), однако её использование слишком длительно, дорогостояще или сопряжено с другими трудностями. Например, для однозначной диагностики рака простаты потребуется биопсия материала опухоли и цитологическое исследование, а для констатации канцерогенного действия химиката потребуется длительное хроническое воздействие на лабораторных животных с последующим их полным патоморфологическим исследованием.

> **К сведению**. Добиться 100%-ной чувствительности достаточно просто: например, если мы будем подозревать онкозаболевание у всех пациентов, то не пропустим ни одного больного. Однако при этом мы получим нулевую специфичность. Аналогично можно легко добиться 100%-ной специфичности: достаточно считать всех пациентов здоровыми. Но при этом пострадает чувствительность, поскольку мы не выявим ни одного больного. Таким образом, идеальный диагностический тест должен обладать 100%-ными и чувствительностью, и специфичностью: чтобы все больные были верно распознаны как больные, а все здоровые — как здоровые.

На практике ситуация идеальной диагностики редка, поэтому пользуются такой мерой качества диагностики, как ***диагностическая эффективность*** (в узком смысле), которая рассчитывается как среднее между чувствительностью и специфичностью:

$$\text{Диагностическая эффективность} = \frac{\text{Чувствительность} + \text{Специфичность}}{2}.$$

П р и м е р. Рассмотрим данные примера из лабораторной работы № 13 по диагностике рака простаты у 130 мужчин по показателю отношения ПСА свободный/общий. Напомним, что с помощью модели бинарной логистической регрессии нами было получено пороговое значение 0,278, меньше которого следовало относить пациента к онкобольным, а выше которого — к не имеющим рака простаты.

Задание. Рассчитать показатели чувствительности, специфичности и диагностической эффективности показателя «Отношение ПСА свободный /общий» для диагностики рака простаты при пороговом значении 0,278.

В пакете Excel

① Открыть файл «ПСА.xls».

② Путь: Данные — Сортировка и фильтр — Сортировка. Сортировать по [Рак] — Добавить уровень — Затем по [ПСА своб./общ.].

③ Заходим в фильтр столбца [Рак ▼] и оставляем галочку только в ☑ 0. В результате отобразятся только значения пациентов без рака.

④ Заходим в фильтр столбца [ПСА своб./общ. ▼] — Числовые фильтры — Меньше или равно … вбиваем значение 0,278. Подсчитываем число отображаемых ячеек, выделив их мышью. Результат отображается в нижней части экрана в статистике выделенного фрагмента как «Количество». Количество: 26. Меняем фильтр на «Больше 0,278» и подсчитываем ячейки: Количество: 41.

⑤ Заходим в фильтр столбца [Рак ▼] и оставляем галочку в ☑ 1. Повторяем подсчёт в фильтре столбца [ПСА своб./общ. ▼]. Меньше или равно 0,278, Количество: 47, Больше 0,278 Количество: 16.

⑥ Сводим полученные данные в таблицу и проверяем краевые суммы и общую сумму на предмет возможной ошибки: 63 + 67 = 130 (верно).

Группа	ПСА своб./общ.		Всего
	≤ 0,278	> 0,278	
С раком (1)	**47**	16	63
Без рака (0)	26	**41**	67

⑦ В нашем случае меньшее значение соответствует большему риску, поэтому для (1) выделим ячейку 47, а для (0) — ячейку 41.

⑧ Рассчитываем показатели чувствительности, специфичности и диагностической эффективности:

чувствительность = 47 / 63 = 0,746, или 74,6 %;

специфичность = 41 / 67 = 0,612, или 61,2 %;

диагностическая эффективность = (0,746 + 0,612) : 2 = 0,679, или 67,9 %.

⑨ Поскольку и чувствительность, и специфичность — это частоты, рассчитываем для них 95% ДИ, как на лабораторном занятии № 3. **Задание**: самостоятельно рассчитайте 95 % ДИ по Джеффрису.

⑩ **Вывод**. При использовании для диагностики рака простаты порогового значения отношения «ПСА свободный/общий», равного 0,278, чувствительность и специфичность [95% ДИ] составили соответственно 74,6 % [62,9; 84,1] и 61,2 % [49,3; 72,2]. Диагностическая эффективность теста — 67,9 %.

* * *

Следует признать, что полученные оценки диагностической эффективности являются невысокими. **ВАЖНО!** При этом нет никаких оснований ожидать, что модель бинарной логистической регрессии дала нам оптимальное значение по показателям чувствительности и специфичности, поскольку её подгонка осуществлялась с использованием другого критерия — минимизации ошибки логита. Однако с использованием компьютера искать другие оптимальные значения можно и без модели — простым перебором всех вариантов. То есть мы можем последовательно выбирать из данных все значения показателя «ПСА свободный/общий» и, используя их в качестве порогового значения

(как только что сделали для значения 0,278), рассчитывать показатели чувствительности и специфичности. Такой анализ называется ***ROC-анализом***, от английского *Receiver Operating Characteristic* — рабочая характеристика приёмника. Этот термин пришёл из области военной радиолокации, когда по характеристике радиосигнала необходимо было принять решение о том, является ли его источником вражеская цель или какой-то другой объект. В ходе ROC-анализа рассчитанные при всех значениях маркёра пары значений чувствительности и специфичности откладываются на графике в координатах «(1-Специфичность)» и «Чувствительность» (рис. 14.2).

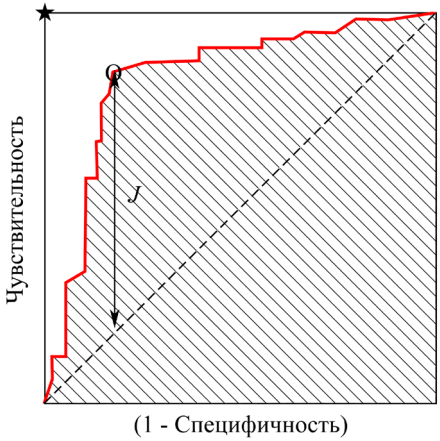


Рис. 14.2. ROC-кривая, площадь под ней (штриховка) и индекс Юдена J с точкой O максимальной диагностической эффективности

В результате получается так называемая ***характеристическая кривая***, или ***ROC-кривая*** (*ROC curve*). В таком виде максимальной чувствительности (1 или 100 %) и максимальной специфичности (1 или 100 %) соответствует точка в верхнем левом углу — на рис. 14.2 отмечена звёздочкой. Поэтому, чем ближе кривая подходит к этой точке, тем выше диагностическая эффективность метода.

В качестве количественной характеристики качества диагностики в ROC-анализе выступает ***площадь под ROC-кривой*** (*Area under curve, AUC*). На рис. 14.2 она обозначена штриховкой. Максимально возможная площадь будет 1 × 1 = 1, то есть

чем ближе AUC к 1, тем выше эффективность теста. На графиках обычно отчерчивают также диагональ, соответствующую *AUC* = 0,5. Чем ближе ROC-кривая к диагонали, тем меньшей диагностической эффективностью обладает тест.

Преимущество ROC-анализа заключается в том, что мы можем оценивать и сравнивать диагностическую эффективность разных методов не для одного порогового значения, а во всём диапазоне значений маркёра. При этом можно найти на кривой такую точку, в которой достигаются максимальные чувствительность и специфичность, и посмотреть, какому значению маркёра соответствовала эта пара. На рис. 14.2 такая оптимальная точка обозначена буквой *O*, а двусторонней стрелкой обозначено расстояние, известное как ***индекс Юдена*** *J* (*Youden's index*), который изменяется от 0 до 1:

$$J = \text{Чувствительность} + \text{Специфичность} - 1$$

Поскольку ROC-анализ является весьма узкоспециализированным методом, он имеется только в специальных статпакетах. Одним из наиболее удобных коммерческих приложений является пакет MedCalc, а из бесплатных пакетов — пакеты для статистической среды R, например: pROC и OptimalCutpoints. Мы проведём расчёт в бесплатной онлайновой программе easyROC, которая не уступает лучшим коммерческим решениям.

П р и м е р. Продолжаем работать с данными по диагностике рака простаты по показателю «ПСА свободный/общий».

Задание. Провести ROC-анализ. Определить площадь под ROC-кривой, найти оптимальное пороговое значение по критерию Юдена и рассчитать для него показатели чувствительности и специфичности. Построить графики.

В пакете Excel

① Откройте файл ПСА.xlx. Измените русские названия на английские: PCA1, PCA2, Cancer и сохраните его в текстовом формате (тип файла: Текст MS-DOS) как PCA.txt в английской раскладке клавиатуры.

② Откройте файл PCA.txt в Блокноте Windows и убедитесь, что он нормально открывается и читается. Часто на этом шаге требуется заменить десятичный разделитель в виде запятой

на точку: Правка — Заменить — Вбить в поля «Что» и «Чем» запятую и точку — Заменить всё. Но в нашем случае этого можно не делать.

В браузере

③ В строке браузера введите: http://www.biosoft.hacettepe.edu.tr/easyROC или найдите поисковиком онлайновую программу easyROC.

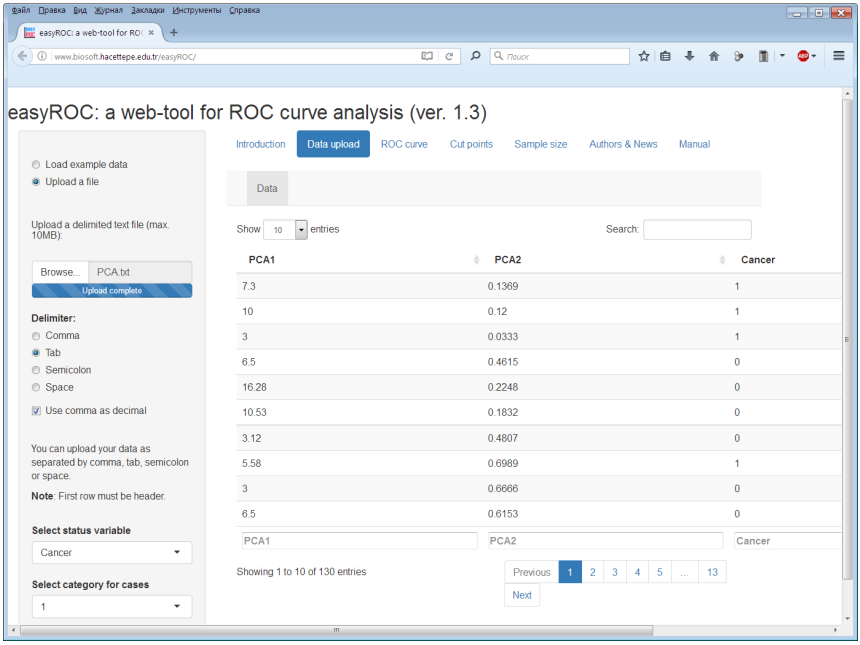
④ Ввод данных. **Раздел [Data upload]** (Ввод данных).

4.1. Радиометка в ◉ Upload a file.

4.2. Browse — Указать путь к файлу PCA.txt. В центральной части окна появятся первые 10 строк файла — значит, данные успешно считаны. Если этого не произойдёт, возможно, нужно поменять Delimiter (Разделитель значений) или в самом файле остался текст на кириллице.

4.3. Поставить галочку в ☑ Use comma as decimal, так как в качестве десятичного разделителя мы оставили запятую.

4.4. Ниже выбираем в качестве переменной статуса Cancer, а в качестве метки онкобольного — 1.



⑤ Настройка анализа. **Раздел [ROC curve]** (ROC-кривая).

5.1. Select markers (выбор маркёров) — PCA2 (так мы обозначили отношение ПСА свободный/общий).

5.2. По умолчанию программа ставит галочку в ☑ Higher values indicate risks, что обозначает, что большему риску соответствует большее значение показателя. В нашем случае это не так, что можно увидеть по вогнутой в обратную сторону характеристической кривой на рисунке. Поэтому снимаем галочку.

5.3. ☑ Advanced options (расширенные опции).

По умолчанию программа проводит ROC-анализ по Делонгу, что обеспечивает сопоставимость результатов с большинством других программ. В принципе, можно ничего не менять и сразу выписать результаты. Однако easyROC позволяет провести более тонкую настройку.

5.3.1. Select a method for curve fitting (Выбор метода для подгонки кривой). Оставляем непараметрический (◉ Nonparametric). Параметрический способ обеспечивает подгонку модели ***бинормальной кривой***, расчёт параметров для которой проводится

в предположении, что и в группе больных, и в группе здоровых распределение показателя-маркёра нормальное. Это очень сомнительное предположение, которое к тому же не даёт явных преимуществ для трактовки и практического использования результатов анализа. **Задание**: попробуйте параметрический способ, а затем верните непараметрический.

5.3.2. Select a method for SE estimation (Выбор метода для расчёта стандартной ошибки. Можно выбрать биномиальный способ (◉ Binomial), хотя это не особенно принципиально.

5.3.3. Select a method for Conf. Interval (Выбор метода для расчёта доверительного интервала). Поставьте радиометку в ◉ Binomial Exact.

⑥ Выписываем и оформляем результаты расчёта:

Площадь под ROC-кривой *AUC* = 0,767±0,0371 (95% ДИ: от 0,685 до 0,867).

Значимость отличия *AUC* от 0,5: *z*-критерий = 7,21; $p << 0{,}001$.

⑦ Поскольку графиков будет несколько, их настройку лучше проводить не в разделе **[ROC curve]**, а в следующем разделе — **[Cut points]**. Переходим в этот раздел.

Cut point (*cut-off value*) — ***точка отсечения*** (пороговое значение) соответствует такому значению маркёра, при использовании которого в качестве граничного для разделения групп на не имеющих статусного показателя (0) и имеющих его (1) достигается максимальная диагностическая эффективность. Мы уже имели дело с такой величиной в лабораторной работе № 13, где для этих же данных рассчитывали оптимальное пороговое значение по результатам логистической регрессии.

В ситуациях, когда и чувствительность, и специфичность принимаются одинаково важными для диагностики, в качестве статистики для поиска оптимальной точки отсечения обычно используется индекс Юдена. Однако следует отметить, что придание равных весов чувствительности и специфичности — это не обязательно лучший вариант. В случае заболевания, верная диагностика ранних стадий которого может спасти жизнь человеку, с этической точки зрения правильнее будет пожертвовать специфичностью в пользу чувствительности, то есть лучше несколько чаще подозревать заболевание у здоровых, зато пропустить меньше больных. Однако это может быть расточительным с экономической точки зрения, поскольку — как ни цинично это звучит — человеческая

жизнь имеет свою стоимость (в России — около 4 млн долл.). Это приводит к более сложным критериям оптимальности, которых в пакете easyROC больше, чем в коммерческих специализированных программах: 34(!) вместе с индексом Юдена. Так, например, можно учесть в расчётах цены ложных положительных и отрицательных результатов (*cost of a false positive*, *cost of a false negative*) или учесть ***превaленс*** — показатель распространённости заболевания в популяции.

Копируем и вставляем в протокол анализа данные по диагностической эффективности при использовании индекса Юдена:

method for optimal cut-off

Table 1. Cut-off Results

Optimal cut-off method:	Youden
Optimal cut-off point:	0.4
Optimal criterion:	0.4738214

Table 2. Performance Measures

Value	Lower	Limit	Upper Limit
Sensitivity:	0.937	0.845	0.982
Specificity:	0.537	0.411	0.660
Positive Predictive Value:	0.656	0.534	0.878
Negative Predictive Value:	0.900	0.769	0.938
Positive Likelihood Ratio:	2.024	1.551	2.641
Negative Likelihood Ratio:	0.118	0.045	0.313

Из этой таблицы видно, что в качестве критерия оптимальности используется критерий Юдена. Его максимальное значение $J = 0{,}474$ достигается при значении маркёра «ПСА свободный/общий», равном 0,4. При использовании данного значения в качестве точки отсечения:

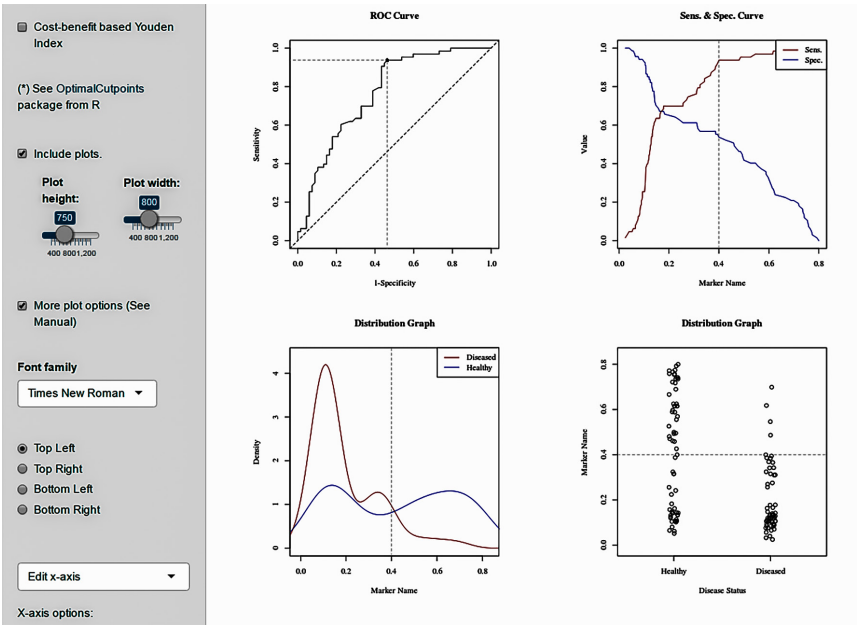
чувствительность = 0,937 (95% ДИ от 0,845 до 0,982);

специфичность = 0,537 (95% ДИ от 0,411 до 0,660);

диагностическая эффективность = (0,937 + 0,537) / 2 = 0,737, или 73,7 %.

Задание: проинтерпретируйте эти результаты в терминах процентного отношения верно классифицированных онкобольных и не имеющих рака простаты. Сравните этот результат со значениями для точки отсечения 0,278 (см. выше).

Таким образом, ROC-анализ позволил нам провести иное и несколько более качественное разделение, чем логистическая регрессия. Были верно распознаны почти все больные (93,7 %),



однако примерно у половины здоровых (53,7 %) также подозревалась онкология.

⑧ Построение графиков. В конце раздела **[Cut points]** ставим галочки в ☑ Include plots (Включить графики) и ☑ More plot options (See Manual) (Больше графических возможностей (см. Руководство)).

Программа автоматически строит четыре рисунка, и мы можем настроить общие и частные их элементы.

8.1. Делаем графики квадратными, двигая ползунки. Например, 750×800.

8.2. Выбираем Font family (Семейство шрифтов) [Times New Roman]

8.3. Далее редактируем по отдельности только нужные графики.

◉ Top Left — Верхний левый. Обязателен для включения в работу.

○ Top Right — Верхний правый. Чувствительность и специфичность. Необязателен.

○ Bottom Left — Нижний левый. Необязателен для включения в работу, но важен для интерпретации. На нём изображены функции плотности распределения для больных (Deseased)

и здоровых (Healthy). Обратите внимание, что в нашем случае оба графика бимодальные, то есть наши данные неоднородны и представлены какими-то двумя подгруппами. Учёт этих подгрупп позволит провести более качественную диагностику. На основании литературных данных можно предположить, что эти две группы сформированы мужчинами разного возраста.

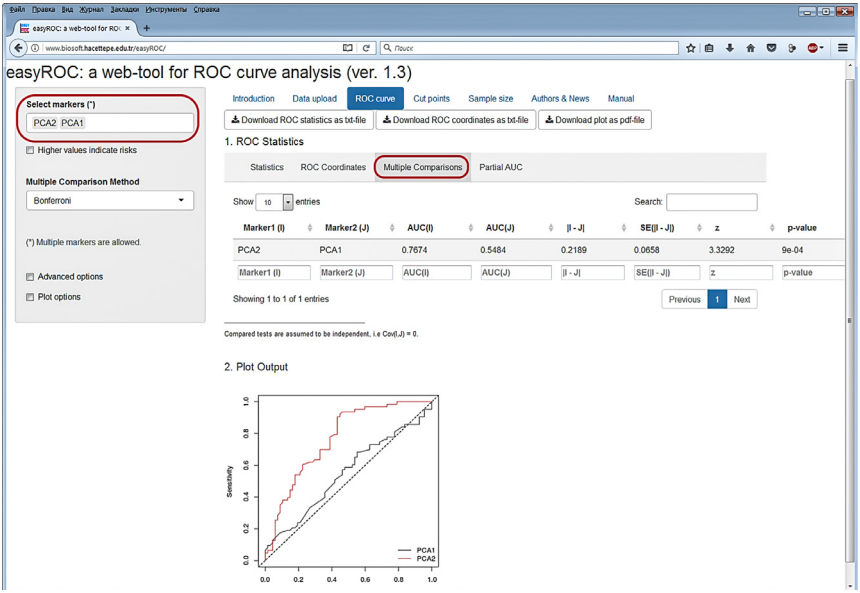
○ Bottom Right — Нижний правый. На нём изображена так называемая ***точечная диаграмма*** (*Dot diagrams*). Каждая точка — отдельный пациент. Здесь наглядно видно, что относительно порогового значения 0,4 группа здоровых делится примерно пополам, а вот в группе больных неверно классифицированными остаются только 4 человека. Значит, верно распознаны 63 – 4 = 59 больных, то есть чувствительность составила 59 / 63 = 0,937 (см. выше). Таким образом, данный график весьма нагляден и полезен, а потому также рекомендуется включить его в результаты.

Для каждого графика следует зайти в: Edit x-axis (Редактирование оси X), Edit y-axis (Редактирование оси Y), Other options (Другие опции). Изменить названия на русские, обязательно увеличить размеры шрифтов, можно сделать какие-то элементы цветными.

⑨ Сохранение графиков. Нажмите правой клавишей мыши на любой области графиков и выберите «Сохранить изображение». В указанном месте будет сохранён рисунок в формате *.png, из которого в любом растровом графическом редакторе (например, в Paint) нужно аккуратно вырезать нужные графики и сохранить отдельно.

⑩ Программа easyROC позволяет не только оценивать показатели диагностической эффективности отдельных тест-систем, но и сравнивать их между собой. Если необходима оценка одновременно нескольких показателей, то в разделе **[ROC curve]** выбирается одновременно несколько показателей, а на закладке MultipleComparisons (Множественные сравнения) получают результаты сравнения (рис. на с. 210). По умолчанию сравнение проводится методом Бонферрони с получением площади под *z*-кривой стандартного нормального распределения.

Задание. Сравните самостоятельно диагностическую эффективность показателей ПСА общий (PCA1) и отношения ПСА свободный/общий (PCA2).



⑪ Оформление в квалификационной работе (вариант).

11.1. Раздел «Материал и методы».

Оценку диагностической эффективности проводили по показателям чувствительности, специфичности, а также площади под характеристической кривой (ROC-кривой). Для первых двух мер 95% ДИ рассчитывали методом Джеффриса, для последней — точным биномиальным методом. Значимость отличия площади от 0,5 проводили по *z*-критерию, а точку оптимального разделения групп по величине показателя-маркёра находили по критерию Юдена. Расчёты и графические построения выполнены в пакете easyROC (version 1.3. (Дать ссылку на источник)).

11.2. Раздел «Результаты и обсуждение».

В таблицах можно привести абсолютные частоты, по которым проводился расчёт чувствительности и специфичности, и обязательно — сами показатели диагностической эффективности с 95% ДИ. Также привести и обсудить графики.

Обсуждение обычно строится вокруг величины граничного значения, поскольку именно от его выбора и зависит успех диагностики. Также отмечаются другие интересные детали. В нашем примере к таковым относятся достаточно низкая специфичность,

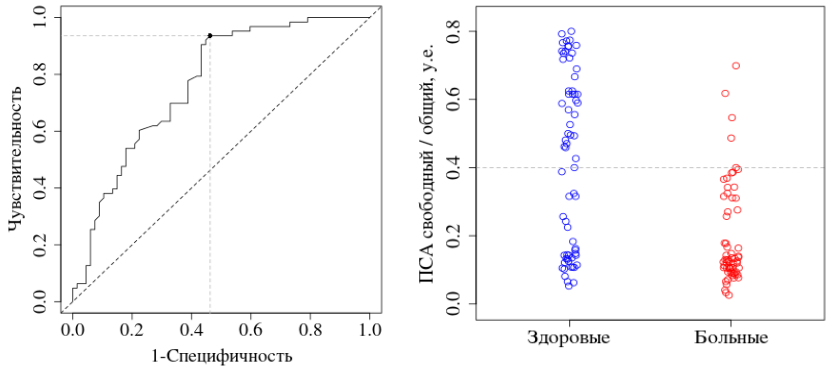


Рис. 14.3. Результаты ROC-анализа: слева — ROC кривая с точкой максимальной диагностической эффективности, справа — диаграмма распределения пациентов относительно оптимального порогового значения показателя

а также бимодальное распределение показателя в обеих группах, которое требует объяснения. Поскольку в подобных исследованиях обычно одновременно оценивается качество сразу нескольких диагностических систем, в результатах и обсуждениях приводится также информация о результатах их статистического сравнения.

11.3. Раздел «Выводы».

В ходе ROC-анализа была доказана диагностическая значимость показателя «Отношение ПСА свободный/общий»: площадь под характеристической кривой *AUC* = 0,767 (95% ДИ: от 0,685 до 0,867); $p << 0{,}001$. При значении показателя 0,4 достигалась максимальная диагностическая эффективность (73,7 %), в том числе чувствительность — 93,7 % (95% ДИ: от 84,5 до 98,2) и специфичность — 53,7 % (95% ДИ: от 41,1 до 66,0).

ЛАБОРАТОРНАЯ РАБОТА № 15

Работа с пространственными данными. Построение карт-схем

Тема 13. Некоторые специфические задачи в биологических исследованиях.
Количество часов: 2.
Цель: Освоить ряд методов интерполяции пространственных данных и научиться строить карты-схемы с полигонами Вороного, с цветными ячейками и цветными изолиниями. Работа на ПК.

► **Пространственные данные** (географические данные, геоданные, *spatial data*) — данные о пространственных объектах и их наборах, состоящие из двух частей: координатных данных и атрибутивных данных.

Координатные данные определяют позиционные характеристики пространственного объекта, то есть описывают его местоположение в установленной системе координат. В экспериментальных исследованиях могут использоваться свои системы координат, тогда как в натурных исследованиях это обычно знакомые всем географические координаты в системе WGS84 (World Geodetic System 1984): широта и долгота в градусах, минутах и секундах: например, 55° 39’ 09,50’’ северной широты и 61° 25’ 06,57’’ восточной долготы. Именно эти координаты отображаются на географических картах и выдаются навигаторами GPS и/или ГЛОНАСС.

Атрибутивные данные представляют собой совокупность непозиционных характеристик (атрибутов) пространственного объекта в виде количественных и/или качественных данных. В географии такие данные могут характеризовать, например: высоты, глубины, температуры, типы почв, распространение полезных ископаемых; в биологии — обилие видов на участке ареала, степень химического или биологического загрязнения территории, антропогенную деформацию экосистемы; в медицине — распространённость синдрома или заболевания на какой-либо территории и т. п.

Процедура задания связи атрибутивных данных с координатными, то есть «привязка» объектов к координатам, называется ***геокодированием***.

Методы обработки пространственной информации постоянно совершенствовались и к настоящему времени выделились в самостоятельную науку — ***геоинформатику***. Одним из её направлений является *интерполяция* пространственных данных и их отображение на картах.

Для интерполяции значений между точками пространственных данных на пространственный объект накладывается сеть с определённым размером ячеек, которая у картографов называется ***гридом*** (*grid*). В случае двумерного объекта это 2D-грид (2D-grid). Для построения изображения специальные алгоритмы рассчитывают значения для каждой ячейки грида. Понятно, что чем меньше ячейка, тем более детализированные получаются карты, чем больше ячейка — тем более грубые.

Количество таких алгоритмов (методов, моделей) пространственной интерполяции велико. С определённой долей условности их можно разделить на *детерминированные* (используют геометрические функции), *геостатистические* (задействуют информацию о статистических закономерностях распределения данных в пространстве) и *стохастические* (усредняют результаты реализации большого числа равновероятных моделей пространственного распределения данных). На этом занятии мы научимся трём способам графического отображения данных на картах и картах-схемах, которых должно хватить для обычно скромных потребностей биологов и медиков.

Используем для задачи интерполяции условно бесплатную программу 3DField (version 4.3.9.1, автор — Владимир Галушко). Незарегистрированная версия пакета обрабатывает не более 50 пространственных точек и не печатает 3D-вид, о чём сообщается в окне с предложением регистрации при каждом запуске программы. В отличие от географов биологи редко имеют даже два-три десятка пространственных точек, поэтому возможностей незарегистрированной версии многим должно хватить.

П р и м е р. Рассмотрим распределение содержания меди в донных отложениях Аргазинского водохранилища (Челябинская область). Данный водоём находится в ближайшей зоне влияния Карабашского металлургического комбината, куда медь попадает с водами, дренирующими пиритсодержащие хвостохранилища комбината (так называемый Рыжий ручей, далее р. Сак-Элга

и далее — р. Миасс). В 10 точках дночерпателем были отобраны донные отложения; в них атомно-абсорбционным спектрометрическим методом определено содержание меди. Получены данные:

Станция	Концентрация Cu, мг/кг сухого вещества
1	1365
2	1954
3	2067
4	395
5	1643
6	1692
7	603
8	405
9	174
10	20,5

Карта с указанием точек отбора проб прилагается (https://yadi.sk/d/g50i73pt3J6pAa). Требуется построить карту загрязнения донных отложений Аргазинского водохранилища медью.

Этап 1. Геокодирование

В пакете 3DField

① Включаем русскоязычный интерфейс (при первом запуске программы). Путь: View — Language — Russian.

② Выбор фоновой карты. На панели слева от окна программы: Задание — Выбрать изображение — … указываем место с файлом … Аргази.jpg.

③ Задание границ объекта. Путь: Объекты — Граница — Определить границу.

Вид стрелки над изображением меняется. Далее нужно обвести границу водоёма, кликая правой клавишей мыши. Это нужно сделать не слишком грубо, но и не слишком подробно. Когда обведёте полностью, нужно протянуть последнюю линию к форме с кнопками и нажать кнопку [Добавить к объекту].

Далее выбираем [Цвет линии — чёрный], [Ширина] — 2, и сохраняем границу в файл: [Записать в файл] … указываем место для файла … Аргази.brd.

④ Задание пространственных точек со значениями. Путь: Объекты — Точки — Добавить.

Становимся указателем мыши в центр кружка станции № 1 и дважды кликаем. В форме «Добавить точку» появляются координаты X и Y. В поле [Значение] вбиваем первое значение — 1365, в поле [Название] — 1 и нажимаем кнопку [Да].

В форме «Точки данных» появляется первая строка с данными, а на карте — точка и значение в ней (рис. 15.1). **Задание**. Введите аналогично оставшиеся 9 точек. В программе в качестве десятичного разделителя используется точка, поэтому последнее значение 20,5 не введётся (попробуйте), но введётся 20.5.

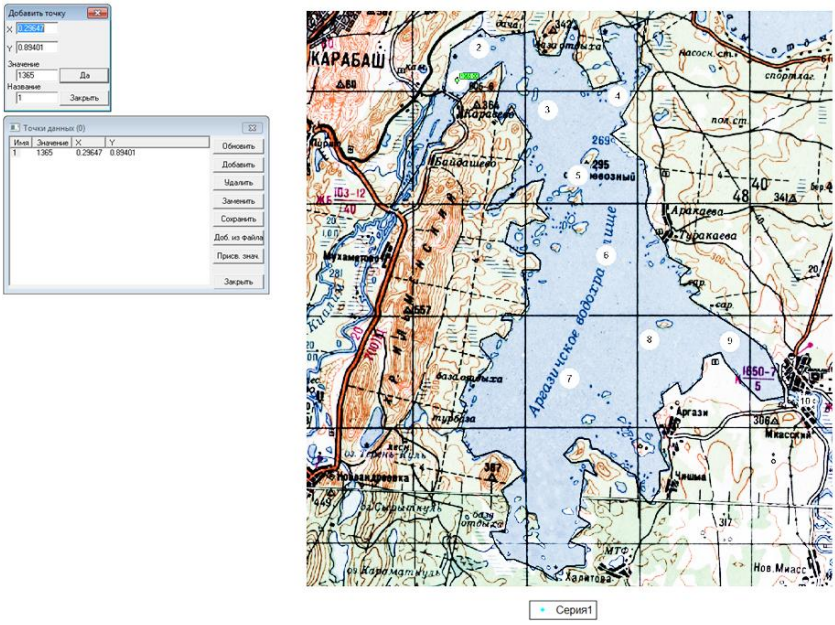


Рис. 15.1. Геокодирование в программе 3DField

Если при вводе была допущена ошибка, то удалите соответствующую строку в форме «Точки данных», нажмите [Обновить] и вводите снова; порядок здесь необязателен, главное — ничего не пропустить. Когда будут безошибочно введены все 10 точек, нажмите [Обновить] (индикация точек на карте изменится) и [Сохранить] … указываем место для файла … Медь.dat

⑤ Сохранение проекта. Путь: Файл — Запомнить документ как … указываем место для файла … Аргази_медь.3DF.

Этап 2. Анализ пространственной информации

Закрываем форму «Точки данных» и отключаем отображение карты: Объекты — Фоновая карта — Снимаем галочку ☐ Показать на карте. В окне остаётся только граница объекта и точки в нём.

2.1. Построение карты-схемы с полигонами Вороного

Выбираем в левом меню: Задание — Карты — ***Полигоны Вороного*** (*Voronoi polygon, Voronoi diagram*). В результате объект будет разделён на регионы таким образом, что любое положение внутри такого региона будет находиться ближе к значению этого региона, чем к значению любого другого региона. Таким образом, для *нерегулярных сеток* данных — как в нашем случае — получается наиболее обоснованное разделение площади исходя из имеющейся информации.

Если цвета, размеры и шрифты для обозначения точек и меток к ним не устраивают, это можно изменить: Формат — Значки точек.

Далее сохраняем изображение: Файл — Запомнить картинку — Медь_полигоны.png (или другой растровый формат: *.bmp или *.tif).

Карту с полигонами можно доработать в любом графическом редакторе, например — в Paint из набора Стандартных программ во всех версиях Windows. Если имеются какие-то градации признака, например «низкий уровень», «средний уровень» и «высокий уровень» содержания вещества, то можно раскрасить полигоны в соответствующие цвета, используя инструмент «Заливка цветом» (рис. 15.2). В результате получится карта-схема с градациями содержания вещества, разбитая на полигоны способом, который можно считать оптимальным для имеющегося набора пространственных точек. Поскольку информация на такой карте представлена весьма упрощённо, полученное изображение правильнее называть не «картой», а «картой-схемой».

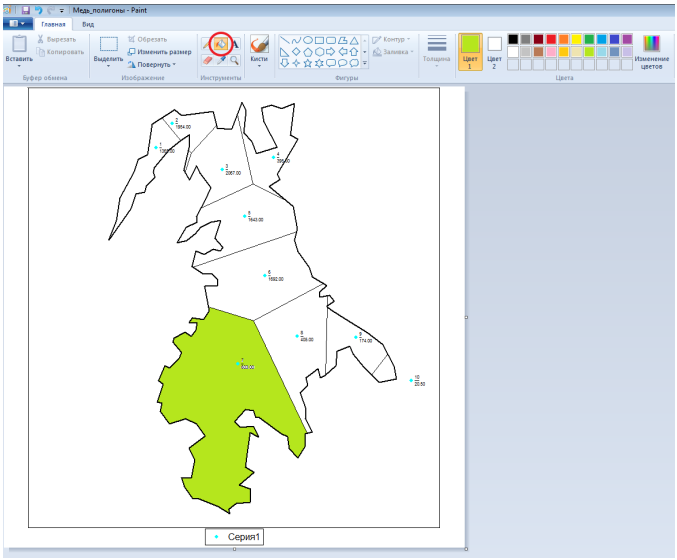


Рис. 15.2. Раскраска полигонов Вороного в пакете Paint для Windows 7

2.2. Построение карты с цветными ячейками

Путь в левом меню: Задание — Карты — Цветные ячейки.

Видно, что результат выглядит очень выигрышно, «профессионально». Программа разбивает поверхность на сеть ячеек, которые раскрашивает в цвета от тёмно-синего (соответствует минимуму) до красного (соответствует максимуму). Размер ячеек по умолчанию нас устраивает (его можно и изменить в: Объекты — Сетка — Размеры сетки), однако цветные ячейки выходят за границы объекта, поэтому границу лучше отключить: Объекты — Граница — Снимаем галочку ☐ Показать на карте.

По умолчанию для ***пространственной интерполяции*** (*spatial interpolation*) между значениями на карте программа использует систему линейных уравнений. Это можно увидеть в меню с пиктограммами над изображением. Рассмотрим бегло остальные интерполирующие функции. Для этого будем выбирать в меню другие способы интерполяции, а затем дважды кликать мышью на «Цветные ячейки» в левом меню экрана (Задание — Карты — Цветные ячейки).

1. ***Система линейных уравнений*** (*system of linear equations*).

Относительно простой с вычислительной точки зрения метод.

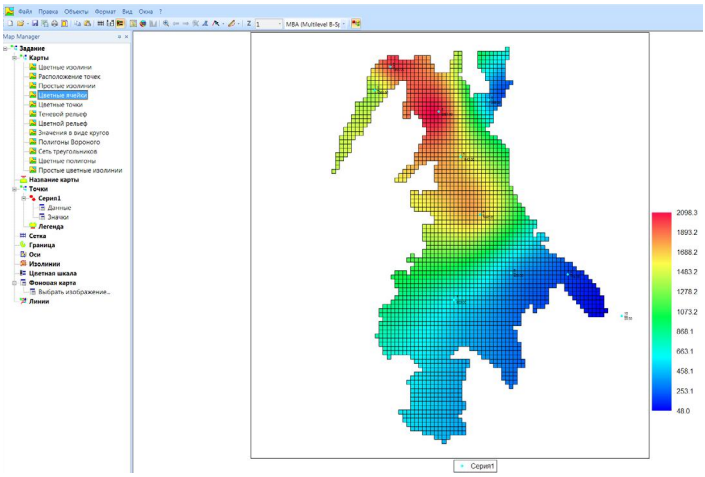


Рис. 15.3. Интерполяция методом MBA с использованием цветных ячеек в пакете 3DField

Интерполирует не очень хорошо, поскольку видно, что точки окружены достаточно яркими ореолами, как будто именно в этих местах концентрация отличалась от окружающего пространства. В других аналогичных пакетах данный метод не применяется.

2. ***Метод треугольников*** (***триангуляция***, *triangulation*).

Классический, но весьма требовательный к объёму данных метод. Как уже видно по результату, он интерполирует жёстко в пределах имеющегося полигона данных, а нам бы хотелось, чтобы метод ещё и *экстраполировал* значения к границам объекта.

3. ***Метод обратного расстояния*** (*inverse distances*).

Алгоритм метода относительно простой: значения в ячейках грида получаются путём усреднения значений вблизи этой ячейки; причём чем ближе точка к ячейке, тем больший вес она получает. Данный метод имеет известный недостаток — склонность к образованию на изображении так называемых «бычьих глаз» (*bull's eyes effect*) — больших округлых ореолов вокруг точек данных. Такой результат выглядит не очень естественно.

4. ***Кригинг*** и ***блочный кригинг*** (*kriging*, *block kriging*).

Это единственные собственно *геостатистические* техники в пакете. Результат выглядит очень неплохо, однако не является оптимальным. Пакет 3DField не имеет развёрнутого геостатистического блока и в качестве параметров значений кригинга авто-

матически использует логарифмически линейную модель зависимости ½ дисперсии данных γ от расстояния между объектами x вида $\gamma = a + \lg x$. Соответствующую ***вариограмму*** (*variogram*) можно посмотреть, кликнув правой кнопкой мыши в левом меню на Данные (Точки — Серия 1 — Данные), а далее проследовать в Анализ данных — Вариограмма.

Чтобы настроить модель кригинга более точно, нужно в ***геостатистических пакетах*** (*geostatistical packages*) провести построение и анализ вариограмм, используя координаты точек из пакета 3DField. Затем выписать параметры наиболее подошедшей модели из числа тех, что предлагает геостатистический пакет. Чаще всего используют ***сферическую модель*** (*spherical model*), но не обязательно она будет лучшей. Далее в 3DField нужно зайти в: Объекты — Интерполяция — Кригинг — Опции и внести полученные параметры модели в поля Variogram, Range, Nugget и Sill. Учитывая, что геостатистические пакеты также нуждаются в освоении и содержат собственный арсенал графических средств, это может оказаться слишком трудоёмким, поэтому пока в 3DField кригингом следует пользоваться с осторожностью.

5. ***Минимальное искривление*** (*minimum curvature*).

По длительности процесса и информации в открывшемся дополнительном окне можно понять, что алгоритм этого метода итерационный (Iteration). Он состоит из двух этапов: 1) локальной интерполяции и 2) глобальной экстраполяции. Специалисты могут настроить параметры этого метода в: Объекты — Интерполяция — Минимальное искривление — Опции, а для наших целей можно использовать настройки метода по умолчанию.

6. ***Метод «естественного соседа»*** (*natural neighbours*, *Sibson's method*, метод естественной окрестности, интерполяция Сибсона) находит самое близкое подмножество входных образцов к пространственной точке и для интерполяции использует взвешенные значения, основанные на пропорциональных областях, вычисленных по Сибсону.

7. *RBF* (*radial based function*, ***радиальная функция***) и блочная RBF. Радиальная функция хорошо подходит, когда точно известно, что пространственный объект имеет округлую форму. **ВАЖНО**: например, это может быть рудное тело, озеро, зона аэрозольного загрязнения при слабом ветре. В нашем случае данный тип интерполяции не подходит.

8. MBA (*multilevel B-spline Approximation*, ***многоуровневая аппроксимация сплайнами***). Мы кратко знакомились со сглаживанием сплайнами в случае зависимости вида $y = f(x)$. В данном случае используется аналогичный способ сглаживания, но только применительно к пространственным данным. Образно это можно представить так: на точки карты помещаются штыри, высота которых пропорциональна концентрации меди, а сверху на штыри накладывается пластина из упругого материала и прижимается так, чтобы она соприкасалась со всеми штырями. В результате минимизируется общая кривизна поверхности, а сама поверхность проходит точно через точки исходных данных. Это простой метод, не требующий специальных знаний в области геоинформатики и часто дающий неплохой результат.

9. *Zones* (***метод зон***) — малоизвестный метод. На момент написания лабораторного практикума информации об этом методе не содержалось ни в помощи к пакету, ни на сайте проекта. Поэтому пока данным способом пользоваться не следует.

10. В других пакетах могут присутствовать и другие методы. Например, для моделирования аэрозольных загрязнений широко применяются специальные *модели «факела»*. Поскольку такие модели разрабатываются с учётом специфики явления, они более точны, чем рассмотренные нами интерполяторы, а входящие в них параметры могут иметь конкретный физический смысл.

Вернёмся к варианту MBA и остановимся на нём как на достаточно простом и эффективном. На эффективность косвенно указывает отсутствие явных ореолов вокруг точек с данными и логичность наблюдаемой картины уменьшения концентрации по мере удаления от источника — металлургического комбината и его хвостохранилищ, находящихся к северо-западу от водохранилища. Также нужно отметить сложный контур загрязнения: поскольку преобладающая часть меди поступает в водоём в составе твёрдых глинистых и органических частиц, контур несколько вытянут и деформирован в направлении старого русла реки Миасс, по которому скорость движения воды максимальна.

Полученное изображение нужно доработать и сохранить:

1. Изменение значков и подписей к точкам данных. Дважды кликаем на области [Серия 1] под рисунком или заходим по пути: Формат — Значки точек. Изменяем размеры и цвет

значков, шрифта; отказываемся от лишних десятичных знаков. Выбираем также количество отображаемой на карте информации: нужны ли номера станций отбора проб, нужны ли собственно значения концентрации меди или эта информация даётся в работе в ином месте и т. д. От отображения прямоугольника [Серия 1] можем отказаться, сняв галочку в ☐ Включить в изображение.

2. Изменяем значения на шкале легенды. Для этого дважды кликаем на область цифр в легенде, перемещаем её внутрь рисунка, настраиваем минимум и максимум таким образом, чтобы шаг деления шкалы был удобным для восприятия, например кратным 10. Также настраиваем тип, размер и цвет шрифта. **Внимание!** Пакет предоставляет на выбор пользователя большое число палитр. С этим инструментом важно не переусердствовать: вы строите научную графику, а не рисуете рекламный буклет. Помните, что жертвовать информативностью карты в погоне за яркостью недопустимо, это признак непрофессионализма. В нашем случае палитра № 0 подходит как нельзя лучше. Однако, например, для карты глубин водоёма логичнее будет выбрать палитру с градациями синего типа палитр № 38 и 56. При этом нужно будет поместить глубину со значением 0 (ноль) на границу объекта (путь: Объекты — Граница — Значение, ☑ Учитывать значение на границе; в качестве значения выставить 0), а для интерполяции использовать радиальную функцию RBF.

3. Сохраняем изображение. Для возможности внесения правок в будущем: Файл — Запомнить документ как … — выбираем формат 3DF. Для вставки карты в работу: Файл — Запомнить картинку … — выбираем растровый формат типа *.png или *.tif.

2.3. Построение карты с цветными изолиниями

1. Построение изображения. Процедура очень сходна с ячейками, многое делается аналогично. Выбираем в левом меню: Задание — Карты — Цветные изолинии.

2. Изменение цветов. Кликаем на пиктограмму «Изменить цветную шкалу» или на область цифр в легенде. В уже знакомом меню закладки 1 выбираем подходящую палитру, например № 37, и изменяем порядок цветов на обратный: Интерп. [Reverse]. Также вручную корректируем значения на требуемых изолиниях (рис. 15.4). Здесь же можно выбрать [Число цветов],

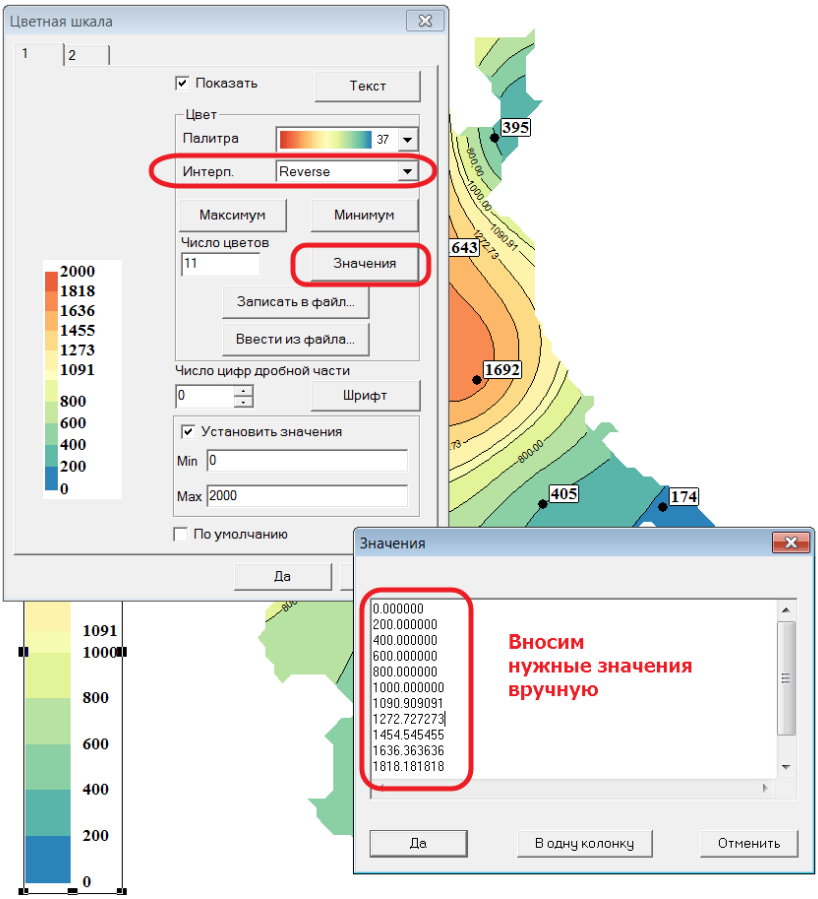


Рис. 15.4. Изменение цветной шкалы при интерполяции цветными изолиниями в пакете 3DField

если нас не устраивает предлагаемое по умолчанию значение. Выберем 8 цветов, чтобы шкала содержала значения от 0 до 2 000 с шагом 250 мг/кг. При этом некоторые этапы настройки, возможно, придётся повторить. На закладке 2 можно ☑ включить рамки для цветов.

3. Надписи. На карте с изолиниями логично отказаться от значений станций и значений концентраций в точках, чтобы не перегружать изображение цифрами. Поэтому заходим: Формат — Значки точек и в окне Цифровое значение параметра выбираем [Не показывать].

4. Формат изолиний. На левой панели управления находим раздел Изолинии и дважды кликаем на любой ненулевой изолинии, например, «изолиния 250». Попадаем в меню изолиний, где выбираем ширину линий, изменяем шрифт и т. д. (рис. 15.5). После всех настроек ставим галочку [V] Все изолинии для глобального применения настроек.

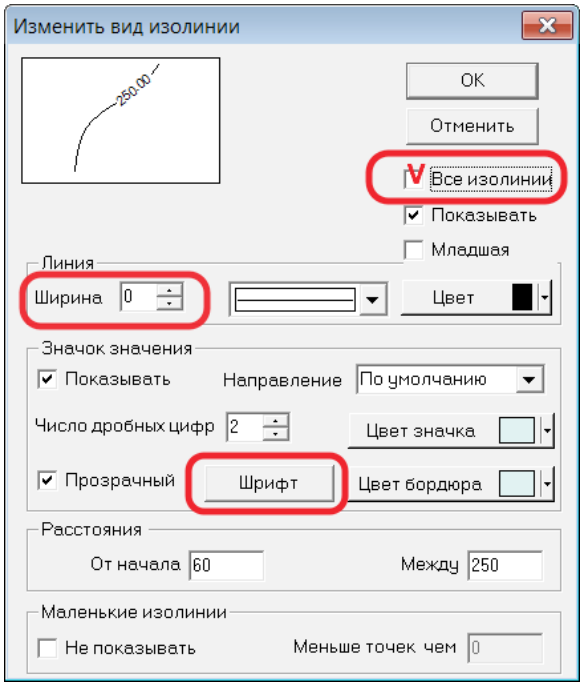


Рис. 15.5. Настройка изолиний в пакете 3DField

5. Граница. Можем вернуть границу: Объекты — Граница — [V] Показать на карте. Но тогда, в целях улучшения качества изображения, рисунок придётся дорабатывать в растровом редакторе для заполнения непрокрашенных областей вблизи границы. Учитывая широкий функционал, простоту, русскоязычный интерфейс и условную бесплатность пакета 3DField, такая доработка — меньшая из всех возможных проблем на пути создания качественной графики. Доработанный вариант см. на рис. 15.7.

6. Сохраняем проект и изображение в растровом формате.

Этап 3. Оформление в квалификационной работе

3.1. Статистическая часть раздела «Материалы и методы»

Для построения карт загрязнения донных отложений водоёма тяжёлыми металлами проводили многоуровневую интерполяцию сплайнами значений между изученными точками (Multilevel B-Spline Approximation, MBA). Геокодирование и графические построения выполнены в пакете 3DField (version 4.3.9.1 (Дать ссылку на источник)).

3.2. Раздел «Результаты и обсуждение»

Приводятся карты либо с цветными ячейками (рис. 15.6), либо с изолиниями (рис. 15.7). Оба типа карт давать не следует. Здесь это сделано только в целях демонстрации доработанных образцов таких карт.

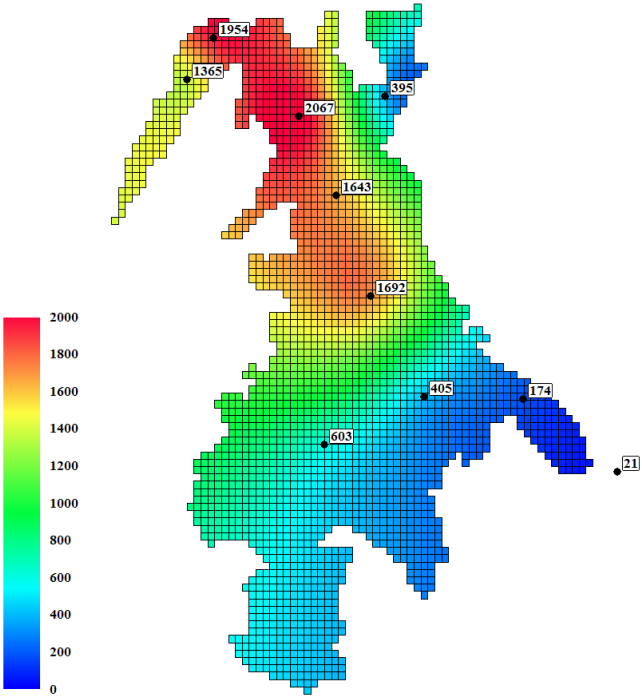


Рис. 15.6. Концентрация меди (мг/кг сух. вещ-ва) в донных отложениях Аргазинского водохранилища. Интерполяция методом MBA (цветные ячейки)

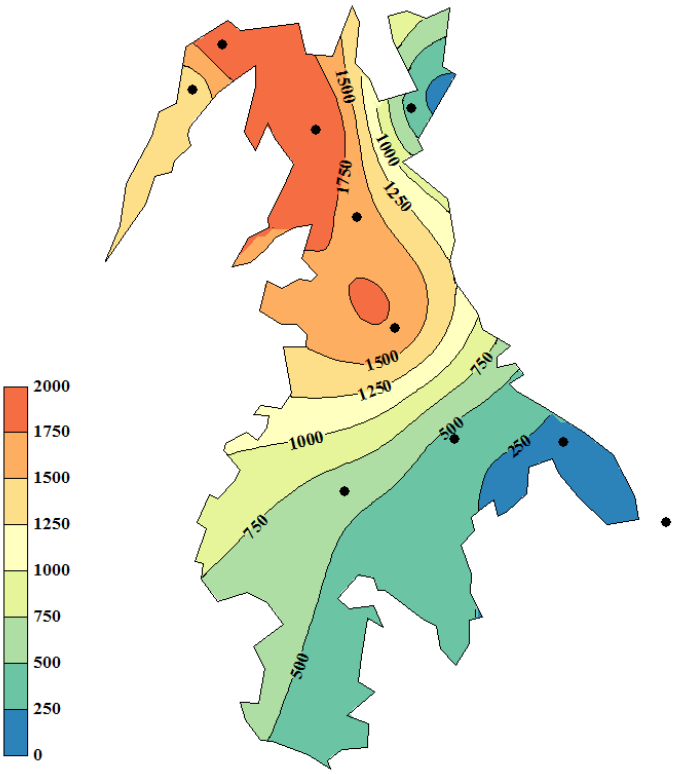


Рис. 15.7. Концентрация меди (мг/кг сух. вещ-ва) в донных отложениях Аргазинского водохранилища. Интерполяция методом МВА (изолинии)

В ходе обсуждения выделяются наиболее существенные моменты и специфика. Например, в нашем примере к общим моментам относится:

1) закономерность снижения концентрации с удалением от источника;

2) факт загрязнения элементом донных отложений всей акватории водохранилища, поскольку фоновое значение 20,5 мг/кг зарегистрировано в реке Миасс уже после плотины.

В качестве специфики можно обсудить:

1) ситуацию в отдельных курьях водохранилища, например, меньшую загрязнённость курьи Баик в северо-восточной части водоёма;

2) роль островов, создающих барьер на пути распространения загрязнений, что видно на рисунке с изолиниями.

Следует особо отметить, что все эти моменты стали заметны в результате интерполяции лишь по 10 значениям, что указывает на высокую информативность использования специфических средств обработки пространственных данных.

3.3. В разделе «Выводы»

В этом разделе помещаются наиболее важные с теоретической или практической точки зрения положения.

ЛАБОРАТОРНАЯ РАБОТА № 16

Кластерный анализ, анализ главных компонент и анализ главных координат

Тема 14. Многомерные методы разведочного анализа данных.
Количество часов: 2.
Цель: Освоить выбор и технологию использования методов кластерного анализа, анализа главных компонент (PCA) и анализа главных координат (PCoA). Работа на ПК.

Многомерные методы разведочного анализа данных очень многочисленны и разнообразны, но их можно разделить на четыре группы:

1) методы кластерного анализа;
2) ординационные (проекционные) методы;
3) методы интеллектуального анализа, или «Добычи данных» (*data mining*);
4) методы визуализации и редукции многомерных данных на основе нейронных сетей.

На этом лабораторном занятии мы познакомимся с тремя методами из первых двух групп.

1. Кластерный анализ

► **Кластерный анализ** (КА, *claster analyses*) — группа методов для обнаружения сходных объектов в многомерном пространстве признаков. В таком наиболее частом применении КА является классической ***Q-техникой***. Тем не менее если КА применяется к транспонированной матрице данных, то он может быть использован для нахождения наиболее сходных признаков в многомерном пространстве объектов, то есть в качестве ***R-техники***. Различные методы КА широко используются в области ***машинного обучения*** (*machine Learning*), где они относятся к ***методам обучения без учителя***.

Существует большое число вариантов кластерного анализа, но на практике статистические пакеты предлагают обычно три шага для настройки КА:

1) ***иерархический КА*** или ***метод К-средних***. Задача первого — найти сходные объекты и построить *дендрограмму сходства*, тогда как второго — разделить все объекты на заранее известное

число наиболее сходных групп. **ВАЖНО:** для разведочного анализа необходимо выбрать иерархический КА;

2) алгоритм кластеризации — математическое правило, по которому будут находиться сходные объекты. Наиболее часто применяемыми алгоритмами являются ***метод невзвешенного попарного среднего*** (*Unweighted Pair Group Method with Arithmetic Mean — UPGMA*) и ***метод Уорда*** (*Ward's method*);

3) расстояние между объектами, то есть единицы измерения сходства объектов. Более подробно о расстояниях см. теоретический материал, в том числе [8; 14], здесь лишь отметим, что обычно наиболее важным для классификации является соотношение признаков, а не степень их выраженности или размеры. Поэтому в качестве мер расстояния лучше использовать меры корреляции и ассоциации, а не линейные (евклидовы) расстояния.

Необходимо отметить, что несмотря на стандартные алгоритмы их программная реализация может иметь нюансы, что на практике нередко приводит к несколько различающимся видам дерева в разных пакетах — к этому нужно быть готовым и обязательно указывать в описании анализа конкретный использованный пакет и его версию.

П р и м е р. Рассмотрим иерархический кластерный анализ на примере данных по гаплогруппам Y-хромосомы человека у народов Европы. Гаплогруппы представляют собой наборы схожих вариантов определённых аллелей хромосомы, возникших в результате мутационного изменения нуклеотидов в последовательности ДНК и наследуемых совместно. Поскольку Y хромосома не участвует в кроссинговере, генетические изменения в ней передаются строго по отцовской линии: от отца сыновьям. Это делает Y-хромосому удобным объектом *генетической генеалогии* — установления родства по генетическим меркёрам.

Данные: информация о частотах встречаемости (в процентах) 12 гаплогрупп Y-хромосомы у народов Европы, взятая с ресурса: http://www.eupedia.com/europe/european_y-dna_haplogroups.shtml (23.03.2014). Во всех странах, кроме России, данные по входящим в состав страны этническим группам были удалены, знаки «?» заменены на 0. Такая версия данных находится в папке «Данные» по адресу https://yadi.sk/d/g50i73pt3J6pAa, файл «Гаплогруппы.doc». Более актуальную информацию читатели могут скачать самостоятельно с указанного сайта.

Страна/народность	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Албания	2	12	1,5	9	16	1,5	19,5	2	27,5	1	0	0
Австрия	12	7	2,5	19	32	7,5	9	1	8	1	0,5	0,5
Беларусь	5,5	17,5	1	51	5,5	1,5	2,5	1	4	0	0	10
Бельгия	12	3	4,5	4	61	4	4	1	5	1	0,5	0
Босния и Герцего-вина	3	55,5	0	15	3,5	1,5	4	0,5	12	1	2	2
Болгария	4	20	2	17	11	5	11	3	23,5	1,5	0,5	0,5
Хорватия	5,5	37	1	24	8,5	2,5	6	1	10	0,5	1	0,5
Чехия	11	9	4	34	22	5	6	0	6	1	1,5	0,5
Кипр	0	8	0	3	9	9	37	6	20	5	0	0
Дания	34	2	5,5	15	33	2,5	3	0	2,5	0	1	1
Англия	14	2,5	4,5	4,5	67	1,5	3,5	0	2	0,5	0,5	0
Эстония	15	3	0,5	32	8	0	1	0	2,5	3,5	0,5	34
Финляндия	28	0	0,5	5	3,5	0	0	0	0,5	0	0	61,5
Франция	8,5	3	3,5	3	58,5	5,5	6	1,5	7,5	1	0,5	0
Германия	16	1,5	4,5	16	44,5	5	4,5	0	5,5	1	0,5	1
Греция	3,5	9,5	1,5	11,5	15,5	6,5	23	3	21	4,5	0	0
Венгрия	8,5	16	2	29,5	18,5	3,5	6,5	3	8	0	0	0,5
Исландия	29	0	4	23	42	0	0	0	0	0	1	1
Ирландия	6	1	5	2,5	81	1	1	0	2	0	0	0
Италия	4,5	3	2,5	4	39	9	15,5	3	13,5	2,5	0	0
Косово	5,5	2,5	0	4,5	21	0	16,5	0	47,5	0	0	0
Латвия	6	1	1	40	12	0	0,5	0	0,5	0,5	0,5	38
Литва	6	6	1	38	5	0	0	0	1	0,5	0,5	42
Македония	3	23	1,5	13,5	12,5	4	14	2	21,5	1,5	0,5	0,5
Мальта	1	10	1	3,5	32,5	6,5	21	8	9	4,5	1	0
Молдова	5	21	3	30,5	16	1	4	4	13	1	0	1,5
Молдова_гагаузы	4	20	3	19	12,5	13,5	5,5	2	11	3	0	1
Черногория	6	29,5	1,5	7,5	9,5	2,5	9	0,5	27	0	2	1,5
Нидерланды	16,5	1	6,5	4	49	4,5	3,5	0,5	3,5	1	0	0
Норвегия	31,5	0	4,5	25,5	32	1	0,5	0	1	0	1	2,5
Польша	8,5	5,5	2	57,5	12,5	1,5	2,5	0	3,5	0,5	0,5	4
Португалия	2	1,5	3	1,5	56	6,5	9,5	3	14	2,5	0,5	0
Румыния	4,5	26	2,5	17,5	12	5	13,5	1,5	15	0,5	0,5	0,5
Россия (РФ)	5	10,5	0	46	6	1	3	0	2,5	1,5	1,5	23
РФ_башкиры	0	0,5	0	26	47,5	0,5	3	0	0,5	0	0	17
РФ_чуваши	7,5	1,5	2,5	31,5	4	0	0	0	0	0	0	28
РФ_коми	3	1	0	32	8	0	0	0	0	0	0	51
РФ_марийцы	5	1	1	37,5	2	2	0	0	0	2	0	49,5
РФ_мордвины	12	2,5	5	26,5	13,5	0	0	0	0	0	0	19,5
РФ_татары	2	5,5	0	31,5	7,5	11	9,5	0	3	5	2	20,5
РФ_удмурты	3	0,5	0,5	19	5,5	1,5	0	0	1,5	1,5	0	67
Шотландия	9	1	4	8,5	72,5	0,5	2	0	1,5	0,5	0,5	0
Сербия	8,5	33	0,5	16	8	2	8	0,5	18	1	1,5	2
Словакия	6,5	16	1,5	41,5	14,5	4	2	1	6,5	0,5	0,5	3
Словения	9	20,5	1,5	38	18	1,5	2,5	0	5	1	0	0
Испания	1,5	4,5	1	2	69	3	8	1,5	7	2,5	0	0
Швеция	37	1,5	3,5	16	21,5	1	2,5	0	3	0	2,5	7
Швейцария	14	1,5	8	3,5	50	7,5	3	0,5	7,5	0,5	1,5	1
Турция_Анатолия	1	4	0,5	7,5	16	11	24	9	11	2,5	2	4
Украина	3,5	13	0,5	45	7,5	2,5	7	1	5,5	1	1	7,5
Уэльс	7	0,5	2	1	83,5	2,5	1	0	2	0,5	0	0

В пакете PAST

Импорт данных:

① Открыть пустой лист и поставить галочки в ☑ Row attributes и ☑ Column attributes.

② Открыть текстовый файл Гаплогруппы.doc и скопировать все данные в буфер.

③ В пакете PAST встать на пересечение строки Name и колонки Name и вставить данные из буфера.

④ Снять галочки и сохранить файл как Гаплогруппы.dat.

Анализ:

① Выделить область значений.

② Путь: Multivariate — Clustering — Classical.

③ Algorithm (Алгоритм) — Paired group UPGMA, то есть используем метод невзвешенного попарного среднего.

④ Similarity index (Мера сходства) — выбираем Rho (ρ, «ро»). Это обычный коэффициент корреляции Спирмена. Данная мера рекомендуется к использованию для количественных и порядковых признаков, поскольку устойчива к асимметричным распределениям и учитывает нелинейность связи (см. теоретический материал). При анализе совстречаемости организмов, когда данные представлены качественными альтернативными признаками, закодированными как «0» (нет объекта) и «1» (есть объект), следует в первую очередь попробовать индексы Жаккара (Jaccard) и Раупа — Крика (Raup-Crick). **ВАЖНО!** Это может быть особенно полезно экологам и микробиологам для поиска ассоциаций организмов. С этими мерами мы познакомимся на следующем занятии.

⑤ Нажать кнопку [Compute] (Рассчитать) и развернуть окно на весь экран. Полученный в ходе анализа рисунок называется ***дендрограммой*** (рис. 16.1). Также пакет рассчитывает коэффициент ***кофенетической корреляции*** *(cophenetic corelation)*, который изменяется от 0 до 1 и показывает, насколько полно парные меры сходства/различий большого числа признаков удалось отразить на единственной дендрограмме.

Интерпретация:

5.1. Чем ближе располагаются объекты на ветвях такого дерева, тем более они сходны. В нашем дереве наиболее сходными народами (странами) являются чуваши и мордвины России. Видно также, что они близки к латвийцам, а вместе с народом коми —

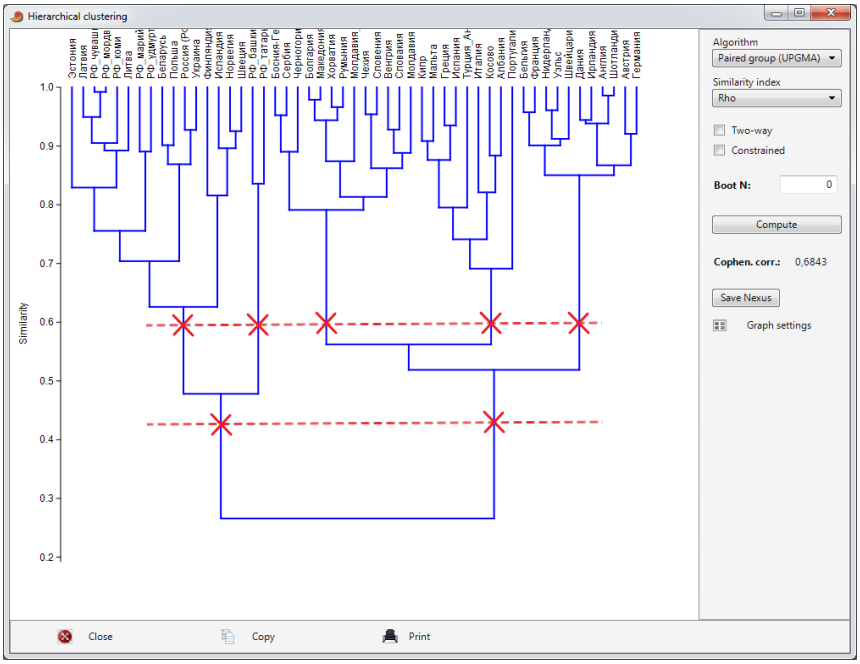


Рис. 16.1. Выделение кластеров при визуальной оценке дендрограммы

и к другим прибалтийским народам: литовцам и эстонцам. Генетически очень близки англичане и шотландцы. Найдите ещё несколько самых близких народов самостоятельно. **Задание!** Найдите страны, наиболее генетически близкие России.

5.2. Чем длиннее ветвь, на которой располагаются объекты, тем более отделена эта группа от других в многомерном пространстве признаков. На нашем дереве такой спецификой отличаются российские татары и башкиры, заметно отстоящие от группы народов России, восточной и северной Европы и Прибалтики. Также заметно отличаются от других народов жители стран центра и северо-запада Европы — они сгруппировались в своеобразный «зонтик» на длинной ножке в правой части графика.

5.3. Процесс выделения кластеров заключается в мысленном отсечении ветвей дерева на разных уровнях сходства (на рис. 16.1 для наглядности показано пунктиром) и подсчёте числа срезанных ветвей (крестики). Видно, что изначально все страны разделились на два больших кластера: в первый вошла Прибалтика, Россия,

Восточная и Северная Европа, а также российские башкиры и татары. Во второй кластер вошли все остальные страны/народы. Далее находим следующую область отсечения так, чтобы она прошла максимально возможно по длинным ветвям дерева. Она находится приблизительно на уровне сходства $\rho = 0{,}6$. Здесь выделяется пять кластеров: 1) Прибалтика, Россия, Восточная и Северная Европа; 2) башкиры и татары России; 3) страны Балканского полуострова за исключением его юга (Турция, Греция, Албания); 4) страны Средиземноморья и юга Балканского полуострова; 5) страны центральной и северо-западной Европы.

5.4. На основе знаний в предметной области выдвигаются гипотезы, объясняющие наблюдаемую кластерную структуру, и даётся интерпретация кластеров. Мы не являемся специалистами в антропологии и генетике человека, но можем предположить, что генетическая близость народов объясняется географической близостью и/или наличием устойчивых путей миграции населения (морские и сухопутные торговые пути, вхождение территорий в состав иных существовавших в историческое время государств и т. п.). Уже на предыдущем этапе, в попытке как-то обобщить группы входящих в кластеры стран, мы, по сути, уже сразу и интерпретировали кластеры по географическому принципу. Такая интерпретация представляется вполне логичной.

5.5. После интерпретации кластеров можно пойти дальше и обсуждать положение и/или специфику тех или иных интересующих объектов. Например, обращает на себя внимание высокое генетическое разнообразие народов первого кластера и низкое — последнего. Также интересно отметить противоречия между биологической близостью и политическими разногласиями ряда стран (Великобритания — Шотландия, Россия — Польша, Прибалтика, Украина).

⑥ Дендрограмму для вставки в работу необходимо доработать: перевести ось Y (Similarity — сходство), изменить весь шрифт на Times New Roman, подобрать размер шрифта так, чтобы в окончательном документе его размер выглядел аналогично шрифту основного текста или был меньше его на 1–2 пункта. Частично это можно сделать в PAST (в правой части формы с рисунком — Graph settings), но лучше сохранить правленый рисунок в формате *.svg и доработать окончательно в векторном графическом редакторе (TpX, Inkscape, Corel Draw и т. п.).

⑦ **Оформление в квалификационной работе или статье (вариант)**.

7.1. Статистическая часть раздела «Материал и методы».

Для нахождения сходных по сочетанию гаплогрупп народов стран Европы использовался иерархический кластерный анализ, который был проведён методом UPGMA. В качестве меры близости использовался коэффициент корреляции Спирмена. Расчёты и графические построения выполнены в пакете PAST (version 3.19, Hammer et al., 2001).

7.2. Раздел «Результаты и обсуждение».

Результаты кластерного анализа представлены на рис. N. Анализ дендрограммы показывает, что… Далее проводится описание наблюдаемой кластерной структуры, кластеры по возможности интерпретируются; проводится обсуждение результатов кластеризации с привлечением литературных данных.

7.3. Раздел «Выводы».

В ходе кластерного анализа было установлено, что по соотношению 12 гаплогрупп в народах Европы выделяется четыре кластера: 1) Прибалтика, Россия, Восточная и Северная Европа с примыкающим к ним кластером башкир и татар России; 2) страны Балканского полуострова за исключением его юга (Турция, Греция, Албания); 3) страны Средиземноморья и юга Балканского полуострова; 4) страны центральной и северо-западной Европы.

* * *

Таким образом, обычный кластерный анализ — весьма простой, быстрый и полезный метод. Строго говоря, это даже не статистический метод, а визуализация математического правила, находящего близко расположенные объекты в многомерном пространстве. Однако он имеет ограниченные возможности для интерпретации: в нашем случае осталось непонятным, в чём именно заключалось сходство и различие тех или иных народов на дендрограмме? Чтобы ответить на этот вопрос, необходимо далее рассчитать усреднённые показатели по странам, входящим в кластер, и сравнивать их с другими кластерами, что весьма трудоёмко.

Более принципиальным недостатком кластерного анализа, про который в литературе обычно не упоминается, является сильная зависимость результатов от набора признаков. Например, если в наборе из 10 признаков 8 отражают размерно-возрастные раз-

личия объектов и только 2 чётко характеризуют половые различия, дендрограмма сходства будет отражать именно размерно-возрастные кластеры, а ветвление по половой принадлежности будет подавлено. Поэтому лучше проводить кластеризацию не самих исходных данных, а меток обобщающих переменных, полученных ординационными техниками. Такие техники разведочного анализа сложнее, но гораздо информативнее. Познакомимся с некоторыми из них.

2. Ординационные методы

► **Ординация** (*ordination*) — процесс уменьшения размерности многомерных данных путём получения из наблюдаемых переменных меньшего числа новых переменных, которые содержат бóльшую часть информации исходных данных. Пространство новых переменных можно рассматривать как проекцию многомерного пространства исходных данных, а потому ординационные методы называют также ***проекционными методами***.

В зависимости от того, что лежит в основе анализа — матрица сходства признаков или матрица расстояний между объектами, ординационные техники делятся на соответственно: *основанные на вычислении собственных чисел* (*eugenvalue-based methods, eigenanalysis-based methods*) и *основанные на расстояниях* (*distance-based methods*). Такое разделение — скорее историческое, поскольку один и тот же метод математически можно выразить с помощью обоих подходов, а результаты использовать как для анализа сходства признаков по набору объектов (R-техника), так и для анализа сходства объектов в пространстве набора признаков (Q-техника). С практической точки зрения важнее различать следующие две группы методов:

1) ***неограниченные*** (*unconstrained*), или ***методы непрямого градиентного анализа***. На анализ варьирования показателей не накладывается никаких ограничений: анализируется вся изменчивость, присущая одному набору данных. К таким методам относятся: ***анализ главных компонент*** (*principal component analysis, PCA*), ***анализ главных координат*** (*principal coordinate analysis, PCoA, PCO*), ***анализ соответствий*** (*correspondence analysis, CA*);

2) ***ограниченные*** (*constrained*), или ***методы прямого градиентого анализа***. Данные в этом случае состоят из двух разнока-

чественных блоков: блока независимых переменных (регрессоров) и блока зависимых переменных (откликов). Анализируется не вся изменчивость в наборах, а только та часть изменчивости откликов, которая объясняется регрессорами. Таким образом, данные техники являются по сути многомерным вариантом регрессии. К таким методам относятся: ***анализ избыточности*** (*redundancy analysis, RDA*), ***регрессия методом частных наименьших квадратов*** (*partial least squares, PLS, PLS-regression*), ***канонический анализ соответствий*** (*canonical correspondence analysis, CCA*).

На этом занятии мы познакомимся с двумя неограниченными многомерными техниками: анализом главных компонент и анализом главных координат.

2.1. Анализ главных компонент

Анализ главных компонент, АГК (*principal component analysis, PCA*) — исторически первый и наиболее распространённый многомерный метод анализа данных. Впервые был предложен К. Пирсоном в 1901 г., а затем независимо переоткрыт и разработан Г. Хотеллингом в 1930-х гг. Метод является классической *R-техникой,* в основе которой лежит анализ *матрицы сходства* признаков, а мерой сходства является корреляция Пирсона. В ходе такого анализа из большого числа каким-либо образом связанных между собой признаков получают меньшее число несвязанных между собой ***главных компонент*** (ГК), которые являются *линейной комбинацией* этих признаков и обычно обусловлены действием нескольких различных непосредственно ненаблюдаемых ***латентных переменных — факторов***.

Например, мы можем выйти на улицу и предложить тысяче случайных прохожих поучаствовать в исследовании. У каждого человека (мужчины и женщины, взрослые и дети) будем измерять 100 морфометрических признаков: рост, длина конечностей, обхваты туловища и конечностей, межглазничное расстояние и т. д. В результате мы получим матрицу данных 1 000 строк × 100 столбцов. Рассмотрим упрощённый вариант: только 2 признака из 100 и 12 человек из 1 000, для которых вычислена обычная корреляция Пирсона (рис. 16.2, *а*).

Найдём центр этой системы *O* и проведём через него две перпендикулярные прямые — главные компоненты (ГК): ГК 1 —

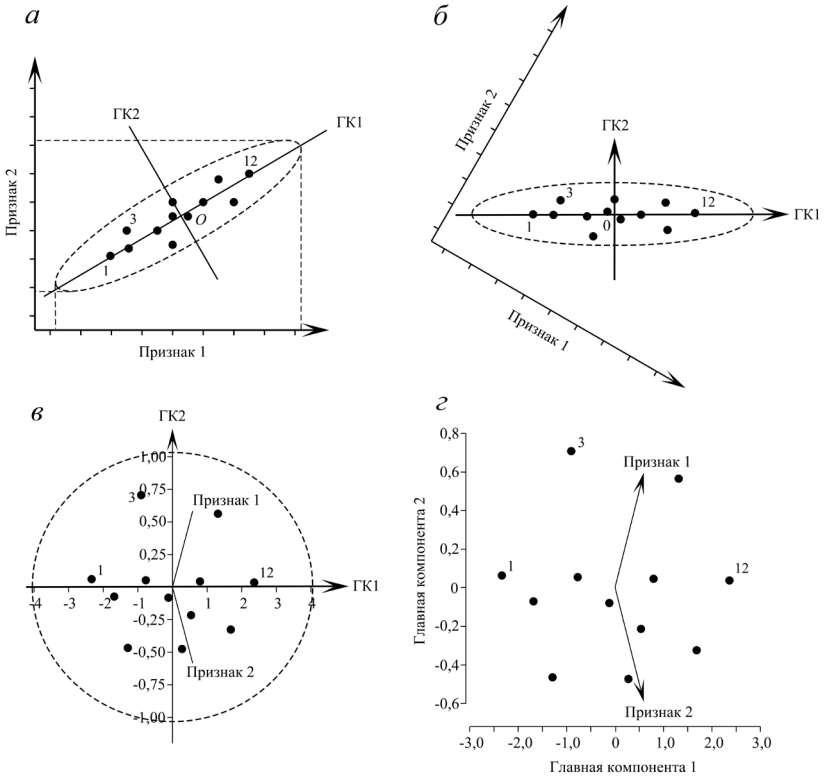


Рис. 16.2. Выделение главных компонент в корреляционном облаке двух признаков. Пояснения — в тексте

через максимальный диаметр корреляционного эллипса, а ГК 2 — через минимальный. Обратите внимание, что всё общее, что связывало два признака, вобрала в себя ГК 1, а необъяснённая общей связью изменчивость (математически — дисперсия) вошла в ГК 2. По проекциям (опущенный на оси пунктир) можно также видеть, что первый признак вкладывает в ГК 1 несколько больше, чем признак 2. Теперь повернём эту систему так, чтобы ГК 1 стала осью абсцисс, ГК 2 — осью ординат, а точка *O* — нулём, то есть началом новой системы координат (рис. 16.2, *б*). Это и будет система главных компонент, которая после нормировки на долю объясняемой дисперсии станет не эллипсом, а кругом. В нём откладывают два вектора признаков таким образом, что

по величине и знаку полученное значение соответствует корреляции признака с данной ГК (рис. 16.2, *в*). Обычно круг не изображается, а векторы признаков изображаются стрелками (рис. 16.2, *в*). Полученный график называется ***ординационной диаграммой***, а если на нём изображены одновременно и векторы признаков, и точки объектов, то он называется ***биплотом*** (*biplot*). Таким образом, из двух коррелирующих признаков мы получили две ГК, первая из которых содержит общее двух признаков, а вторая — оставшуюся необъяснённой изменчивость.

Аналогично мы можем рассмотреть взаимную корреляцию не двух, а трёх признаков. В пространстве трёх признаков получим корреляционный *эллипсоид*, который будет похож на батон хлеба. В нём мы сможем аналогично найти центр и провести три перпендикулярные друг другу оси: ГК 1 вберёт в себя максимум общей изменчивости (дисперсии) всех трёх признаков, ГК 2 — максимум оставшейся общей дисперсии, ГК3 — оставшуюся дисперсию.

Мы не можем представить себе пространство большей мерности, чем три, но математика там будет точно такая же. Так, в 100-мерном ***гиперпространстве*** наших 100 морфометрических признаков можно аналогично найти центр — ***центроид*** — и провести через него 100 перпендикулярных друг другу ГК, которые будут объяснять всё меньше и меньше общей дисперсии в системе. Польза от такого подхода состоит в том, что за совместным варьированием всех 100 признаков стоит не так много процессов. **Задание**. Как вы думаете, сколько и каких? Какие факторы будут в первую очередь обусловливать изменчивость размеров и формы? В действительности основных источников мало: 1) размерно-возрастная изменчивость; 2) половые различия; 3) конституциональные различия; 4) этнические различия. Эти четыре латентных фактора обусловливают почти целиком всю наблюдаемую нами изменчивость всех 100 признаков. АГК помогает установить количество таких факторов, а поэтому часто рассматривается как вариант ***факторного анализа*** или предшествует ему.

После интерпретации ГК мы должны суметь интерпретировать их. А далее можно вычислить значение ГК для каждого человека и использовать его в качестве нового обобщающего признака, то есть вместо 100 признаков мы получим только четыре

сложных индекса, но таких, которые вобрали в себя максимум информации из всех 100 признаков и по своей информативности не уступают им, а даже превосходят за счёт уменьшения шума случайной вариабельности. Таким образом, в ходе АГК происходит сокращение пространства признаков — ***редукция данных с обобщением***, это и есть цель анализа. В нашем примере мы спроецировали 100-мерное пространство признаков на четырёхмерное пространство компонент.

В статистических пакетах АГК находится обычно в модулях ***многомерного анализа*** (*multivariable, multivariate analysis*): либо в качестве самостоятельного метода (*principal component analysis, PCA*), либо в качестве варианта факторного анализа (*factor analysis*).

П р и м е р. Рассмотрим АГК на том же примере с гаплогруппами для экономии времени, поскольку в нашем случае использование корреляции Пирсона для данных, представленных в процентах, строго говоря, некорректно. Но поскольку мы не будем оценивать статистическую значимость корреляции Пирсона, а лишь используем её в качестве одного из способов найти новые оси в многомерном пространстве данных, чтобы разобраться в их структуре, — в этом нет грубой ошибки. Так часто делают в разведочном анализе, хотя нужно отдавать себе отчёт в том, что при работе с асимметрично распределёнными признаками и/или нелинейными корреляциями результаты АГК могут давать искажённую картину.

В пакете PAST

① Файл Гаплогруппы.dat открыт, данные выделены.

② Путь: Multivarite — Ordination — Principal components (PCA).

③ В открывшейся форме на закладке Summary выбираем Matrix — Correlation, то есть корреляционную матрицу, а не рассчитанную по умолчанию и популярную в экологии ковариационную матрицу (Variance-covariance) и нажимаем кнопку [Recompute] (Пересчитать).

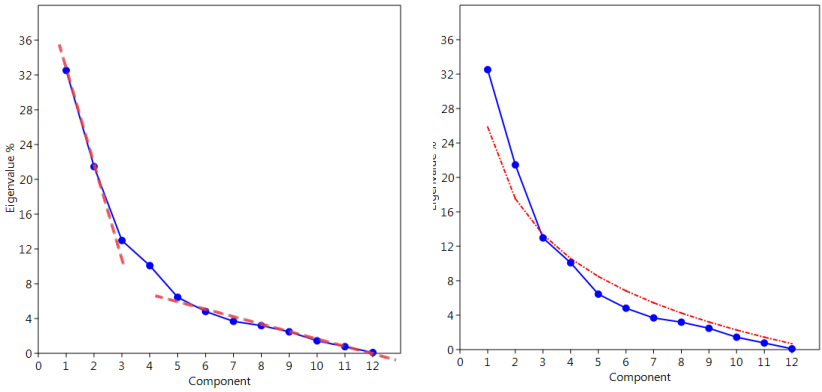
④ Закладка Summary. Видно, что для 12 признаков (гаплогрупп) вычислено 12 главных компонент (PC). Как уже упоминалось, АГК относится к подходам, основанным на вычислении собственных чисел, так как в ходе анализа было решено *характеристи-*

ческое уравнение двенадцатой степени и были получены 12 его корней, которые называются ***собственными числами*** (*eigenvalue*). Они пропорциональны доле объясняемой ГК дисперсии и дают в сумме число признаков (в нашем случае — 12). Поэтому легко вычислить долю объясняемой каждой ГК дисперсии в процентах — см. последнюю колонку — % variance. Видно, что ГК1 объясняла около трети всей изменчивости в данных (3,90279 : 12 = 0,3252325, или 32,5 %), а две первые ГК — более половины (32,5 + 21,5 = 54,0 %); то есть в первых двух компонентах содержится информации больше, чем в 6 признаках из 12. Далее распишем последовательно этапы анализа.

Этап 1. **Определение числа наиболее важных компонент**. Если АГК используется как вариант разведочного факторного анализа, то необходимо установить, сколько латентных факторов (процессов, явлений) определяет наблюдаемую изменчивость признаков, то есть отделить полезную информацию — «сигнал» от бесполезного «шума» данных. Способов предложено много, рассмотрим три:

а) ***критерий «каменистой осыпи Кэттелла»*** (*Cattell's scree test*). Закладка Scree plot. На этом графике, который напоминает склон горы с осыпью, представлены данные таблицы Summary, то есть доли объясняемой дисперсии для всех компонент, образующие убывающий ряд и соединённые ломаной линией. Визуально определяются две группы ГК: 1) расположенные «на горе» — они и есть самые важные, 2) расположенные «в осыпи» у подножья горы — это какая-то остаточная изменчивость, не требующая интерпретации. Изломы бывают чёткими, а бывают — нет. Можно мысленно провести прямые линии через точки ГК склона горы и через точки ГК осыпи (на рис. 16.3, *а* показаны пунктиром) — это поможет определиться. В нашем случае «на горе» бесспорно находятся ГК 1 и 2, а в осыпи — ГК 5 и последующие. Таким образом, критерий Кэттелла подсказывает важность выделения двух или четырёх компонент. Психологи обычно считают важной для интерпретации также первую компоненту осыпи, то есть они бы выбрали решение с тремя или пятью ГК;

б) ***критерий «сломанной трости»*** (*broken stick test*). Ставим галочку ☑ Broken stick в правой части формы с осыпью. Метод сравнивает распределение *n* ГК на графике осыпи с распределением обломков стержня такой же длины, сломанного в *n* – 1



а) *б)*

Рис. 16.3. Определение числа наиболее важных компонент на графике «каменистой осыпи»:
а) критерий Кэттелла, б) критерий «сломанной трости»

случайно выбранных точках. ГК, находящиеся над линией «сломанной трости», демонстрируют *нетривиальную дисперсию*, которую нельзя объяснить случайностью. В нашем случае красная (штрих-пунктирная) ломаная линия отсекает сверху ГК 1 и ГК 2, то есть указывает на решение с двумя компонентами. Тем не менее можно заметить, что она почти касается ГК 3 и 4, но однозначно отстоит от ГК 5. Таким образом, критерий «сломанной трости» также подсказывает нам решение с двумя (строго) или четырьмя (либерально) главными компонентами;

в) ***критерий Кайзера*** (правило Кайзера, *Kaiser's rule*). Согласно ему следует рассматривать только ГК с собственными числами (СЧ) более 1. Поскольку сумма СЧ чисел равна числу признаков (в нашем случае 12), в среднем приходится по одному признаку на СЧ. Поэтому главные компоненты с СЧ>1 содержат информации больше, чем один признак, с СЧ = 1 — столько же, сколько один признак, а ГК с СЧ<1 — меньше чем один признак. Таким образом, при СЧ ≤ 1, компоненты не являются обобщающими, а потому нет смысла их интерпретировать. Обычно этот критерий выделяет слишком много ГК, поэтому лучше сочетать его с другими. Возвращаемся на закладку Summary, смотрим столбец Eigenvalue. Видно, что для ГК 4 СЧ = 1,21, а для ГК 5 СЧ = 0,77 (< 1). Таким образом, критерий Кайзера подсказывает решение с 4 ГК.

По результатам проверки тремя методами мы получили: двумя методами — две или четыре ГК, и одним методом — четыре ГК. Чтобы не потерять важную информацию, окончательно остановимся на **необходимости и достаточности** выделения четырёх ГК.

ВАЖНО! Считается, что редукция данных с обобщением успешна, если выбранное количество ГК объясняет около 80 и более процентов дисперсии данных. В нашем случае первые четыре ГК объясняют 32,523 + 21,463 + 12,973 + 10,089 = 77,048 % дисперсии — это довольно хорошо.

Этап 2. **Выбор метода факторного анализа и вращение решения.** Строго говоря, АГК не является методом факторного анализа (см. теоретический материал), однако обычно является первым его этапом. На этапе 2 нужно решить, в рамках какой модели проводить дальнейший анализ: а) пытаться объяснить дисперсию данных меньшим числом компонент — собственно метод АГК или б) пытаться объяснить матрицу корреляций между показателями меньшим числом факторов — методы факторного анализа. После этого проводится анализ с выбранным на этапе 1 числом латентных переменных (в нашем случае — 4), которое нужно ввести в окошко программы вручную. После анализа результаты сохраняются в черновике и проводится ещё один такой же анализ, но с ***вращением*** (*rotation*). Процедура вращения накладывает на полученное решение ряд искусственных ограничений, которые на практике повышают интерпретируемость результатов. Для разведочного анализа данных хорошо использовать ***прямоугольные вращения***, предполагающие, что латентные факторы полностью ортогональны, то есть не коррелируют друг с другом. Обычно используют ***вращение «Варимакс»*** с нормализацией Кайзера (варимакс-вращение, *Varimax*, *Varimax with Kaiser Normalization*). К сожалению, пока в пакете PAST нет ни вариантов факторного анализа, ни возможности вращать полученное решение. Попробуем интерпретировать решение из четырёх факторов без вращения.

Этап 3. Интерпретация главных компонент (факторов) и их наименование. Закладка Loadings (Нагрузки). *Факторные нагрузки* представляют собой коэффициент корреляции Пирсона между признаком и ГК или фактором. Они показывают, с какой силой и каким знаком показатель «участвует» в латентной

переменной. Сохраняем таблицу нагрузок на черновике и доводим до окончательного вида. Для этого:

1) копируем таблицу в буфер — в нижней части формы кнопка [Copy];

2) вставляем в открытый лист Excel или аналогичную электронную таблицу;

3) удаляем ненужные колонки, оставляя только выбранное на этапе 1 число латентных переменных (в нашем примере — 4);

4) меняем десятичную точку на запятую: Выделить таблицу — Правка — Заменить — Найти [.], Заменить на [,] — Заменить всё — ОК;

5) округляем до тысячных: Формат — Ячейки — Закладка Число — Числовой — Число десятичных знаков [3] — ОК;

6) копируем таблицу в Excel и вставляем в Word или аналогичный текстовый редактор;

7) создаём последнюю строчку таблицы, куда вписываем долю объясняемой ГК дисперсии в процентах, округлённых до десятых;

8) выделяем высокие нагрузки жирным шрифтом или цветом. По абсолютной величине коэффициента корреляции мы интерпретировали связи $|r| < 0{,}3$ — как слабые, $0{,}3 \le |r| \le 0{,}7$ — как средней силы и $|r| > 0{,}7$ — как сильные; аналогично интерпретируем и нагрузки. Если анализ проводится на большом массиве данных (сотни объектов) для интерпретации полезно выделять нагрузки более 0,25, если до 100 объектов — 0,3–0,5 и выше, если менее 30 — 0,7 и выше.

В итоге для нашего примера получаем следующую таблицу факторных нагрузок.

Таблица 1 — Нагрузки 12 показателей на четыре главные компоненты

Гаплогруппа	Главная компонента			
	1	2	3	4
I1	–0,292	0,247	0,208	**0,311**
I2*/I2a	0,174	–0,249	**0,557**	–0,140
I2b	–0,163	**0,501**	0,108	0,104
R1a	–0,200	**–0,430**	0,044	0,178
R1b	–0,059	**0,531**	–0,146	–0,181
G	**0,361**	0,158	–0,098	**0,385**

Окончание таблицы

Гаплогруппа	Главная компонента			
	1	2	3	4
J2	**0,466**	0,031	–0,042	–0,031
J*/J1	**0,409**	0,041	–0,125	0,135
E1b1b	**0,340**	–0,018	0,267	**–0,424**
T	**0,364**	–0,052	**–0,317**	**0,339**
Q	0,044	–0,004	**0,505**	**0,576**
N	–0,223	**–0,358**	**–0,398**	0,122
Доля объясняемой дисперсии, %	32,5	21,5	13,0	10,1

Примечание: жирным шрифтом выделены нагрузки ≥0,3.

Эту же информацию можно отобразить графически, если перейти на закладку Loadings plot (График нагрузок) и выбирать в окошке последовательно интересующие компоненты.

Интерпретация латентных переменных — наиболее сложный процесс, поскольку требует глубоких знаний в предметной области. Однако сам принцип интерпретации прост: исследователь пытается объяснить, по какой причине показатели, выделенные жирным шрифтом, изменяются согласованно: с одинаковым знаком — коррелируют между собой положительно, с разными знаками — отрицательно. Так, из табл. 1 видно, что ГК 1 обусловлена согласованным изменением доли обладателей гаплогрупп G, J2, J*/J1, E1b1b и T. Мы не являемся специалистами в генетике человека, но даже беглое ознакомление со сведениями из Википедии позволяет ответить на вопрос, что удерживает данные гаплогруппы вместе: они наиболее распространены на Ближнем и Среднем Востоке, в Северной и Восточной Африке и на Кавказе. Таким образом, данный набор (***паттерн***) гаплогрупп по отношению к Европе можно обозначить как «Южные гаплогруппы». Латентный фактор, проявившийся в ГК 2, отчётливо ***биполярный***: чем больше у народов доля гаплогрупп I2b и R1b, тем меньше доля R1a и N. И наоборот: знак в АГК произволен, и мы можем умножить столбец нагрузок на (–1) для удобства интерпретации. Анализ карт распределения данных гаплогрупп показывает, что ГК 2 маркирует направление с северо-востока

(R1a и N) на юго-запад Европы, потому её можно обозначить как «Клинальная изменчивость гаплогрупп в направлении СВ–ЮЗ». Аналогично анализируются и называются все существенные ГК. Если чёткой интерпретации и/или информационно ёмких названий не получается, то можно ограничиться рабочими названиями ГК или факторов, данными по показателю с наибольшим вкладом (нагрузкой).

Этап 4. Построение ординационной диаграммы. Данный этап можно совмещать с этапом 3 для облегчения интерпретации латентных переменных. Закладка Scatter plot (Диаграмма рассеяния). Ставим галочки в ☑ Row labels (Названия строк) и ☑ Biplot (совмещённый график, или ***биплот***). На биплоте совмещены как точки объектов, так и векторы самих переменных в пространстве выбранных ГК (рис. 16.4). Проекции этих векторов показателей на ГК пропорциональны их нагрузкам на ГК. Видно, что наибольший положительный вклад в ГК 1 дают гаплогруппы J2, J*/J1, G, E1b1b и T, а с отрицательным знаком в ней участвуют показатели, находящиеся по другую сторону от нуля, особенно: гаплогруппы I1 и N. Анализ проводится так: смотрится, какие показатели и объекты находятся по одну сторону от нуля для данной ГК, а какие — по разные стороны. Так, видно, что гаплогруппы J2, J*/J1, G, E1b1b и T удерживает вместе то, что они наиболее часто встречаются совместно у населения Кипра, Турецкой Анатолии, Мальты и Греции. Напротив, данные гаплогруппы крайне редки у российских мордвинов, жителей Норвегии, Швеции, Финляндии, Латвии — эти страны/народы занимают противоположный отрицательный полюс ГК; но у последних выше доля гаплогрупп I1 и N. **Задание**. Определите по ординационной диаграмме наиболее типичные для жителей России и Ирландии гаплогруппы и проверьте правильность вывода по исходной таблице частот гаплогрупп (с. 229). Действительно, получается, что у жителей России преобладают гаплогруппы R1a (46 %) и N (23 %), а у жителей Ирландии — R1b (81 %), I2b (5,0 %) и I1 (6,0 %).

Если снять галочку ☐ Biplot и поставить в ☑ Min. spanning tree, то программа отобразит ***минимальное остовное дерево***, которое соединит ближайшие объекты многомерного пространства признаков в проекции на выбранные ГК. **Задание**. Население какой страны наиболее близко России, если судить по минимальному остовному дереву в пространстве двух первых ГК?

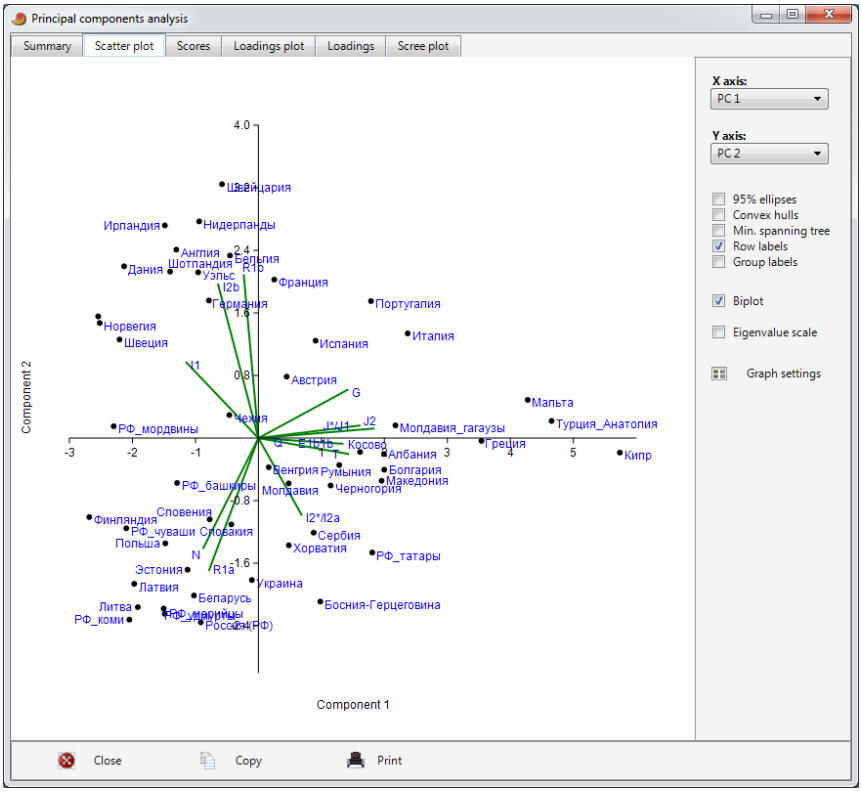


Рис. 16.4. Ординационная диаграмма результатов анализа главных компонент в пакете PAST

Ординационную диаграмму пред вставкой в работу необходимо доработать: перевести названия осей, дополнить долей объясняемой дисперсии, изменить весь шрифт на Times New Roman, подобрать кегль шрифта таким образом, чтобы в окончательном документе его размер составлял 0,8–1 размера шрифта основного текста. Частично это можно сделать в PAST (в правой части формы с рисунком — Graph settings), но лучше сохранить правленый рисунок в формате *.svg и доработать окончательно в векторном графическом редакторе (TpX, Inkscape, Corel Draw и т. п.). Там же можно снабдить векторы признаков стрелочками на концах, а накладывающиеся друг на друга названия раздвинуть в пространстве для удобства чтения графика (рис. 16.5).

Этап 5. Расчёт индивидуальных значений латентных переменных. В некоторых задачах важно не только разобраться в структуре большого массива данных, но и рассчитать значения латентных переменных для всех интересующих объектов. Такие индивидуальные ***факторные метки*** *(factor scores)* находятся в таблице на закладке Scores (Метки или Значения компонент). Эти данные можно скопировать в буфер (Copy), вставить и сохранить в новом файле, а далее работать с ними как с новыми, *обобщающими переменными*: смотреть распределение, сравнивать по ним группы, проводить кластерный анализ. Если у нас есть данные о внешних факторах, которые могли повлиять на использованные в анализе показатели, то проводится корреляционный или регрессионный анализ меток с такими факторами. В экологических исследованиях это часто помогает выявить изменчивость, привносимую экологическими факторами: градиентами температур, глубин, расстояний от источников загрязнения и т. п. Таким образом, АГК позволяет выявить влияние таких факторов косвенно, не включая их в анализ — именно поэтому такие методы называют методами непрямого градиентного анализа.

Оформление в квалификационной работе или статье

1. Статистическая часть раздела «Материал и методы».

Для выявления наиболее общих закономерностей распределения гаплогрупп по странам Европы данные были подвергнуты анализу главных компонент. При этом в качестве меры близости показателей использовали корреляцию Пирсона, а для выявления числа наиболее важных компонент, необходимых и достаточных для описания данных, применяли критерии Кэттелла, Кайзера и «сломанной трости». Расчёты и графические построения выполнены в пакете PAST (version 3.11, Hammer et al., 2001).

2. Раздел «Результаты и обсуждение».

Это самый большой раздел, поэтому, если объём публикации позволяет, можно дать достаточно подробное описание:

1) привести доработанную диаграмму осыпи — для подтверждения вывода о количестве необходимых и достаточных для интерпретации ГК;

2) привести таблицу факторных нагрузок и проинтерпретировать ГК. При интерпретации желательно ссылаться на работы предшественников, которые подтверждают и/или объясняют

наблюдаемое число латентных факторов и корреляционную структуру показателей в каждой ГК;

3) привести доработанную ординационную диаграмму с двумя наиболее интересными ГК. Часто это ГК 1 и 2 (рис. 16.5), но необязательно. Данный рисунок лишь частично дублирует данные таблицы нагрузок, но позволяет более детально интерпретировать данные.

3. Раздел «Выводы».

В ходе анализа главных компонент были выделены четыре латентные переменные, объясняющие в сумме 77,0 % общей дисперсии. Далее нужно привести их названия, полученные в ходе интерпретации ГК на этапе 3.

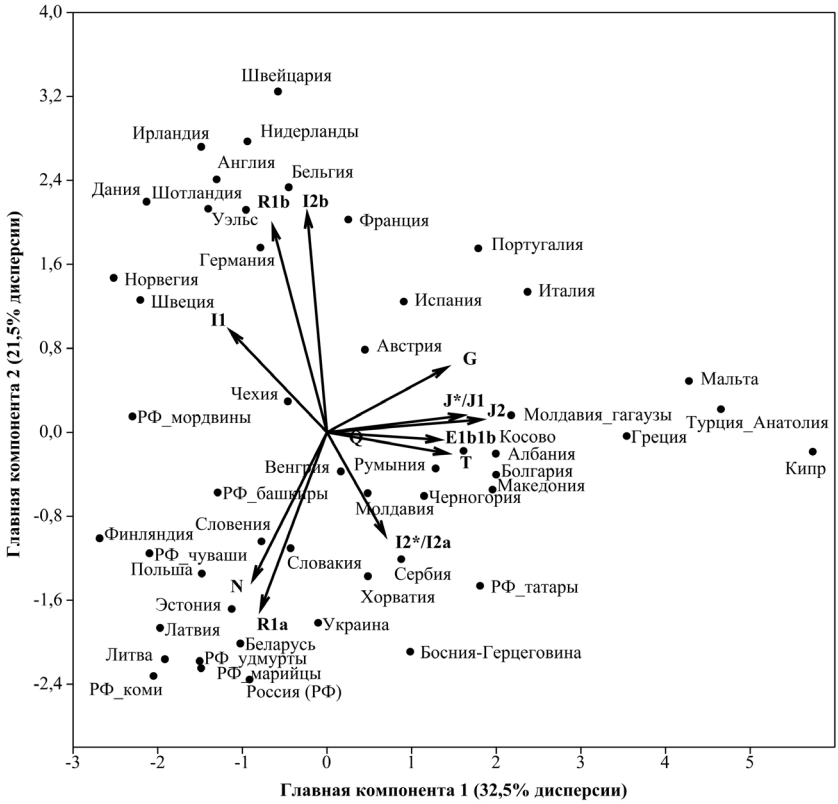


Рис. 16.5. Страны и народы Европы в пространстве двух первых главных компонент гаплогрупп Y-хромосомы

* * *

Проделанная работа показывает, что анализ главных компонент является более информативным методом по сравнению с кластерным анализом: удалось не только определить взаимную близость объектов в многомерном пространстве, но и объяснить такое расположение. Единственным возможным недостатком такого подхода является то, что он основан на корреляции Пирсона и *многомерном нормальном распределении*: чем сильнее отклоняются данные от этой модели, тем менее правдоподобными становятся и результаты анализа. Поэтому АГК желательно проводить на предварительно преобразованных данных, например методом Бокса — Кокса; такое преобразование хотя и не приведёт к многомерной нормальности, но по крайней мере устранит аберрации анализа, вызванные эффектом шкалы. Другим возможным подходом к редукции данных с обобщением является анализ главных координат, в котором могут быть использованы любые меры сходства.

2.2. Анализ главных координат

Анализ главных координат, АГКО (*principal coordinate analysis, PCoA, PCO*), или ***классическое многомерное шкалирование*** (*classical multidimensional scaling, CMDS*) — другая ординационная техника, также относящаяся к методам непрямого градиентного анализа. В отличие от АГК данный подход изначально был основан не на вычислении собственных чисел, а на *вычислении расстояний между объектами* (*eigenanalysis-based approache*), хотя современные алгоритмы реализации этого метода могут включать в себя расчёт собственных чисел. Соответственно, в АГКО анализируется не матрица сходства признаков, а матрица расстояний между самими объектами. При этом в качестве расстояний могут выступать произвольные меры, например: расстояния Махаланобиса, величины $(1 - r)$ корреляций Пирсона, Спирмена, Кендалла, $(1 - I)$ индексов сходства Жаккара, Съёренсена и др. (см. теоретический материл [8]). Цель анализа заключается в проекции объектов из многомерного пространства в пространство меньшей размерности с максимально возможным сохранением расстояний между объектами. Если в АГКО использовать обычные евклидовые расстояния между объектами, то его результаты будут пропорциональны результатам

АГК, проведённому по матрице ковариаций. В целом, несмотря на иную философию метода, интерпретация главных координат проводится аналогично анализу главных компонент.

Проведём этот анализ на тех же данных по гаплогруппам, но будем использовать в качестве меры сходства корреляцию Спирмена.

В пакете PAST

① Файл Гаплогруппы.dat открыт, данные выделены.

② Путь: Multivarite — Ordination — Principal coordinates (PCoA).

③ В открывшейся форме на закладке Summary выбираем Similarity index — Rho, то есть используем в качестве меры сходства коэффициент корреляции Спирмена ρ («ро») и нажимаем [Recompute] (Пересчитать).

④ Закладка Summary. Видно, что для 12 признаков (гаплогрупп) вычислено 12 главных координат. Определиться, какие из них самые важные, можно аналогично АГК. В этом модуле пакет не строит график «каменистой осыпи» и «сломанной трости», поэтому можно ориентироваться по величине собственных значений более 1 — получается три главных координаты, причём первые две из них объясняют 41,182 + 17,534 = 58,716 % общей изменчивости данных. Таким образом, если мы будем рассматривать только две первые ***размерности*** (главные координаты, оси), то объясним более чем половину всех расстояний между странами Европы по соотношению гаплогрупп. Представим это графически (рис. 16.6).

⑤ Закладка Scatter plot (Диаграмма рассеяния). Ставим галочку в ☑ Row labels (Названия строк). Поскольку для АГКО важны расстояния между объектами, в этом анализе минимальное остовное дерево представляется более уместным, поэтому можно поставить также галочку в ☑ Min. spanning tree для его отображения. Полученную диаграмму нужно доработать в модуле Graph settings и сохранить в формате *.svg для окончательной правки в векторном редакторе.

⑥ Интерпретация главных координат. Проводится аналогично АГК. Если вы хорошо знакомы с географией, то можете легко увидеть, что вдоль ГК 1 проявились различия в направлении с юга на север, а вдоль ГК 2 — в направлении с запада на восток. Таким образом, можно говорить о том, что различия между

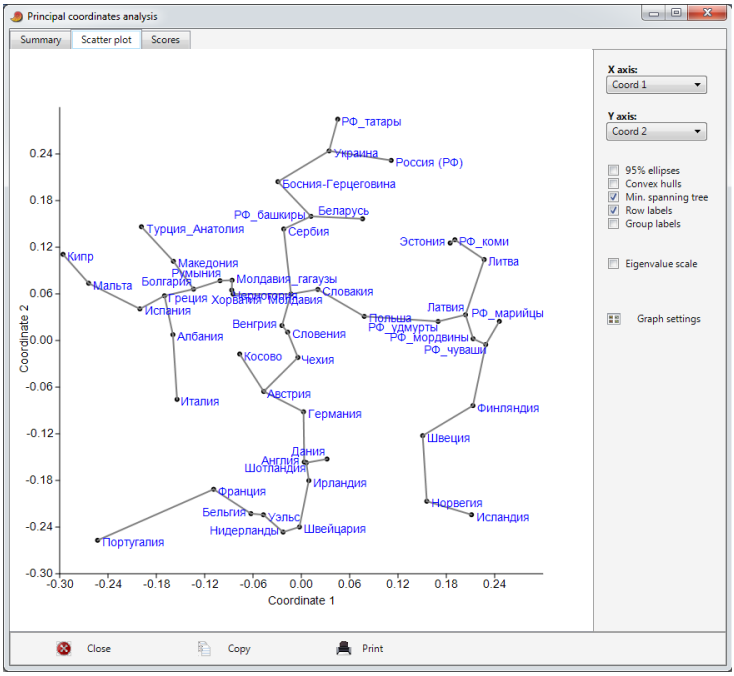


Рис. 16.6. Ординационная диаграмма результатов анализа главных координат в пакете PAST

странами и народами Европы по соотношению 12 гаплогрупп примерно на 60 % можно объяснить географическим расположением. Саму ординационную диаграмму желательно развернуть так, чтобы направления совпали с географической картой (рис. 16.7) — в данном частном случае это можно сделать, поменяв значения Coord 1 и Coord 2 местами в правом верхнем углу формы с диаграммой (рис. 16.6), а в общем случае это делается отображением и вращением диаграммы и надписей в векторном редакторе.

⑦ **Оформление в квалификационной работе или статье**.

7.1. Статистическая часть раздела «Материал и методы».

Для выявления наиболее общих закономерностей распределения гаплогрупп по странам Европы данные были подвергнуты анализу главных координат с использованием в качестве меры близости показателей корреляции Спирмена. Для визуализации расстояний между странами в проекции на главные координа-

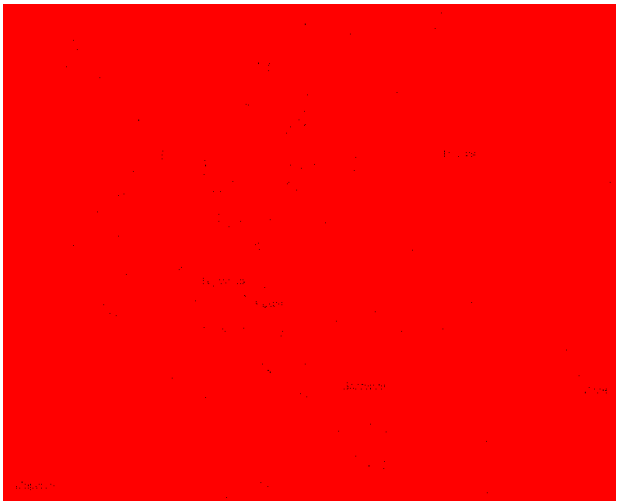


Рис. 16.7. Карта стран Европы

ты использовали построение минимального остовного дерева. Расчёты и графические построения выполнены в пакете PAST (version 3.11, Hammer et al., 2001).

7.2. Раздел «Результаты и обсуждение».

Обычно достаточно привести ординационную диаграмму, сопроводив её интерпретацией главных координат и обсуждением (рис. 16.8).

В обсуждении можно затронуть не только самые общие моменты, но и интересующую специфику. Например, может представлять интерес, что наиболее близкими к усреднённым данным по Российской Федерации на минимальном остовном дереве были российские татары и украинцы, или что Португалия отличается высокой генетической спецификой и оказалась ближе не к Испании, а к Франции и другим странам северо-запада Европы, омываемым Атлантическим океаном, или что российские коми генетически мало отличимы от эстонцев и т. п. Если объём публикации позволяет, можно привести и график осыпи для наглядного подтверждения выбранного для интерпретации числа координат, однако его придётся строить вручную. Для этого на закладке Summary нужно скопировать данные (Copy), вставить их в пустую область открытого окна пакета или, создав новый файл, заменить десятичные точки запятыми и строить график осыпи,

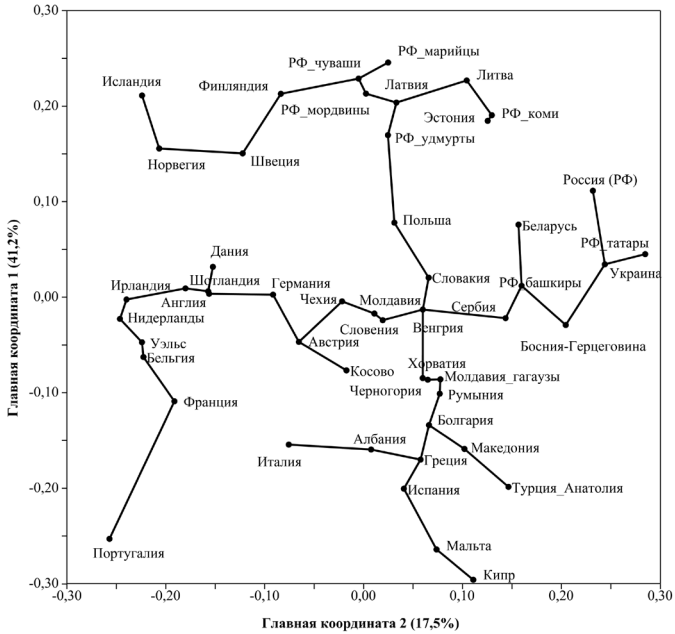


Рис. 16.8. Страны и народы в пространстве двух первых главных координат гаплогрупп Y-хросомомы. Расстояние — корреляции Спирмена, линии — минимальное остовное дерево

используя любую из двух колонок по желанию, — смысл есть как в приведении собственного числа (Eigenvalue), так и приведении доли объясняемой координатой дисперсии в процентах (Percent).

7.3. Раздел «Выводы».

Различия между странами и народами Европы по соотношению 12 гаплогрупп примерно на 60 % объяснялись географическим расположением. Наиболее близкими народами к … были: …, …, … (или иная специфика).

ЛАБОРАТОРНАЯ РАБОТА № 17

Многомерные методы разведочного анализа данных для качественных признаков

Тема 14. Многомерные методы разведочного анализа данных.
Количество часов: 2.
Цель: Познакомиться с методами анализа соответствий и канонического анализа соответствий.

Классификация направлений разведочного анализа данных, приведённая на предыдущем занятии, справедлива для любых данных, в том числе сочетающих сразу несколько шкал: 1) отношений; 2) интервалов; 3) рангов; 4) наименований. На этом занятии мы познакомимся с особенностями методов кластерного анализа и ординационных техник, применяемых для анализа качественных номинальных признаков (шкала наименований).

В отличие от шкал количественных признаков, всегда содержащих только результаты измерений, данные по номинальным признакам могут быть организованы в файле двумя способами.

1. Частотная форма. В таком виде в ячейках таблицы находятся абсолютные частоты (в штуках, единицах) категорий признаков — как в таблицах сопряжённости. Например, если в строках находятся разные биотопы, а в колонках — виды встречающихся в них организмов, то в ячейках будут находиться количества обнаруженных в данном биотопе организмов: от нуля до максимального. Для данных, представленных в частотной форме, удобной мерой связи являются корреляции, особенно Спирмена (в пакете PAST — Rho); она хорошо подходит для анализа соотношения частот.

2. Бинарная форма. В ячейках таблицы находятся только нули и единицы: 0 — если признак в данной ячейке отсутствовал; 1 — если признак в данной ячейке имелся.

Для бинарных данных особенно полезны ***индексы сходства Жаккара*** (в пакете PAST — Jaccard), ***Съёренсена*** (синонимы: Дайса, Чекановского — Съёренсена; в пакете PAST — Dice) и ***Раупа — Крика*** (в пакете PAST — Raup-Crick). Поясним кратко,

чем хороши индексы Жаккара и Съёренсена и в чём их отличие от индекса Раупа — Крика.

Предположим, мы регистрируем наличие (1) или отсутствие (0) видов растений на нескольких участках или видов микроорганизмов в биоматериале нескольких пациентов. Нас интересует обнаружение ассоциаций таких видов (микро)организмов. Для любых двух видов ситуацию можно представить в виде следующей таблицы частот:

		Вид 2	
		1	0
Вид 1	1	11	10
	0	01	00

Коэффициенты сходства Жаккара и Съёренсена не учитывают ячейку 00, поскольку предполагают, что отсутствие обоих видов нельзя считать показателем сходства. Иначе наиболее сходными могут оказаться виды, которые в большинстве местообитаний вообще не встречались. Так, индекс Жаккара C_J будет рассчитываться как отношение числа местообитаний, где встречались оба вида, к числу местообитаний, где встречался хотя бы один вид: $C_J = \frac{11}{(11+10+01)}$. Индекс Съёренсена похож на него и линейно с ним связан (см. теоретический материал).

В отличие от данных индексов и распространённых коэффициентов ассоциации (фи, Юла, Крамера и др.), индекс Раупа — Крика является вероятностной мерой сходства, который вычисляется с использованием рандомизационной процедуры Монте-Карло. Поэтому все его значения $\geq 0{,}95$ являются статистически значимыми на 5%-ном уровне ($P \leq 0{,}05$). Данный индекс в неявном виде задействует в анализе и ячейку 00.

> **Комментарий**. Для любой пары качественных признаков индекс Раупа — Крика представляет собой отношение наблюдаемого числа случаев совместного присутствия признаков у объектов выборки к отношению числа таких случаев в перемешанных случайным образом данных. Алгоритм рандомизационной процедуры заключается в том, что данные по первому признаку остаются неизменными, а по второму признаку перемешиваются случайным образом между строками (местообитания, пациенты). В результате цифровые значения данных остаются такими же, как в исходной выборке, однако корреляционная связь между двумя признаками разру-

шается. После вычисления индекса сходства $C_{R\text{-}C}$, процедура рандомизации повторяется и $C_{R\text{-}C}$ вычисляется заново. Поскольку перемешивание значений происходит случайным образом, полученный во втором случае индекс будет немного отличаться от первого. Операция рандомизации с вычислением $C_{R\text{-}C}$ повторяется многократно (в PAST v 3.19 — 10 000 раз) и окончательно индекс сходства рассчитывается как отношение числа совместной встречаемости видов в реальных данных к таковой в усреднённых рандомизированных данных.

К сведению. В пакете PAST таблицы парных значений индексов сходства можно получить по пути: Multivariate — Similarity and distance indices.

Познакомимся с рядом многомерных техник анализа номинальных данных.

1. Кластерный анализ

2. Анализ главных координат

Все основные положения этих методов анализа, рассмотренные на предыдущем занятии для количественных признаков, справедливы и для случая качественных признаков. Единственным отличием является возможность использования специфических для номинальных данных мер связи (см. выше).

3. Анализ соответствий

Анализ соответствий (*correspondence analysis, CA*) — многомерная ординационная статистическая техника, которая может рассматриваться в качестве аналога метода главных компонент для категориальных данных. Классический АС был предложен Хёршфилдом в 1935 г., развит Бензекри́ в 1960–1970-х гг. и был популяризован французскими исследователями, хотя сходные методы разрабатывались независимо другими авторами из разных стран (оптимальное шкалирование, оптимальная оцифровка, взаимное усреднение, квантификационный метод, анализ однородности). Традиционно метод применяется для анализа многовходовых таблиц сопряжённости и разлагает на ортогональные факторы статистику хи-квадрат. Строки и столбцы матрицы данных обрабатываются в ходе анализа аналогично. На анализ варьирования показателей не накладывается никаких ограничений, то есть АС — *неограниченная* (*unconstrained*) техника, или *метод непрямого градиентного анализа*.

П р и м е р. В травматологическом отделении клиники регистрировались виды травм и отмечался пол пациента. Также определялась группа крови, поскольку пациентам с тяжёлыми

травмами могла потребоваться гемотрансфузионная терапия. Данные в частотной форме находятся в файле «Травмы.xls» по адресу: https://yadi.sk/d/g50i73pt3J6pAa.

Задание: выявить ассоциации различных травм, используя анализ соответствий.

> **Комментарий**. В данном варианте анализа можно не обращать внимания на другие показатели в наборе (пол, группы крови). Ассоциации травм будут находиться без явного учёта таких индикаторных показателей, хотя неявно они будут влиять на результаты, поскольку группировка материала проводилась именно по ним.

В пакете PAST

① Откройте файл «Травмы.xls», перенесите данные в PAST, сохраните и выделите только ту часть, которая относится к травмам.

② Путь: Multivariate — Ordination — Correspondence (CA).

③ Закладка Summary. Как и в анализе главных компонент (АГК), здесь представлены ***собственные числа*** (*eigenvalue*) и объясняемая ими доля изменчивости, которая в АС традиционно называется ***инерцией*** (*inertion*). Видно, что первая ***ось*** *(axis)* анализа соответствий объясняла 57,4 % инерции, вторая — 18,9 %. В отличие от АГК, уменьшение доли объясняемой инерции в АС носит, как правило, плавный характер, поэтому явных изломов на линии каменистой осыпи не бывает. Поэтому при выборе числа обсуждаемых в исследовании осей следует ориентироваться на их интерпретируемость.

④ Рассмотрим ординационную диаграмму первых двух осей: закладка Scatter plot. Поскольку мы ищем ассоциации видов травм, которые представлены в колонках, выставляем [Biplot scaling]: Column principal. Оставляем галочки в ☑ Plot columns и ☑ Labels. Дорабатываем рисунок в [Graph settings].

Из рис. 17.1 видно, первая ось разделила травмы, полученные, вероятно, в результате конфликтов (левая отрицательная зона) от прочих (правая положительная зона). Вторая ось позволила выделить порезы и резаные раны преимущественно рук. В первой четверти координатной плоскости находится сборная группа травм, имеющих, вероятно, бытовой характер.

Если анализ соответствий позволяет обнаружить естественным образом группирующиеся объекты, то их можно пробовать сгруппировать и в таблице сопряжённости, которая в результате уменьшится.

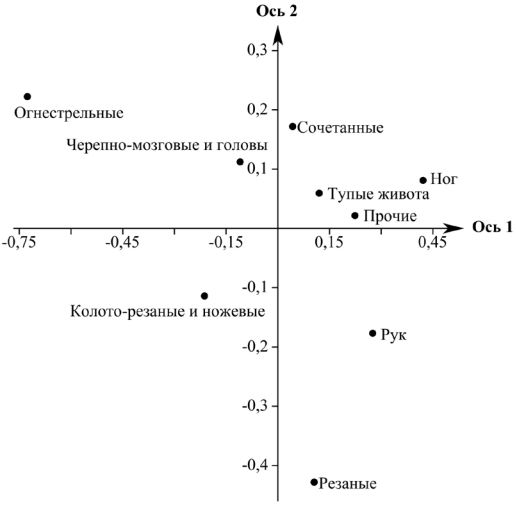


Рис. 17.1. Виды травм в пространстве двух первых осей соответствий

4. Канонический анализ соответствий (КАС)

Канонический анализ соответствий (*canonical correspondence analysis, CCA*) — ограниченная многомерная техника, или метод прямого градиентного анализа. Как отмечалось в лабораторной работе № 16, данные для таких техник состоят из двух разнокачественных блоков: блока независимых переменных (регрессоров) и блока зависимых переменных (откликов). При этом анализируется не вся изменчивость в наборах, а только та часть изменчивости откликов, которая объясняется регрессорами. Следовательно, ограниченные техники могут рассматриваться в качестве многомерного варианта регрессии.

В качестве откликов в КАС анализируются данные качественных признаков в частотной или бинарной форме. В качестве независимых регрессоров выступают количественные или качественные данные. Поскольку метод нашёл очень широкое применение в экологических исследованиях, такие регрессоры в программах часто называют ***средовыми переменными*** (*environmental variables*): классическое применение КАС — объяснить несколькими средовыми переменными (влажность, температура, pH, тип почвы и т. п.) данные по численности большого числа видов.

Если средовые переменные являются не количественными, а качественными, то из них формируют так называемые ***фиктивные переменные*** (*dummy variables*). Такие индикаторные переменные состоят только из нолей (0 — отсутствие данной категории в строке) и единиц (1 — наличие категории в строке).

П р и м е р. Рассмотрим данный вариант анализа на том же примере с травмами. Только теперь задействуем также и регрессоры: пол и группы крови. Постараемся объяснить только ту часть изменчивости, которая связана с этими показателями.

> **Комментарий 1**. Файл должен быть организован таким образом, чтобы колонки с регрессорами (средовыми переменными) предшествовали колонкам зависимых переменных.

> **Комментарий 2**. Включение в анализ групп крови в качестве регрессора требует объяснения. Для экономии места отправляем читателей к нашей работе: Нохрин Д. Ю., Рязанова Л. А., Тишевская Т. В. Группы крови и характер: виктимологический подход с анализом статистики травматизма // Вестник Челябинского государственного университета. 2015. № 21 (376). Биология. Вып. 3. С. 128–137.

В пакете PAST

① Файл «Травмы.dat» открыт, выделена область всех данных. Обратите внимание на то, как организованы фиктивные переменные «пол» и «группа крови».

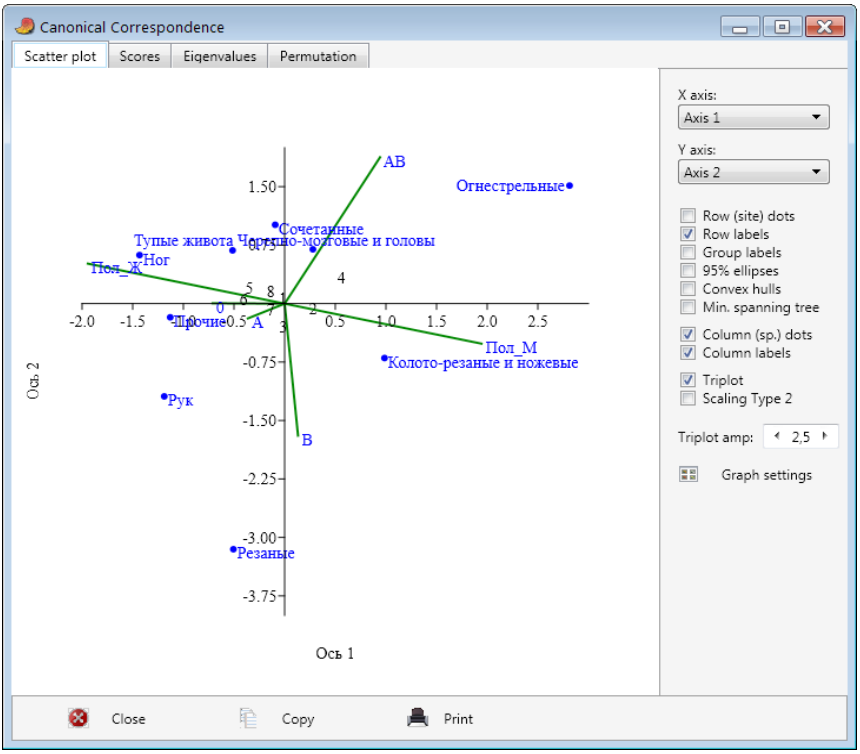
② Путь: Multivariate — Ordination — Canonical correspondence (CA).

③ В появившемся окне вводим число регрессоров (No. of environ. vars). В нашем случае — 6 (два пола и четыре группы крови).

④ Закладка Eigenvalues. Смотрим доли инерции, связанные с набором регрессоров. К сожалению, текущая версия PAST (3.19) не выдаёт долю инерции, объясняемую регрессорами, в общей инерции. Поэтому пока неясно, насколько хорошо регрессоры объясняют изменчивость показателей основного набора. Тем не менее качественно это можно сделать по оценке статистической значимости рандомизационной техникой в Permutation. Для этого нажмите [Compute] и посмотрите значимость первой оси. Если $P \leq 0{,}05$, имеет смысл смотреть значимость следующей оси, затем следующей и т. д. В нашем случае для первой оси $P = 0{,}307$, поэтому оценивать значимость последующих осей нет смысла, даже если в них $P < 0{,}05$ (это аберрации, свойственные рандомизационной технике в данном случае).

⑤ Закладка Scatter plot.

Обязательно нужно поставить галочку в ☑ Triplot, чтобы на ординационной диаграмме стали отображаться векторы регрессоров. В окне Triplot amp. можно ввести коэффициент, на который будут умножены векторы регрессоров при отображении на диаграмме: это нарушает математическую точность изображения, но повышает наглядность для удобства интерпретации.



⑥ Интерпретация. Интерпретация осей проводится по проекциям векторов, максимальных для данной оси. Так, в нашем примере видно, что ось 1 вобрала в себя половые различия: пациентам мужского пола были более свойственны огнестрельные, колото-резаные и ножевые ранения, тогда как пациентам женского пола — травмы ног, рук и прочие. Вдоль оси 2 проявились различия между обладателями групп крови AB и B: для первых были несколько более характерны огнестрельные ранения, а для вторых — резаные раны.

На основании интерпретации ординационной диаграммы выдвигаются гипотезы о структуре связей и зависимостей в наборе данных. Для подтверждения наиболее интересных закономерностей возможна свёртка большой таблицы сопряжённости исходных данных в таблицу 2×2 и расчёт относительных рисков. Так, в нашем случае логично свернуть таблицу по категориям «Пол» (мужской и женский) и «Конфликтные травмы» (огнестрельные, колото-резаные и ножевые ранения против всех остальных):

Травмы / Пол	Конфликтные	Прочие
Мужчины	186	346
Женщины	27	141

Далее методами лабораторной работы № 6 можно рассчитать риски и отношения шансов. Относительный риск получения «Конфликтных травм» был выше у мужчин: $RR = 2{,}18$ (95% ДИ от 1,51 до 3,13), $P = 2{,}9 \times 10^{-5}$.

⑦ **Оформление в квалификационной работе или статье (вариант)**.

7.1. Статистическая часть раздела «Материал и методы».

Для выявления ассоциаций девяти видов травм, а также их зависимости от пола и группы крови системы АВ0 был использован соответственно анализ соответствий и канонический анализ соответствий. Расчёты и графические построения выполнены в пакете PAST (version 3.19, Hammer et al., 2001).

7.2. Раздел «Результаты и их обсуждение».

В разделе приводятся ординационные диаграммы. В ходе описания указываются доли объясняемой выделенными осями инерции. Даётся интерпретация осей. Возможен переход к другим методам анализа для статистического подкрепления выявленных на графике закономерностей.

7.3. Раздел «Выводы».

В ходе анализа соответствий было установлено, что виды травм могут быть сгруппированы в три категории: 1) огнестрельные, колото-резаные и ножевые ранения; 2) порезы, колото-резаные и ножевые ранения и 3) бытовые травмы. Канонический анализ соответствий показал, что первая категория травм была свойственна преимущественно мужчинам: относительный риск $RR = 2{,}18$, $P < 0{,}001$, и т. д.

ЛАБОРАТОРНАЯ РАБОТА № 18

Планирование научного исследования

Тема 15. Планирование научного исследования.
Количество часов: 2.
Цель: Научиться использовать литературные источники и данные пилотных экспериментов для расчётов объёмов выборок. Проведение рандомизации. Работа на ПК, решение задач.

Существует большое число самых разных способов проведения натурных исследований и лабораторных экспериментов, которые призваны решать большой спектр различных задач. Потому невозможно дать универсальное пошаговое руководство по планированию качественного научного исследования — такого, которое содержит минимум ошибок измерения и обеспечивает максимальную уверенность в результате. Однако знание основных принципов планирования позволит для каждого конкретного случая разрабатывать оптимальную схему исследования. На этом занятии мы подытожим знания об уже известных нам принципах и познакомимся с новыми.

I. Принципы планирования научного исследования

Принцип 1. Знание способов статистического анализа данных

При планировании научного исследования у исследователя должен быть чёткий, но одновременно гибкий план последующего статистического анализа полученных данных. Мы неслучайно рассматриваем тему планирования на последнем занятии, а не на первом: не зная, как можно обработать данные, не познакомившись с примерами грамотных исследований, вы бы отнеслись к данной теме, как к абстрактной теории. Сейчас вы знаете основные задачи научного исследования, свойства выборок и получаемых на их основе оценок параметров генеральной совокупности, знаете типы данных, умеете оценить характер распределения признаков и выбрать для анализа параметрические или непараметрические методы, имеете представление

о мощности статистических критериев и т. д. Этого вполне достаточно, чтобы самостоятельно спланировать исследование. Возможно, что-то будет сделано по аналогии с рассмотренными нами примерами или примерами из учебников, что-то не будет самым оптимальным, однако основа для планирования у вас уже имеется. **Задание**: перечислите все рассмотренные нами 7 задач исследования и назовите известные вам методы их решения:

- для количественных признаков с нормальным распределением;
- количественных признаков с ненормальным распределением;
- порядковых признаков;
- качественных упорядоченных категориальных признаков;
- качественных номинальных признаков.

(Можно пользоваться тетрадями для лекционных и практических занятий или записями, сделанными по мере самостоятельного прохождения курса).

Принцип 2. Организация статистических повторностей

► **Повторность** (*replication*) в статистике — это повторение экспериментального измерения в одинаковых условиях. Если повторности осуществляются только в отношении одного и того же субъекта — ***внутри субъектов*** (*within subjects*), то источником изменчивости будет сам процесс измерения (точность методики, прибора и т. п.). Если же в повторных измерениях участвуют разные субъекты, то любые вариации ***между субъектами*** (*between subjects*) будут обусловлены как изменчивостью процесса измерения, так и изменчивостью самих субъектов. Организация повторностей позволяет нам оценивать эту изменчивость, разбивать её на составляющие компоненты и таким образом бороться с неопределённостью. Как вы помните, и оценка различий между двумя группами (например, критерием Стьюдента), и между несколькими группами (например, дисперсионным анализом) осуществлялись на основе соотнесения межгрупповой и внутригрупповой изменчивости. Вообще говоря, все статистические методы требуют выборки, а элементы выборки в отношении генеральной совокупности представляют собой не что иное, как повторности.

Сколько объектов должно быть в повторности, чтобы данные можно было статистически обработать? Одно-единственное на-

блюдение нельзя обработать статистически. Однако если оно оценивается в совокупности с другими однородными данными — то уже можно. Например, можно сравнить единственное наблюдение с выборкой для его проверки на выброс. Два наблюдения позволяют оценить среднее и отклонения от него. Через две точки можно уже провести линию, а по трём наблюдениям можно оценить как линейную регрессию, так и отклонение от неё. Точность таких оценок будет крайне низка, но технически это возможно. Правильным подходом к оценке объёма выборки является подход с учётом: 1) имеющейся информации о степени неопределённости в изучаемой системе и 2) величины эффекта, который необходимо обнаружить. Такой подход мы рассмотрим в конце этого занятия.

Принцип 3. Организация зависимых выборок

Когда нам необходимо выявить действие какого-либо фактора, изменчивость элементов выборки всегда препятствует этому, создавая информационный шум, на фоне которого эффекты слабых по величине факторов становятся неразличимыми. Эффективный способ противодействовать такой изменчивости между субъектами заключается в оценке каждого субъекта на всех уровнях всех изучаемых факторов. Например, если мы желаем обнаружить эффект слабой физической нагрузки на частоту сердечных сокращений (ЧСС), то можем поступить двумя способами:

1) разделить испытуемых на две части: контрольную и опытную. Испытуемые контрольной группы не будут подвергаться нагрузке, а испытуемые опытной — будут. В случае таких независимых выборок вся изменчивость, свойственная как первой, так и второй группе, будет присутствовать в данных в ходе анализа, то есть у нас будет два источника ошибок;

2) измерить ЧСС у всех испытуемых до нагрузки и после нагрузки. В случае таких зависимых выборок каждый испытуемый будет иметь своё собственное контрольное значение, и источник изменчивости будет только один — индивидуальная реакция на нагрузку. В таком эксперименте удастся одновременно получить три выгоды:

а) снизить количество источников изменчивости с двух до одного, то есть увеличить разрешающую способность (мощность) эксперимента;

б) увеличить в два раза объём выборки для сравнения;
в) получить данные по индивидуальной реакции на нагрузку, с которыми далее можно продолжать работать другими методами.

Если мы не ограничимся двумя измерениями, а проведём их серию, то получим ***панель данных*** (*panel data*), в которой каждый ряд значений будут относиться строго к одному субъекту. Если это временная серия, то о таких панельных данных говорят как о ***повторных измерениях*** (*repeated measurements*).

Организация зависимых выборок — очень эффективная процедура, поэтому, когда это возможно, рационально планировать именно такие экспериментальные схемы. **Задание**: придумайте примеры исследований, когда организовать зависимые выборки не получится: а) для живых объектов, б) для каких-либо неживых образцов. Обычно такие примеры относятся к ситуациям, когда живой объект гибнет, а образец разрушается целиком. Очевидно, что в остальных случаях можно пытаться организовать зависимые выборки.

Принцип 4. Рандомизация

► **Рандомизация** (*randomization*) в широком смысле — процесс обеспечения случайности чего-либо. В зависимости от контекста это может быть и перемешивание последовательности чисел в случайном порядке, и взятие из популяции случайной выборки, и сам процесс генерирования случайных чисел. ► **Рандомизация в статистике** — процесс ***случайного назначения*** (*random assignment*) субъектов или других экспериментальных единиц в различные экспериментальные группы, который осуществляется с использованием случайных чисел или другого случайного устройства. В результате рандомизации влияние всех неконтролируемых в исследовании факторов случайным образом распределяется между объектами исследования и не приводит к ***смещению*** (*bias*) интересующих оценок.

> **К сведению**. ***Случайные числа*** (СЧ, *random numbers*) могут быть получены двумя различными способами:
> 1) аппаратный генератор СЧ использует измерение какого-либо хаотического физического процесса («атмосферный шум» в радиодиапазоне, тепловой шум, другие электромагнитные и квантовые явления);
> 2) генератор *псевдослучайных чисел* использует различные вычислительные алгоритмы, дающие длинные серии случайных последовательностей цифр.

> Такие последовательности, во-первых, полностью определяются коротким *начальным значением* (*seed value*), а во-вторых, длина истинно случайной серии оказывается ограниченной, а далее серия циклически повторяется. Именно поэтому к названию получаемых таким способом СЧ добавляют начальную составную часть «псевдо».

Если в исследовании имеется фактор, влияние которого на изучаемое явление очевидно, то имеет смысл наложить на полную рандомизацию ограничение. Например, если известно, что на изучаемый признак влияет возраст или пол, то необходимо разбить — ***стратифицировать*** — выборку сначала по этим факторам, а только потом провести рандомизацию. Это позволит избежать смещения оценок, а также обнаружить возможные взаимодействия факторов (см. далее «Принцип 5. Блочное планирование»).

Рандомизацию следует использовать не только на этапе планирования, но и на этапе сбора материала (получения данных). Представим, что у нас есть препараты объектов экспериментальной и контрольной группы, назначение в которые было проведено корректно и с использованием рандомизации. Однако если мы сначала проанализируем препараты одной группы, а затем — другой, то ошибка анализа опять не будет распределена случайно; это приведёт к смещению оценки, которую мы не сможем отличить от влияния экспериментального фактора. Примеры: 1) в начале работы прибор будет разогрет чуть менее среднего, а в конце — чуть более среднего, что влияет на измерение; 2) при титровании мы определяем точку смены окраски в начале серии несколько иначе, чем в конце; 3) при микроскопировании, по мере роста квалификации, проводим учёт регистрируемых событий всё более точно. Для того чтобы возможные артефакты анализа не повлияли на наши суждения относительно оцениваемых явлений, потенциальную ошибку анализа следует также распределить по объектам случайно, то есть следует проводить анализ образцов на приборе или титрование в случайном порядке, а препараты перед анализом шифровать и также перемешивать. Про это **особенно важно** не забывать молодым исследователям, поскольку именно у них рост квалификации происходит наиболее быстро, а подсознательное желание получить ожидаемый результат слишком велико.

П р и м е р. В распоряжении исследователя, планирующего небольшой пилотный эксперимент, имеется 15 добровольцев,

которые необходимо распределить между двумя группами: основной — получающей экспериментальное лечение — и группой сравнения, получающей традиционное лечение. Поскольку поделить 15 поровну нельзя, было решено сделать большей основную группу.

Задание: с помощью процедуры рандомизации назначить 8 добровольцев в основную группу, а 7 — в группу сравнения.

> **Комментарий**. Выполним задание с помощью электронной таблицы Excel. Начиная с версии Microsoft Office 2003 в процессоре электронных таблиц Excel реализован простой и эффективный алгоритм генерации псевдослучайных чисел, который проходит все тесты на случайность. В предыдущих версиях использовался менее совершенный алгоритм RAND, дававший серию случайных чисел длиной около миллиона значений (что, однако, не является принципиальным для нашей серии в 15 значений).

В пакете Excel

① Откройте новый лист Excel и создайте в двух столбцах первой строки два заголовка: Номер добровольца и Случайное число.

② Введите в первых трёх строчках столбца «Номер добровольца» цифры 1, 2, 3, а затем выделите диапазон с цифрами и протяните за нижний правый угол ячейки (курсор меняется на +) по значение 16. В ячейках окажутся цифры от 1 до 15.

③ В ячейку столбца «Случайное число» напротив № 1 поместим функцию категории «Математические» — СЛЧИС: =СЛЧИС(). В ячейке появится случайное число в диапазоне от 0 до 1. Выделите эту ячейку и скопируйте протяжкой в остальные нужные ячейки.

④ Выделите оба столбца целиком и далее: Редактирование — Сортировка и фильтр — Настраиваемая сортировка — Сортировать по [Случайное число]. Номера добровольцев оказываются перемешанными случайным образом.

⑤ Назначаем первые 8 номеров в основную группу, оставшиеся 7 — в группу сравнения.

Принцип 5. Блочное планирование

► **Организация блоков** (*blocking*) — это процесс, при котором группы (блоки) субъектов располагаются в определённых комбинациях уровней разных факторов. Он направлен на разделение эффектов известных и, возможно, неизвестных факторов, которые могут повлиять на исход исследования. С этим принци-

пом мы уже знакомились на лабораторном занятии № 10, когда рассматривали различные модели дисперсионного анализа и его непараметрические аналоги.

II. Расчёты объёмов выборок

Расчёт объёмов выборок n перед проведением исследования позволяет правильно спланировать и распределить имеющиеся ресурсы. Зачастую при выборе n молодые исследователи руководствуются рекомендациями коллег или научного руководителя, которые, однако, часто базируются на личном опыте, а не на расчёте. Распространено мнение, что чем меньше полученное в исследованиях p-значение, тем лучше.

Вместе с тем в серьёзных исследованиях (пока чаще зарубежных) к выбору объёмов выборок подходят весьма основательно. Ведь как слишком большое, так и слишком малое p-значение указывают на плохое планирование, а соответственно — на недостижение цели или излишнюю трату средств.

Далее мы научимся рассчитывать объёмы выборок для нескольких типичных случаев, в том числе — с использованием специфических программ. Несмотря на некоторую условность предпосылок для расчётов (предположение об известном характере распределения, знание наперёд величины предполагаемых различий, выбор значения мощности исследования и т. д.), полученные значения являются обоснованным ориентиром при планировании исследований, а также дисциплинируют исследователя.

1. Предварительные расчёты

Для приводимых ниже расчётов часто требуется знать величину стандартного отклонения s количественных показателей. Кроме как из собственных пилотных исследований, её можно взять из литературных источников. Однако, поскольку в публикациях чаще приводится не s, а стандартная ошибка среднего m или 95% ДИ для среднего, параметр s потребуется вычислить.

П р и м е р 1. В статье приведены данные в форме $\overline{x} \pm m$ и указан объём выборки: 25,2±1,06 для $n = 5$.

Задача: рассчитать стандартное отклонение s.

Решение. Поскольку $m = \frac{s}{\sqrt{n}}$, то $s = m\sqrt{n}$. $s = 1{,}06\sqrt{5} = 2{,}37$.

П р и м е р 2. В статье приведены данные в форме $\bar{x}$ [95% *ДИ*; 95% *ДИ*], то есть среднее с нижней и верхней границами 95% ДИ, а также указан объём выборки: 25,2 [22,3; 28,1] для $n = 5$.

Задача: рассчитать стандартное отклонение *s*.

Решение. Поскольку границы 95% ДИ для нормально распределённого показателя вычисляются как $\bar{x} \pm \Delta$, где $\Delta = t_{n-1;\alpha=0,05} \times \frac{s}{\sqrt{n}}$, то $s = \frac{\Delta \times \sqrt{n}}{t_{n-1;\alpha=0,05}}$.

Рассчитаем Δ как половину разности границ 95% ДИ:

Δ = (28,1 – 22,3) / 2 = 2,9.

По таблице (например [1. С. 130]) или в расчётном файле «Калькулятор распределений.xls» найдём критическое значение *t*-распределения для α = 0,05 и числа степеней свободы $df = n - 1 = 5 - 1 = 4$.

$t_{4;\alpha=0,05} = 2{,}7764$.

Тогда $s = \frac{\Delta \times \sqrt{n}}{t_{n-1;\alpha=0,05}} = \frac{2{,}9\sqrt{5}}{2{,}7764} = 2{,}34.$

2. Кросс-секционные исследования

► **Кросс-секционные,** или **перекрёстные исследования** (*cross-sectional study*) — описательные исследования (*descriptive study*), призванные установить значение параметра в популяции. Это может быть среднее значение некоего количественного показателя или частота качественного признака, например распространённость (превалентность) патологии.

2.1. Количественные показатели

Объём выборки $n = \frac{\left(z_{1-\alpha/2}\right)^2 \times s^2}{d^2}$, где $z_{1-\alpha/2}$ — значение площади под кривой стандартного нормального распределения (1,64 для $P = 0{,}10$; **обычно 1,96 для $P = 0{,}05$**; 2,58 для $P = 0{,}01$), *s* — величина стандартного отклонения показателя (по литературным данным или пилотным исследованиям), *d* — точность оценки.

П р и м е р. Исследователь желает определить среднее значение систолического давления крови у детей своего населённого пункта при величине ошибки I рода в 5 % (α = 0,05) и точности в 5 мм рт. ст. в обе стороны от среднего (d = 5). Согласно опубликованным данным, для этого показателя s = 25 мм рт. ст.

Задача: рассчитать необходимый объём выборки.

Решение: $n = n = \frac{\left(z_{1-\alpha/2}\right)^2 \times s^2}{d^2} = n = \frac{1{,}96^2 \times 25^2}{5^2} = 96{,}04 \approx 96.$

Таким образом, для оценки среднего значения с выбранной точностью потребуется обследовать не менее 96 детей.

2.2. Качественные показатели

Объём выборки $n = \frac{\left(z_{1-\alpha/2}\right)^2 \times p(1-p)}{d^2}$, где $z_{1-\alpha/2}$ — значение площади под кривой стандартного нормального распределения (1,64 для P = 0,10; **обычно 1,96 для P = 0,05**; 2,58 для P = 0,01), p — ожидаемая частота событий в популяции (по литературным данным или пилотным исследованиям), d — точность оценки.

П р и м е р. Исследователь желает определить распространённость артериальной гипертензии у детей своего населённого пункта. Согласно опубликованным данным, обычно она не превышает 15 % (p = 0,15). Величина ошибки I рода выбирается в 5 % (α = 0,05), точность d — также в 5 % (d = 0,15).

Задача: рассчитать необходимый объём выборки.

Решение:

$$n = \frac{\left(z_{1-\alpha/2}\right)^2 \times p(1-p)}{d^2} = \frac{1{,}96^2 \times 0{,}15(1-0{,}15)}{0{,}05^2} = 195{,}9216 \approx 196.$$

Таким образом, для установления распространённости заболевания с выбранной точностью потребуется обследовать, по меньшей мере, 196 детей.

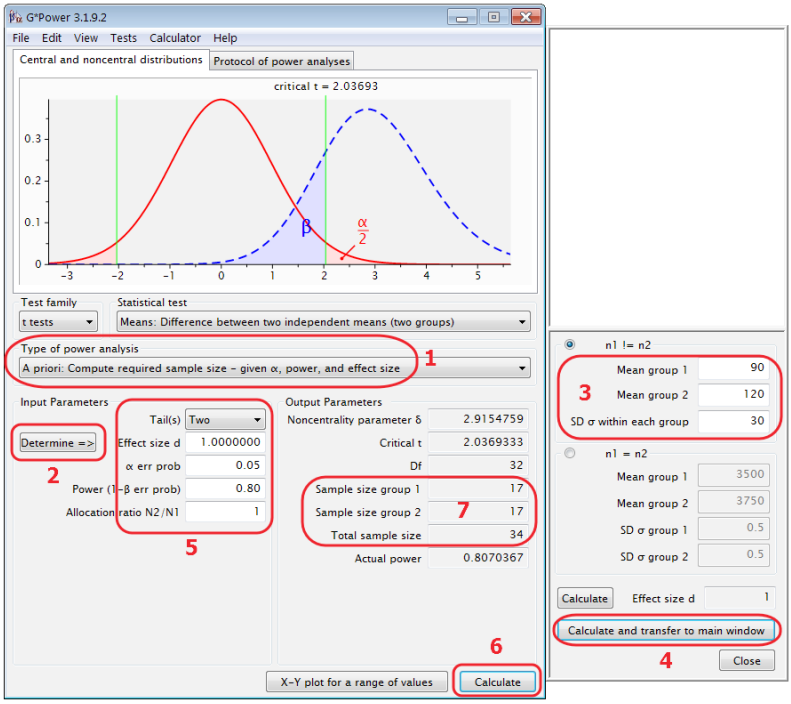
3. Сравнение двух выборок

3.1. Количественные показатели с нормальным распределением

П р и м е р. Стандартное удобрение почвы навозом позволяет получить массу сухого растения кукурузы на стадии молочной зрелости порядка 90 г. Стандартное отклонение s = 30. По данным литературы, современные удобрения позволяют повысить указанный показатель примерно на 30 %, то есть до 120 г. На агробиостанции планируется испытание нового экспериментального удобрения. Сколько растений потребуется высеять на двух грядках (контроль — стандартное удобрение, опыт — новое), отведённых под эксперимент, для сравнительной оценки эффективности нового удобрения?

В пакете G*Power (см. ссылку на с. 12).

① Путь: Tests (Тесты) — Means (Средние) — Two independent groups (Две независимые выборки).



② Выбираем опцию расчёта необходимого объёма выборки — sample size (1).

③ Нажать кнопку [Determine] (2) и задать в выпавшем справа окне средние значения в группах и стандартное отклонение (3). Если предполагаются разные стандартные отклонения, то радиокнопкой следует выбрать вариант ниже.

④ Нажать кнопку расчёта (4) и передать результат в основное окно программы.

⑤ Для большинства задач выбирается двусторонний критерий, поэтому в окне Tail(s) — Хвост(ы) выставляем [Two]. Ошибку II рода β обычно устанавливают в 2–4 раза больше, чем ошибку I рода α. Достаточно взять β = 0,20 и, соответственно, выставить мощность (Power) равной 1 – 0,20 = 0,80. **ВАЖНО!** Обратите внимание, что программа работает с точкой в качестве десятичного разделителя. Если объёмы выборок предполагаются равными, оставляем соотношение Allocation ratio N2/N1 = 1.

⑥ Нажать кнопку расчёта [Calculate] и получить результат в строках области (7). Для нашего примера достаточным объёмом выборки следует считать 17 экземпляров растений в каждой группе. Для сравнения будем использовать критерий Стьюдента.

3.2. Количественные показатели с ненормальным распределением и порядковые показатели

П р и м е р. Воспользуемся данными примера с кукурузой, но учтём, что распределение массы сухого растения может отличаться от нормального. Пакет G*Power позволяет выбрать в качестве планируемого метода сравнения групповых средних ранговый критерий Уилкоксона — Манна — Уитни.

В пакете G*Power

Путь: Tests — Means — Two independent groups: Wilcoxon (nonparametric). В качестве предполагаемого исходного распределения (Parent distribution) можно выбрать: нормальное (Normal), островершинное Лапласа (Laplace) или плосковершинное логистическое (Logistic). Здесь же можно выбрать [min ARE], при котором для расчёта используется метод A.R.E., устанавливающий связь между *t*-критерием и критерием Уилкоксона (см. учебник к пакету).

Задание. Рассчитайте самостоятельно объёмы выборок для данных примера 3.1. Распределение — Normal, затем — min ARE. Мощность — 0,8.

Ответ. При использовании для сравнения критерия Уилкоксона — Манна — Уитни объёмы выборок следует принять в 18 или 20 экземпляров растений в каждой группе.

3.3. Качественные показатели

П р и м е р. Частота клеток с хромосомными аберрациями (ХА) на стадии метафазы в культуре лимфоцитов человека составляет обычно около 1 %. Сколько клеток следует проанализировать, чтобы доказать двукратное увеличение частоты клеток с ХА в эксперименте с влиянием потенциального мутагена?

В пакете G*Power

① Путь: Tests (Тесты) — Proportions (Доли) — Two independent groups: Inequality, z-Test (Две независимые выборки, неравенство, z-критерий) (см. скриншот на с. 273).

② Заполнить поля с долями: p1 (контроль) — 0,01 (1 %), p2 (опыт) — 0,02 (2 %, поскольку предполагается двукратное превышение спонтанного уровня). Критерий — двусторонний (Tail(s) — Two), уровень значимости — 0,05, мощность — 0,8. Кнопка [Calculate].

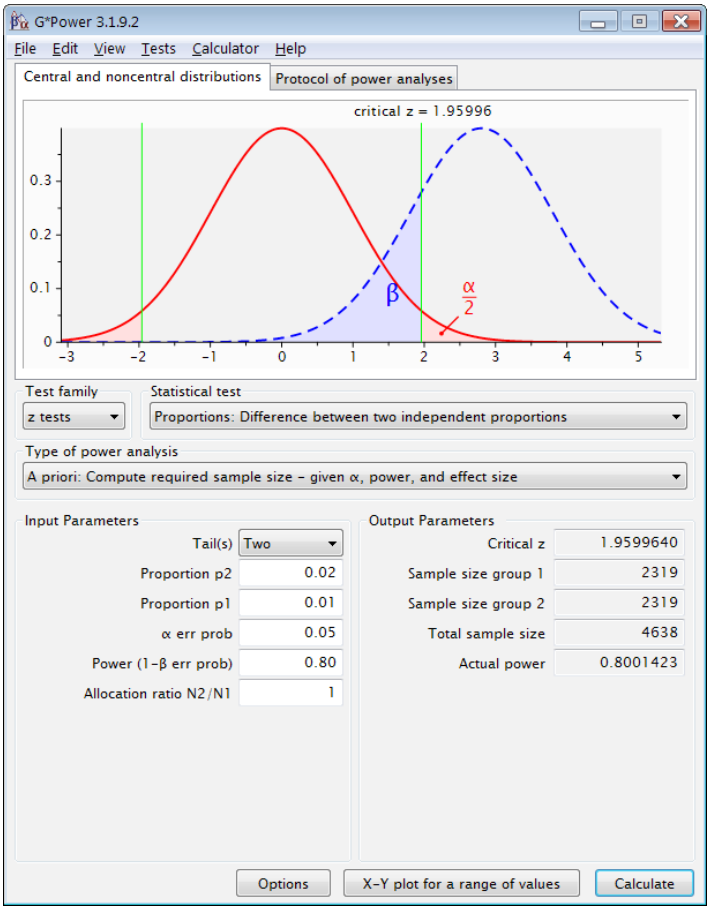
Получаем результат о необходимости анализа в контроле и опыте по 2 319 клеток.

Задание. Рассчитайте самостоятельно объёмы выборок для этих же данных, но с использованием точного метода Фишера, который найдите самостоятельно.

Ответ. По 1 982 клетки в каждом их вариантов опыта.

4. Поиск связей

П р и м е р 1. Согласно распространённой классификации принято считать, что связь показателей с коэффициентом корреляции ≤0,3 является слабой. Сколько пар объектов необходимо исследовать, чтобы доказать наличие слабой связи двух количественных признаков (распределение считать нормальным)?



В пакете G*Power

① Путь: Tests — Correlation and regression — Correlation: Bivariate normal model (Двумерная нормальная модель).

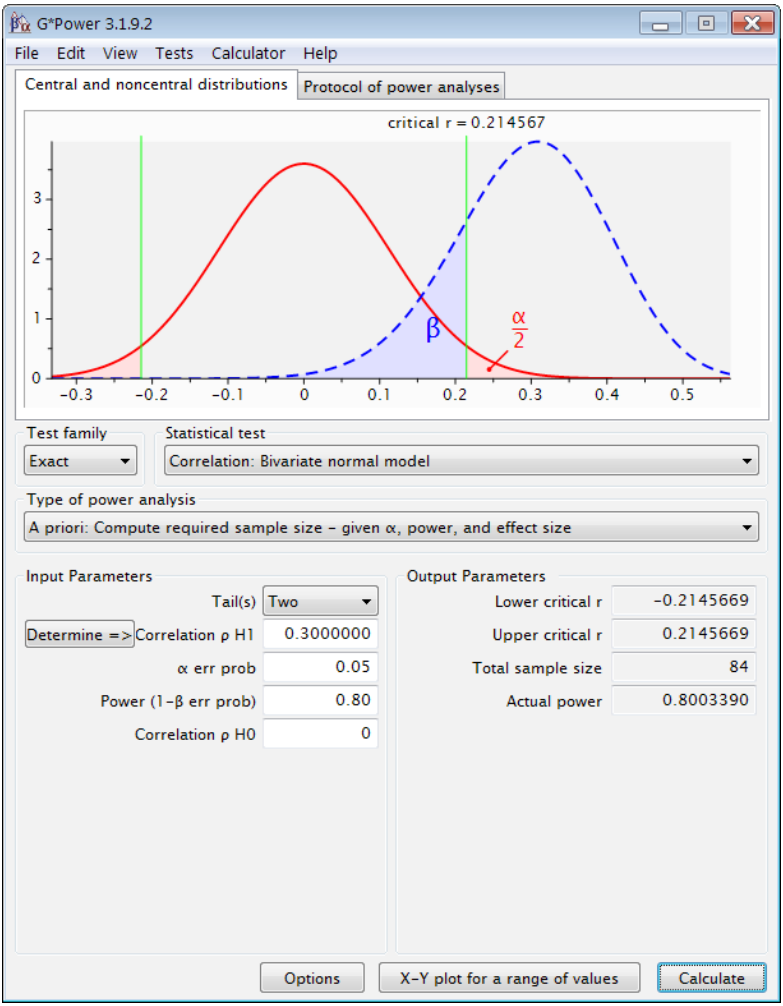
② Устанавливаем двусторонний критерий: Tail(s) — Two. Корреляция — 0,3; уровень значимости — 0,05, мощность — 0,8. Кнопка [Calculate] (см. скриншот на с. 274).

Получаем результат о необходимости анализа, по меньшей мере, 84 пар значений.

П р и м е р 2. Известно, что в норме слабые связи между многими физиологическими показателями в результате стресса

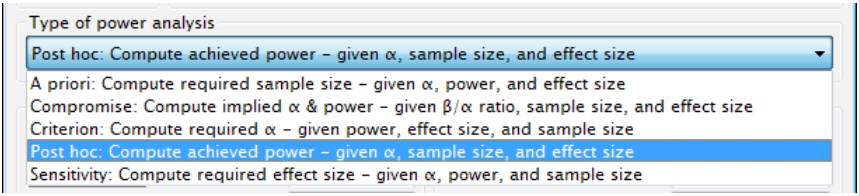
усиливаются. На этом, в частности, основан метод корреляционной адаптометрии. Сколько необходимо исследовать объектов, чтобы доказать для пары признаков увеличение коэффициента корреляции с 0,3 до 0,5 (распределение считать нормальным)?

В пакете G*Power всё настраивается аналогично примеру 1, однако вводятся два значения корреляций: Correlation ρ H1 = 0,5, а Correlation ρ H0 = 0,3. Получаем результат о необходимости исследования 139 пар случаев.



5. Прочие задачи

Мы рассмотрели несколько типичных и достаточно простых примеров планирования объёмов выборок. Тем не менее принцип организации данных и проведения расчётов в пакете G*Power аналогичен и для других задач: сравнения нескольких выборок, анализа регрессий и т. д. Полагаем, теперь вы сможете самостоятельно провести такие расчёты. Отметим только, что данную программу можно использовать не только для расчётов объёмов выборок, но и для других целей:



Например, при отсутствии статистически значимых эффектов полезно оценить достигнутую в исследовании мощность (*Compute achieved power*): если она была мала, скажем, менее 0,8, значит имеющихся данных было просто недостаточно для обоснованного вывода.

Помимо G*Power для расчётов объёмов выборок (*Sample size calculation*) можно использовать ряд других программ, включая онлайновые ресурсы:

Open Epi: http://www.openepi.com/Menu/OE_Menu.htm

EpiTools: http://epitools.ausvet.com.au/content.php?page=home

Free statistics calculators: http://www.danielsoper.com/statcalc/default.aspx

StatPages: http://statpages.info/

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Основная

1. Закс, Л. Статистическое оценивание / Л. Закс ; пер. с нем. В. Н. Варыгина ; под ред. Ю. П. Адлера, В. Г. Горского. М. : Статистика, 1976. 598 с.
2. Ланг, Т. А. Как описывать статистику в медицине. Аннотированное руководство для авторов, редакторов и рецензентов / Т. А. Ланг, М. Сесик ; пер. с англ. под ред. В. П. Леонова. М. : Практ. медицина, 2011. 480 с.

Дополнительная

3. Джонгман, Р. Г. Г. Анализ данных в экологии сообществ и ландшафтов / Р. Г. Г. Джонгман, С. Дж. Ф. Тер Браак, О. Ф. Р. Ван Тонгсен ; пер. с англ. под ред. А. Н. Гельфана. М. : РАСХН, 1999. 306 с.
4. Зорин, Н. А. «Достоверность» или «статистическая значимость» — 12 лет спустя / Н. А. Зорин // Педиатр. фармакология. 2011. Т. 8, № 5. С. 13–19.
5. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. М. : Физматлит, 2006. 816 с.
6. Микиша, А. М. Толковый математический словарь. Основные термины / А. М. Микиша, В. Б. Орлов. М. : Рус. яз., 1989. 244 с.
7. Монтгомери, Д. К. Планирование эксперимента и анализ данных : пер. с англ. / Д. К. Монтгомери. Л. : Судостроение, 1980. 384 с.
8. Мэгарран, Э. Экологическое разнообразие и его измерение : пер. с англ. / Э. Мэгарран. М. : Мир, 1992. 161 с.
9. Нохрин, Д. Ю. Группы крови и характер: виктимологический подход с анализом статистики травматизма / Д. Ю. Нохрин, Л. А. Рязанова, Т. В. Тишевская // Вестн. Челяб. гос. ун-та. 2015. № 21 (376). Биология. Вып. 3. С. 128–137.
10. Орлов, А. И. О применении статистических методов в медико-биологических исследованиях / А. И. Орлов // Вестн. Акад. мед. наук СССР. 1987. № 2. С. 88–94.
11. Петри, А. Наглядная статистика в медицине / А. Петри, К. Сэбин ; пер. с англ. под ред. В. П. Леонова. М. : Геотар-Мед, 2003. 139 с.
12. Реброва, О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. М. : МедиаСфера, 2002. 312 с.

13. Урбах, В. Ю. Биометрические методы / В. Ю. Урбах. М. : Наука, 1964. 415 с.

14. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др. ; пер. с англ. под ред. И. С. Енюкова. М. : Финансы и статистика, 1989. 215 с.

15. Хальд, А. Математическая статистика с техническими приложениями : пер. с англ. / А. Хальд. М. : Иностр. лит., 1956. 664 с.

16. Флейтс, Дж. Статистические методы для изучения таблиц долей и пропорций / Дж. Флейтс. М. : Финансы и статистика, 1989. 208 с.

17. Шитиков, В. К. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R / В. К. Шитиков, Г. С. Розенберг. Тольятти : Кассандра, 2013. 314 с.

18. Энциклопедия статистических терминов : в 8 т. Т. 2 : Инструментальные методы статистики. М. : Федер. служба гос. статистики, 2011. 474 с.

19. Эсбенсен, К. Анализ многомерных данных. Избранные главы. / К. Эсбенсен ; пер. с англ. под ред. О. Е. Родионовой. Черноголовка : Изд-во ИПХФ РАН, 2005. 160 с.

20. Brown, L. D. Interval Estimation for a Binomial Proportion / L. D. Brown, T. T. Cai, A. DasGupta // Statistical Science. 2001. Vol. 16, № 2. P. 101–117.

21. Sokal, R. R. Biometry: the principles and practice of statistics in biological research / R. R. Sokal, F. J. Rohlf. N.-Y. : Freeman & Co, 1995. 850 p.

22. Zar, J. H. Biostatistical analysis / J. H. Zar. 5th ed. New Jersey : Prentice Hall Inc., 2010. 944 p.

23. Conover, W. J. Rank Transformations as a bridge between parametric and nonparametric statistics / W. J. Conover, R. L. Iman // The American Statistician. 1981. Vol. 35, № 3. P. 124–129.

УКАЗАТЕЛЬ ТЕРМИНОВ

Указатель содержит все встречавшиеся в данном практикуме статистические термины. Часть из них раскрывается в тексте достаточно полно, часть — менее полно, часть — только упоминается и подразумевает самостоятельный поиск более полной информации.

Учебное издание

КЛАССИЧЕСКОЕ УНИВЕРСИТЕТСКОЕ ОБРАЗОВАНИЕ

НОХРИН Денис Юрьевич

ЛАБОРАТОРНЫЙ ПРАКТИКУМ ПО БИОСТАТИСТИКЕ

Корректор *М. В. Трифонова*
Вёрстка *М. В. Трифоновой*
Макет обложки *Т. В. Ростуновой*

Подписано в печать 21.05.18.
Формат 60×84 $^{1}/_{16}$.
Бумага офсетная. Гарнитура Sitka.
Усл. печ. л. 16,9. Уч.-изд. л. 15,0.
Тираж 100 экз. Заказ 230.
Цена договорная

Челябинский государственный университет
454001, Челябинск, ул. Братьев Кашириных, 129

Издательство Челябинского государственного университета
454021, Челябинск, ул. Молодогвардейцев, 57б

Д. Ю. Нохрин

ЛАБОРАТОРНЫЙ ПРАКТИКУМ ПО БИОСТАТИСТИКЕ

Издательство
Челябинского государственного университета

978 5 727114872