ПРАВИТЕЛЬСТВО МОСКВЫ
ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ

СОГЛАСОВАНО

Главный внештатный специалист
по лучевой и инструментальной
диагностике
Департамента здравоохранения
города Москвы

_____Ю. А. Васильев
«15» октября 2024 г.

РЕКОМЕНДОВАНО

Экспертным советом по науке
Департамента здравоохранения
города Москвы № 14
«21» октября 2024 г.

ПРОВЕДЕНИЕ СТАТИСТИЧЕСКОГО АНАЛИЗА НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ R В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Часть 1

Методические рекомендации 58

Москва
2024

УДК 519.23
ББК 32.97
П 78

Серия «Лучшие практики лучевой и инструментальной диагностики»
Основана в 2017 году

Организация-разработчик:
Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»

Авторы-составители:
Васильев Ю. А. – канд. мед. наук, главный внештатный специалист по лучевой и инструментальной диагностике ДЗМ, директор ГБУЗ «НПКЦ ДиТ ДЗМ»
Никитин Н. Ю. – канд. физ.-мат. наук, старший научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»
Будыкина А. В. – младший научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»
Памова А. П. – канд. мед. наук, научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»
Бобровская Т. М. – младший научный сотрудник отдела инновационных технологий ГБУЗ «НПКЦ ДиТ ДЗМ»
Арзамасов К. М. – канд. мед. наук, руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

П 78 Проведение статистического анализа на языке программирования R в медико-биологических исследованиях: методические рекомендации. Часть 1 / авт.-сост. Ю. А. Васильев, Н. Ю. Никитин, А. В. Будыкина [и др.]. – Вып. 139. – М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2024. – 80 с.

Рецензенты:
Носовский Андрей Максимович – д-р биол. наук, ведущий научный сотрудник ФБГУН «ГНЦ РФ – ИМБП»
Белиловский Евгений Михайлович – канд. биол. наук, заведующий отделом эпидемиологического мониторинга туберкулеза ГБУЗ «МНПЦ борьбы с туберкулезом ДЗМ»

Методические рекомендации предназначены для подготовки слушателей курсов по статистическому анализу в медико-биологических исследованиях и расчету количества образцов (исследований), необходимых для проведения статистического анализа.

В первой части представлены основные величины, принятые в статистическом анализе, и способы их вычисления. Описаны методы проверки количественных данных на соответствие нормальному закону распределения. Все примеры, приведенные в тексте издания, реализованы на языке программирования R на открытых наборах данных, входящих в пакеты этого языка.

Первая часть методических рекомендаций разработана в рамках выполнения НИОКР «Разработка платформы повышения качества ИИ-сервисов для медицинской диагностики» (№ ЕГИСУ: 123031400006-0) в соответствии с приказом Департамента здравоохранения города Москвы от 21.12.2022 № 1196 «Об утверждении государственных заданий, финансовое обеспечение которых осуществляется за счет средств бюджета города Москвы, государственным бюджетным (автономным) учреждениям, подведомственным Департаменту здравоохранения города Москвы, на 2023 год и плановый период 2024 и 2025 годов»

ISSN

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ

В настоящем документе использованы ссылки на следующие нормативные документы (стандарты):

1. ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятности от нормального распределения.
2. Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть 2. Непараметрические критерии.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ГОСТ – государственный стандарт.

Листинг – распечатка программного кода, написанного на языке R.

ОСТ – отраслевой стандарт.

ПК – персональный компьютер.

ПО – программное обеспечение.

СанПиН – санитарно-эпидемиологические нормы и правила.

СНиП – строительные нормы и правила.

СПИД – синдром приобретенного иммунодефицита.

США – Соединенные Штаты Америки.

$\sum_{i=1}^{N} x_i$ – сумма всех значений *x* с номерами, изменяющимися от 1 до N.

$\sqrt{\frac{1}{N-1}}$ – корень квадратный из деления единицы на N минус 1.

{…} – в математических выражениях в фигурных скобках обозначено «множество» значений, принимаемых какой-либо величиной.

“:” – в математических выражениях означает «определено на…» или «с областью определения…». К примеру, X:{P=P_0} читается, как: «величина X, определенная на множестве P, равном P_0».

“|” – в математических выражениях означает «такое, что…».

“∈” – данный знак читается как «принадлежит». Например, X:{P∈*P*} читается как: «величина X определена на множестве P, принадлежащего к *P”*».

[a; b] – интервал значений от *a* до *b,* включающий значения границ интервалов (т. е. включающий и *a,* и *b*).

(a; b**)** – интервал значений от a до b, не включающий значения границ интервалов (т. е. **не** включается в интервал значение ***a*** и значение ***b***).

(a; b] – полуинтервал значений от *a* до *b,* **не** включающий значение нижней границы (***a***) и включающий верхнюю границу (*b*).

[a; b**)** – полуинтервал значений от *a* до *b,* включающий значение нижней границы (*a*) и **не** включающий верхнюю границу (***b***).

КАК РАБОТАТЬ С МЕТОДИЧЕСКИМИ РЕКОМЕНДАЦИЯМИ

Методические рекомендации предназначены для слушателей курса «Методология планирования и проведения научных исследований. Биостатистика» и не являются самоучителем (в широком смысле данного понятия) по программированию и статистическому анализу данных на языке R, а представляют дополнение к очному курсу по предмету.

Для слушателей, впервые сталкивающихся со статистическим анализом и языком программирования R, решивших самостоятельно освоить предмет (программирование на языке R и статистический анализ данных) с помощью представленных методических рекомендаций, авторы советуют следующий порядок освоения материала:

1. Освоить основной синтаксис языка R с помощью дополнительной литературы, в частности можно обратиться к работе Н. Мэтлоффа [3].
2. Далее последовательно ознакомиться с материалом, изложенным в методических рекомендациях. К каждому статистическому критерию или тесту, изложенному в представленных методических рекомендациях, приведены детальные примеры его применения с кодом, написанным на языке R.
3. Если освоение материала проходит самостоятельно, без возможности прямого или удаленного контакта с консультантом, и представленный в методических рекомендациях материал кажется непонятным, то обучаемому рекомендуется ознакомиться с материалами, представленными в источниках [1], [2] и [4] списка литературы[1].

Для слушателей курса «Методология планирования и проведения научных исследований. Биостатистика», проходящих обучение под руководством преподавателя, методические рекомендации являются дополнительным источником, позволяющим лучше усвоить материал лекций и практических занятий. В этом случае порядок работы с представленными методическими рекомендациями соответствует порядку чтения лекций и проведения семинарских занятий. Аспирантам рекомендуется провести самостоятельное статистическое исследование в соответствии с алгоритмом, представленным на рисунке 2 раздела «Введение», на основе открытых наборов данных. Интернет-ресурсы с открытыми наборами данных указаны в подразделе 2.8 второй части методических рекомендаций.

При использовании электронной версии методических рекомендаций у читателя возникнет желание не вводить код вручную, а копировать программный код из текста и вставлять его в файл скрипта. Настоятельно не рекомендуется это делать по двум основным причинам:

1. При самостоятельном наборе программного кода команды алгоритм работы с данными лучше запоминается.
2. При прямом переносе и немедленном исполнении возникают ошибки, связанные с различиями стандартных шрифтов, принятых при наборе текста в методических рекомендациях и шрифтами, применяемыми в интегрированной среде разработки. Наиболее частая ошибка связана с кавычками, знаком минус (при переносе может быть воспринят как знак дефиса), применяемыми для экранирования строковых констант или значений.

[1] Более глубокого понимания критериев и тестов можно добиться, ознакомившись с оригинальными публикациями, которые можно найти с помощью поисковой системы https://scholar.google.com.

ВВЕДЕНИЕ

Научная методология проведения исследований включает в себя:

1) постановку вопроса;
2) обзор и анализ литературы;
3) формулировку цели исследования;
4) формулировку задач исследования;
5) формулировку предварительной гипотезы;
6) планирование эксперимента (или сбора данных);
7) проведение эксперимента (сбор и структурирование данных);
8) анализ полученных результатов эксперимента (данных);
9) проверку гипотезы на основе полученных данных;
10) построение математической модели[2];
11) проверку математической модели на результатах других экспериментов (на других данных).

В практике научных исследований «возникновение вопросов» по тематике деятельности исследователя является частым явлением[3], что обусловлено многими причинами, например, обычным исследовательским любопытством или клинической (или технической) необходимостью. Возникающие вопросы, как правило, не имеют законченной формы и не представляют собой сформулированную цель исследования, декомпозированную на задачи. Для того чтобы возникший у исследователя вопрос приобрел законченную форму в виде сформулированной цели исследования, необходимо проведение предварительного литературного обзора по тематике вопроса. В рамках литературного обзора исследователю необходимо ответить на несколько вопросов:

1. Существуют ли литературные источники по интересующему исследователя вопросу (исторические источники, научные и научно-популярные публикации, учебная, учебно-методическая литература, ГОСТы, ОСТы, СНиПы, СанПиНы и т. д.)?[4][5]
2. Насколько широко интересующий вопрос представлен в научной литературе (сколько публикаций на данную тему удалось найти, как давно найденные источники были опубликованы, степень и уровень достоверности найденных источников)?
3. Как данный вопрос представлен в научной литературе – как нерешенный или как частично решенный, или считается полностью решенным?
4. Какие методики и подходы применяются для решения исследуемого вопроса (экспериментальные или/и теоретические)?

---

[2] Построение математической модели является обязательным, если конечной целью исследования выступает предсказание каких-либо свойств, описание явлений и процессов (феноменологические модели).

[3] Даже если исследователь обладает широким кругозором и высоким уровнем образования в области возникшего вопроса, предварительное проведение литературного обзора необходимо для понимания актуальности вопроса, в частности: какие части возникшего вопроса уже детально исследованы другими авторами, а какие остались без внимания и почему.

[4] Для поиска соответствующей литературы рекомендуется использовать специализированные поисковые платформы, такие как Google Scholar (https://scholar.google.com/), электронные библиотеки e-library (https://www.elibrary.ru) или (https://pubmed.ncbi.nlm.nih.gov). Также следует использовать каталоги Российской государственной библиотеки (https://www.rsl.ru) и Государственной публичной научно-технической библиотеки России (https://www.gpntb.ru).

[5] В исследовательской практике очень редко рождаются уникальные вопросы – те, с которыми никто ранее не сталкивался. Часто вопрос, возникающий у исследователя, уже был кем-то изучен или изучен не до конца.

На основании проведенного анализа литературных источников исследователь может сформулировать цель исследования и провести ее декомпозицию (если это необходимо) на задачи, которые требуется решить[6], чтобы достигнуть поставленной цели. После постановки цели и задачи исследования и проведения предварительного анализа литературы формулируется предварительная гипотеза (предположение, выдвигаемое исследователем о наличии или отсутствии эффекта или явления).

При частичном подтверждении или опровержении выдвинутой гипотезы результатами проведенного эксперимента или на основании поступивших данных исследователь возвращается к пункту 4, уточняет или выдвигает новую гипотезу и повторяет все нижестоящие пункты.

Финальным этапом выполнения работ является построение математической модели, или статистического описания результатов проведенного исследования. В самом простом случае в качестве математической модели могут выступать вычисленные на основании количественных данных основные параметры выборочного распределения вероятностей, выборочные средние значения, средневзвешенные значения, среднее квадратическое отклонение, квантили и т. д. В целом модель включает величины вычисленных статистических параметров, составляющие основу описательной статистики.

Полученная математическая модель должна быть подвергнута проверке на других данных, имеющих ту же природу и подобные условия получения, что и данные, использованные для построения исходной модели. Математическая модель может носить феноменологический[7] характер и не обладать предсказательной способностью в случае изменений в условиях проведения эксперимента или получения данных.

Каждый из этапов проведения исследований имеет высокую степень важности для получения достоверного конечного результата. Наиболее ресурсоемкими частями являются этапы 6–9, ошибки в этих этапах приводят к большим финансовым потерям и дискредитации исследования как такового. За период развития науки было разработано достаточно большое количество методов, позволяющих провести этапы 6–9 цикла научных исследований. При наличии большого объема данных, факторов, влияющих на исследуемый процесс или свойство, и анизотропии исследуемых свойств в пространстве и/или во времени применение детерминированных методов анализа является крайне затруднительной и ресурсоемкой процедурой. Для проведения подобных исследований применяют методы статистического анализа и/или моделирования.

В данных методических рекомендациях рассматриваются основные статистические методы, рекомендованные к использованию национальными стандартами по метрологии, рекомендациями по проведению статистического анализа результатов эксперимента и в публикациях ряда авторов.

В настоящее время для проведения статистического анализа данных разработано и применяется большое количество программного обеспечения. Наиболее известными специализированными программными продуктами являются:

1. Statistica©[8].
2. SPSS Statistics©[9].

---

[6] Хорошим методическим руководством по решению математических (да и в целом исследовательских задач) является книга «Как решать задачу: понимание постановки задачи, составление и осуществление плана, анализ решения». См.: Пойа Д. Как решать задачу: понимание постановки задачи, составление и осуществление плана, анализ решения / пер. с англ. В. Г. Звонаревой и Д. Н. Белла; под ред. и с предисл. Ю. М. Гайдука. 4-е изд. М.: URSS, 2009. 206 с.

[7] Феноменологический (от слова феномен) – эмпирическое описание наблюдаемого явления или эффекта.

[8] Коммерческий программный продукт, разработанный компанией Dell (США), а настоящим владельцем является компания TIBCO (США).

3. Minitab©[10].
4. Salstat[11].
5. JASP[12].
6. Jamovi[13].
7. Язык программирования R[14] с интегрированной средой разработки (IDE – integrated development environment) RStudio[15©].
8. Язык программирования Python с интегрированной средой разработки Spyder[16].

Обзор всего существующего на сегодняшний день программного обеспечения, позволяющего проводить статистический анализ данных, потребует отдельной книги. Стоит отметить только основные особенности, которые необходимо учитывать при выборе того или иного инструмента:

1. Стоимость коммерческого программного обеспечения. Она, как правило, очень высока, а набор функций, доступных пользователю, ограничен финансовыми возможностями организации.
2. Возможность установки и использования на различных операционных системах таких, как Microsoft Windows, OS Linux, macOS и др.
3. Наличие открытого исходного кода программного обеспечения (ПО). Такое ПО обладает большей гибкостью в части возможности добавления собственных функций и проверки правильности реализации сторонних.
4. Наличие специализированного языка программирования, адаптированного для решения конкретных задач. Это ускоряет решение сложных вычислительных задач, снижает требования к аппаратным ресурсам и повышает гибкость программного обеспечения при решении специфичных задач.
5. Доступность программного обеспечения для исследователей с малым объемом финансирования.
6. Также важно учитывать наличие широкой поддержки данного инструмента статистической обработки академическим сообществом (наличие научно обоснованных и верифицированных алгоритмов, применяемых для проведения статистических тестов и анализа данных).

---

[9] Коммерческий программный продукт, разработанный в Чикагском университете США, в настоящее время права на данный продукт принадлежат компании IBM (США).

[10] Коммерческий программный продукт, разработанный в университете штата Пенсильвания (США), распространяемый компанией Minitab Inc. со штаб-квартирой в Пенсильвании (США).

[11] Свободно распространяемое программное обеспечение, разработчиком которого являются Алан Дж. Салмони и Марк Ливингстон. Выпуск новых версий ПО закончился в 2003 году.

[12] Свободно распространяемый аналог SPSS Statistics со встроенным языком программирования R, поддерживается Университетом Амстердама (Нидерланды).

[13] Свободно распространяемый программный продукт, предназначенный для статистического анализа данных, считается аналогом SPSS Statistics.

[14] Язык программирования высокого уровня, разработанный сотрудниками статистического факультета Оклендского университета (США) для статистического анализа данных. Свободно распространяется.

[15] RStudio не является единственной интегрированной средой разработки на языке R, но практика применения показала, что данная среда наиболее удобна для применения.

[16] В этот язык программирования входит набор библиотек, позволяющих проводить статистический анализ данных, собственно, как и компилируемых языков высокого уровня, таких как Си и Си++. Основным минусом Python для статистического анализа данных является его более общее назначение (в ряде случаев то, что на R решается двумя строчками кода, на Python может потребовать написания кода в несколько десятков строк). Да и в R практически каждая библиотека, набор данных и часто функции сопровождаются публикациями в академической печати, чего не скажешь о библиотеках и функциях, входящих в Python.

Из всего вышеизложенного следует, что во многих случаях наиболее подходящим в широкой практике для проведения статистических исследований является ПО с открытым исходным кодом и/или специализированные языки программирования.

Одним из таких является язык программирования R с интегрированной средой разработки RStudio©, разработанный и поддерживаемый компанией Posit Software, PBC и распространяющийся по открытой лицензии GNU GPL 3[17]. Данная среда разработки не требует специализированных навыков при установке на большинство программно-аппаратных комплексов общего назначения; инструкцию по установке текущей версии RStudio можно найти на сайте проекта[18].

Интерфейс RStudio представлен на рисунке 1.

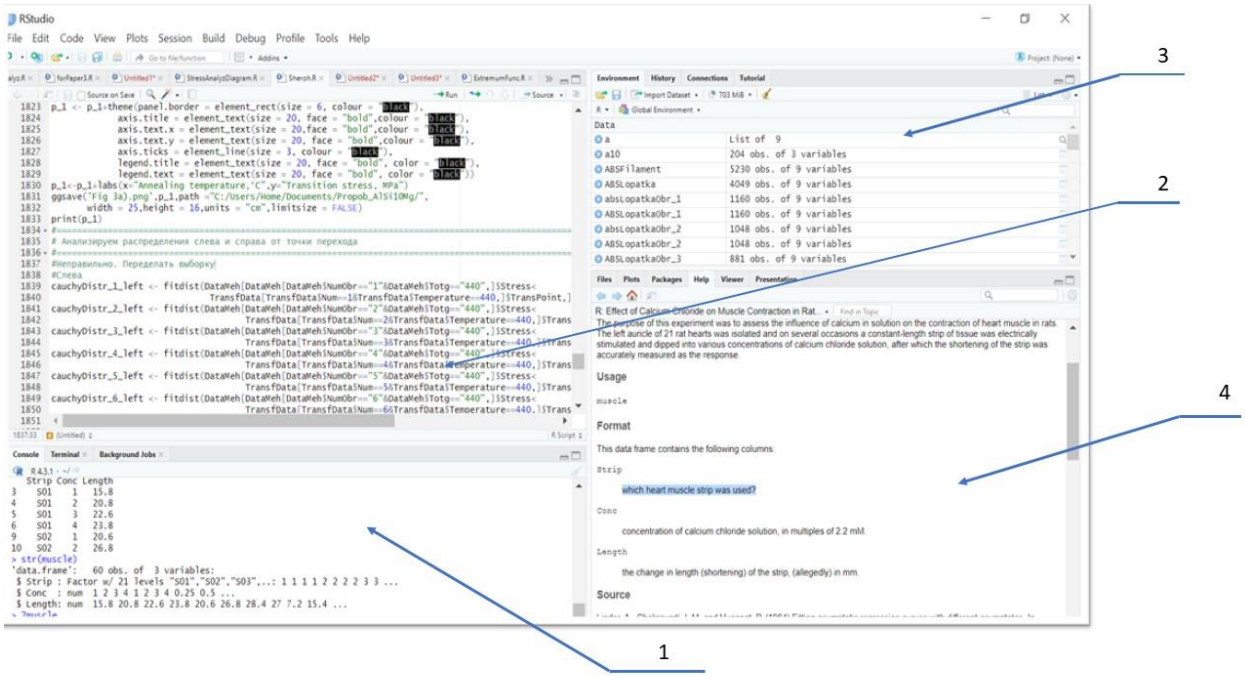


Рисунок 1 – Интегрированная среда разработки RStudio с подключенным компилятором языка программирования R. 1 – интерактивная консоль ввода команд языка R; 2 – поле ввода скриптов на языке R; 3 – область отображения переменных и истории ввода команд в среде RStudio; 4 – область построения графиков, отображения справки (Help) и др.

Детальный обзор интегрированной среды разработки RStudio выходит за рамки настоящих методических рекомендаций и должен рассматриваться на практических занятиях по статистическому анализу данных на языке программирования R[19].

Практическая часть применения статистических методов анализа данных построена на открытых источниках данных, присутствующих в пакете MASS[20] языка R, и на открытом наборе данных, содержащем исследования метрик диагностической

[17] Более подробно о данном типе лицензии можно прочесть: https://www.gnu.org/licenses/agpl-3.0.txt.

[18] См.: https://posit.co/download/rstudio-desktop. В случае, если ваши системы отличны от систем общего назначения (например, исследователь использует ПК с архитектурой arm), то лучше обратиться за помощью к системному администратору.

[19] Обзор системы RStudio представлен на сайте: https://docs.posit.co/ide/user.

[20] Процесс установки пакетов в языке программирования R и IDE RStudio: https://search.r-project.org/R/refmans/utils/html/install.packages.html.

точности 100 врачей[21]. Перечень основных наборов данных, содержащихся в пакете MASS:

1. Aids2[22] – набор данных, содержащий обезличенную информацию о пациентах, у которых диагностирован СПИД в Австралии до 1 июля 1991 г.
2. Cushings[23] – набор данных, содержащий наблюдаемые показатели экскреции с мочой двух стероидных метаболитов у пациентов с подтвержденным диагнозом синдрома Кушинга (гипертензивное заболевание, связанное с избыточной секрецией кортизола надпочечниками).
3. GAGurine[24] – набор данных, содержащий значения концентрации гликозаминогликанов (GAG) в моче у детей в возрасте от 0 до 17 лет.
4. Melanoma[25] – набор данных, содержащий обезличенные данные 205 пациентов в Дании с подтвержденным диагнозом злокачественной меланомы.
5. Pima.te[26] – набор данных, содержащий информацию о популяции женщин в возрасте не менее 21 года, принадлежащих к индейскому племени пима и проживающих в окрестностях Феникса (штат Аризона), которые были обследованы на наличие диабета в соответствии с критериями Всемирной организации здравоохранения. Сбор данных осуществлялся Национальным институтом диабета и болезней органов пищеварения и почек США. Содержит 532 полные записи после исключения (в основном отсутствующих) данных об инсулине в сыворотке крови. Обучающий набор Pima.tr содержит случайно выбранный набор данных 200 испытуемых, а Pima.te – оставшихся 332 испытуемых. Pima.tr2 содержит Pima.tr плюс 100 испытуемых с отсутствующими значениями объясняющих переменных.
6. anorexia[27] – набор данных, содержащий обезличенную информацию об изменении веса молодых женщин, страдающих анорексией.
7. bacteria[28] – набор данных, содержащий тесты на наличие бактерии H. influenzae у детей со средним отитом на северной территории Австралии.
8. biopsy[29] – набор данных о раке молочной железы, полученный из госпиталей Висконсинского университета в Мэдисоне от доктора Уильяма Х. Вольберга. Он оценивал биопсии опухолей молочной железы 699 пациенток до 15 июля 1992 г. Каждый из девяти признаков оценивался по шкале от 1 до 10, известен также исход заболевания; имеются 699 строк и 11 столбцов.

---

[21] Открытый набор данных, доступный на сайте Московского эксперимента: https://www.telemed.ai.

[22] Venables W. N., Ripley B. D. Modern Applied Statistics with S. 4th edition. Springer, 2002.

[23] Aitchison J., Dunsmore I. R. Statistical Prediction Analysis. Cambridge University Press, 1975. Tables 11.1–3.

[24] Venables W., Ripley B. D. S programming. Springer Science & Business Media, 2000.

[25] Andersen Per K., Borgan Ø., Gill R. D., et al. Statistical models based on counting processes. Springer Science & Business Media, 2012.

[26] Smith J. W., Everhart J. E., Dickson, W. C., et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications in Medical Care / ed. R. A. Greenes. Washington, 1988. P. 261–265.

[27] Hand D. J., Daly F., McConway K., et al. A Handbook of Small Data Sets. Chapman & Hall, 1993. P. 229. Data set 285.

[28] Menzies School of Health Research 1999–2000. Annual Report. P. 20. URL: https://www.menzies.edu.au/icms_docs/172302_2000_Annual_report.pdf.

[29] Murphy P. M., Aha D. W. UCI Repository of machine learning databases. [Machine-readable data repository]. Irvine, CA: University of California; Department of Information and Computer Science, 1992.

9. birtwt[30] – набор данных, содержащий факторы риска, связанные с низкой массой тела младенца при рождении. Данные были собраны в Медицинском центре Baystate, Спрингфилд, штат Массачусетс, в течение 1986 года.

10. epil[31] – набор данных о количестве двухнедельных приступов у 59 эпилептиков. Количество приступов регистрировалось в течение базового периода в 8 недель, после чего пациенты были рандомизированы в группу лечения или контрольную группу. Затем подсчеты проводились в течение четырех последовательных двухнедельных периодов. Возраст пациента является единственной ковариатой.

11. gehan[32] – набор данных, содержащий исследования 42 больных лейкемией. Часть из них получала лечение препаратом 6-меркаптопурин, остальные – контрольная группа. Испытание было организовано в виде подобранных пар, обе из которых были выведены из исследования при выходе из ремиссии.

12. muscle[33] – набор данных, содержащий результаты исследования по влиянию концентрации хлорида кальция на сокращение мышц сердца крыс.

13. Indometh[34] – набор данных, содержащий фармакокинетику индометацина.

14. Theoph[35] – набор данных, содержащий фармакокинетику теофиллина.

15. lh[36] – набор данных, содержащий временной ряд изменения концентрации лютеинизирующего гормона в образцах крови.

16. women[37] – набор данных, содержащий средний рост и вес женщин в Америке в возрасте от 30 до 39 лет.

Весь процесс статистического анализа данных можно представить в виде алгоритма, изображенного на рисунке 2.

---

[30] Hosmer D. W., Lemeshow S. Applied Logistic Regression. New York: Wiley, 1989.

[31] Thall P. F., Vail S. C. Some covariance models for longitudinal count data with over-dispersion // Biometrics. 1990. Vol. 46, №3. P. 657–671.

[32] Cox D. R., Oakes D. Analysis of Survival Data. Chapman & Hall, 1984. P. 7. Taken from: Gehan E.A. A generalized Wilcoxon test for comparing arbitrarily single-censored samples // Biometrika. 1965. №52. P. 203–233.

[33] Linder A., Chakravarti I. M., Vuagnat P. Fitting asymptotic regression curves with different asymptotes. In Contributions to Statistics. Presented to Professor P. C. Mahalanobis on the occasion of his 70th birthday / ed. C. R. Rao. Oxford: Pergamon Press, 1964. P. 221–228.

[34] Kwan K. C., Breault G. O., Umbenhauer E. R., et al. Kinetics of Indomethacin absorption, elimination, and enterohepatic circulation in man // Journal of Pharmacokinetics and Biopharmaceutics. 1976. № 4. P. 255–280.

[35] Boeckmann A. J., Sheiner L. B., Beal S. L. Nonmem Users Guide. Part V. Nonmem Project Group; University of California, San Francisco, 1994.

[36] Diggle P. J. Time Series: A Biostatistical Introduction. Oxford, 1990. Table A.1, series 3.

[37] The World Almanac and Book of Facts, 1975.

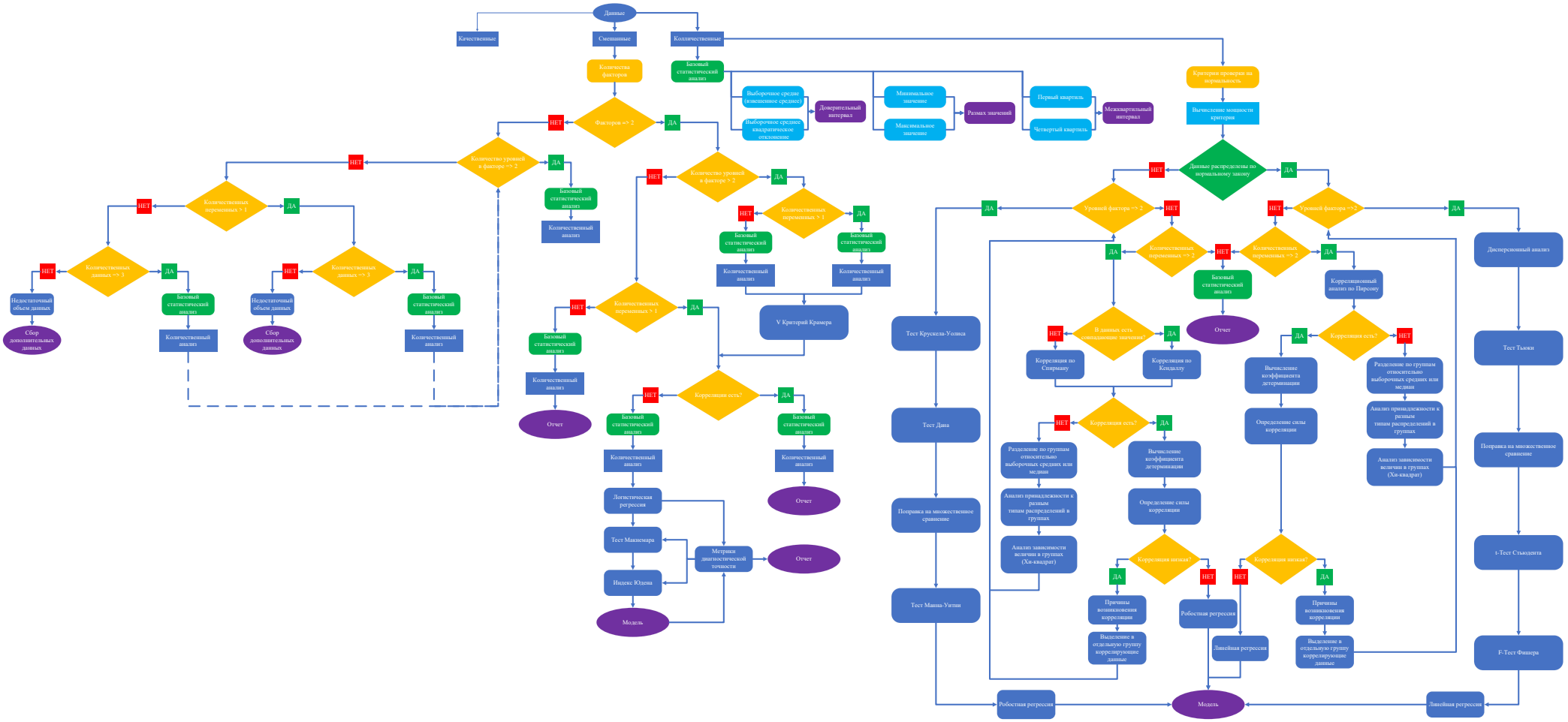


Рисунок 2 – Алгоритм проведения статистического анализа количественных и смешанных данных

Каждый из представленных в алгоритме тестов или методов рассматривается в данных методических рекомендациях. Стоит отметить, что перечень представленных методов не является исчерпывающим, и читатель может самостоятельно расширять и дополнять описанные методы.

Авторы методических рекомендаций выражают благодарность М. Р. Коденко и Р. В. Решетникову за ценные советы, данные при подготовке рукописи.

1. ТИПЫ ДАННЫХ

При проведении статистического анализа данных аналитик сталкивается с тремя основными типами данных:

1) количественными;
2) качественными;
3) смешанными (присутствуют и количественные, и качественные данные).

Количественные данные[38] – численные данные, имеющие шкалу измерения. Они могут быть разделены на дискретные и непрерывные.

Дискретные количественные данные – это числовые данные, полученные путем подсчета какой-либо величины.

Примером дискретных количественных данных является: вес пациента, измеряющийся в килограммах; рост пациента, измеряемый в сантиметрах; количество пациентов; концентрация различных веществ в крови или моче пациента и т.д. (результаты антропометрических, лабораторных и функциональных исследований пациентов).

Непрерывные количественные данные – изменение какой-либо количественной величины, измеренной за определенный период времени.

Примерами непрерывных количественных данных являются электроэнцефалограмма головного мозга, электрокардиограмма сердечной мышцы и др.

В таблице 1 представлен пример дискретных количественных данных.

Таблица 1 – Пример набора данных, содержащего только дискретные количественные данные

Возраст пациента, дней	Концентрация гликозаминогликанов в моче, мкг/мл
0,00	23,0
0,00	23,8
0,00	16,9
3,65	18,6
3,65	17,9

Качественные данные – это данные, описывающие признак предмета исследования естественным языком. Они могут быть номинальными и порядковыми.

Номинальные данные – это подгруппа качественных данных, используемая для именования переменных, не имеющих числового значения.

Примером качественного номинального признака предмета исследования являются пол пациента, наличие или отсутствие вредных привычек, занятие пациентом физической культурой и т.д.

Порядковые данные – это подгруппа качественных данных, имеющих порядок или масштаб. Часто качественные порядковые данные могут быть представлены числами, которым соответствует качественное описание. Примером порядковых качественных данных могут являться порядковый номер пациента, порядковый номер ответа на вопрос и т. д.

Наиболее распространенным примером качественных порядковых данных является ответ на вопрос по пяти-, десяти- и т. д. бальной шкале. Пример пятибалльной шкалы уверенности в каком-либо вопросе, применяющейся при проведении опросов:

1. Да.
2. Скорее да.
3. Затрудняюсь ответить.
4. Скорее нет.

[38] В языке программирования R количественные данные могут быть представлены типами numeric, integer, double (в языке R отсутствует тип данных с одинарной точностью).

5. Нет.

Каждому из пяти представленных ответов может соответствовать число от 1 до 5 в случае применения порядковой шкалы или от 0 до 1 в случае вероятностной шкалы. Выбор шкалы соответствия ограничивается только фантазией исследователя, и единственное требование, которое возникает при планировании сбора качественных данных – единообразие используемых шкал.

В языке программирования R качественное описание объекта, содержащееся в наборе данных, представляется как тип данных «фактор»[39], а значения, принимаемые переменной, имеющей тип «фактор», называются уровнями фактора. В таблице 2 представлен пример факторных данных.

Таблица 2 – Пример набора данных, содержащего только качественные данные

Идентификатор пациента	Пол пациента
1a	М
2a	М
3a	Ж
1b	М
2b	Ж

В данном случае идентификатор представлен численно буквенным кодом, идентифицирующим пациента в некоторой базе данных, а пол пациента обозначен буквами «М» (мужской) и «Ж» (женский).

В практике анализа и проведения эксперимента редко встречаются наборы данных, содержащие только количественные или только качественные данные. Чаще всего наборы данных имеют смешанный характер. В таблице 3 представлен пример смешанного набора данных.

Таблица 3 – Пример набора данных, содержащего качественные и количественные показатели пациента

Идентификатор пациента	Возраст пациента	Пол пациента	Концентрация гликозаминогликанов в моче, мкг/мл
1a2f	15,06	Ж	3,2
2a2f	15,15	М	4,2
3a2f	15,55	М	6,0
1b2f	15,72	Ж	9,7
2b2f	15,86	М	3,4

Фактически набор данных представляет собой совокупность переменных (название каждого столбца), каждая из которых описывает качественную или количественную характеристику исследуемого объекта или явления. Соответственно, переменные, содержащие только качественные значения, называются факторами. Создание наборов медицинских данных детально рассматривается в курсе «Создание наборов данных»[40].

Для каждого типа переменных или их совокупности существует свой набор статистических тестов, позволяющих проводить детальный анализ наличия или отсутствия различий между всевозможными уровнями факторов, наличие связи или

[39] Фактор (лат. factor «делающий, производящий») – причина, движущая сила какого-либо процесса, определяющая его характер или отдельные его черты. См.: Фактор // Большая советская энциклопедия: в 30 т. / гл. ред. А. М. Прохоров. 3-е изд. М.: Советская энциклопедия, 1969–1978.

[40] Васильев Ю. А., Арзамасов К. М., Владзимирский А. В. [и др.]. Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта: учебно-метод. пос. М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2023. 108 с.

отсутствие связей между данными, принадлежащими к разным уровням фактора, которые будут рассмотрены далее. Алгоритм анализа, представленный на рисунке 2, является верным для количественных и смешанных наборов данных. Методы анализа качественных данных не рассматриваются в настоящих методических рекомендациях.

Ниже представлены примеры наборов данных, содержащих смешанные данные, но с превалирующим количеством качественных переменных и с превалирующим числом количественных данных.

1.1. Пример количественных и качественных данных

В практике анализа наборы данных, содержащие только качественные переменные, возникают при сборе методом опроса или анкетирования (даже в этом случае результаты опроса или анкетирования, скорее всего, будут содержать возраст пациента и время (календарную дату) проведения опроса). То же самое относится и к наборам данных, содержащим только количественные переменные: как правило, присутствует порядковый номер образца, и/или идентификатор пациента, и/или гендерный признак пациента. Рассмотрим на примерах некоторые наборы данных, содержащиеся в пакетах языка R, в частности в пакете MASS**.** Знак «**#**» применяется для экранирования (компилятор языка не будет воспринимать текст, находящийся после данного знака, как текст программы) комментариев в тексте программы.

Листинг 1[41]

```
library(MASS) #Подключаем библиотеку, содержащую набор данных bacteria
head(bacteria) # выводим первую часть набора данных
#================================================================
=====
# Результат вывода первой части набора данных
#================================================================
=====
 y ap hilo week ID  trt
1 y p  hi  0 X01 placebo
2 y p  hi  2 X01 placebo
3 y p  hi  4 X01 placebo
4 y p  hi  11 X01 placebo
5 y a  hi  0 X02  drug+
6 y a  hi  2 X02  drug+
#================================================================
====
#Проводим определение структуры набора данных
#================================================================
====
str(bacteria) # Выводим структуру данных, содержащуюся в наборе данных
   # bacteria
#================================================================
====
# Результаты применения функции
#================================================================
====
'data.frame':    220 obs. of 6 variables:
```

[41] Здесь и далее во всем тексте методических рекомендаций примеры программного кода на языке R будут обозначены словом «Листинг» и иметь сквозную нумерацию.

Продолжение листинга 1

```
 $ y  : Factor w/ 2 levels "n","y": 2 2 2 2 2 2 1 2 2 2 ...
 $ ap : Factor w/ 2 levels "a","p": 2 2 2 2 1 1 1 1 1 1 ...
 $ hilo: Factor w/ 2 levels "hi","lo": 1 1 1 1 1 1 1 1 2 2 ...
 $ week: int 0 2 4 11 0 2 6 11 0 2 ...
 $ ID : Factor w/ 50 levels "X01","X02","X03",..: 1 1 1 1 2 2 2 2 3 3
...
 $ trt : Factor w/ 3 levels "placebo","drug",..: 1 1 1 1 3 3 3 3 2 2
...
```

Приведенный набор данных содержит в основном «факторы», его уровни представлены естественным языком. Таким представлением в языке R описываются качественные данные, а сам набор данных содержит только один количественный показатель (week), описывающий порядковый номер недели проведения исследования.

Листинг 2

```
library(MASS) #Подключаем пакет, содержащий набор данных muscle
head(muscle) #Выводим начало набора данных
#=============================================================
=====
#Результат применения команды
#=============================================================
=====
  Strip Conc Length
3  S01  1  15.8
4  S01  2  20.8
5  S01  3  22.6
6  S01  4  23.8
9  S02  1  20.6
10  S02  2  26.8
#=============================================================
=====
#Выводим структуру набора данных
#=============================================================
=====
str(muscle)
#=============================================================
=====
'data.frame':    60 obs. of 3 variables:
 $ Strip: Factor w/ 21 levels "S01","S02","S03",..: 1 1 1 1 2 2 2 2 3 3
...
 $ Conc: num 1 2 3 4 1 2 3 4 0.25 0.5 ...
 $ Length: num 15.8 20.8 22.6 23.8 20.6 26.8 28.4 27 7.2 15.4 ...
```

Набор данных ***muscle*** содержит значения концентрации хлорида кальция, кратные величине 2,2 ммоль (Conc), и длину полосы сокращения сердечной мышцы, измеряемую в миллиметрах (Length) – представленные данные являются количественными. Одновременно набор данных содержит и качественный параметр, а именно обозначение линии сердечной мышцы при проведении измерений (Strip).

2. НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Каждая переменная, входящая в набор данных, имеет совокупность значений, полученных в результате наблюдения или путем прямого измерения. Взаимная зависимость и влияние изменения одной величины на другую являются предметом исследования, в том числе методами статистического анализа.

Если на исследуемую величину действует большое количество независимых переменных и отсутствует возможность выделить доминирующую величину, то распределение исследуемой величины **стремится** к так называемому нормальному или Гауссову закону распределения (при проведении исследования данных выбор применяемого для анализа теста или критерия зависит от того, принадлежит ли исследуемая величина нормальному закону распределения или нет).

На рисунке 3 представлен пример графика (гистограммы[42]) случайной величины, значительно приближенной к нормальному закону распределения.

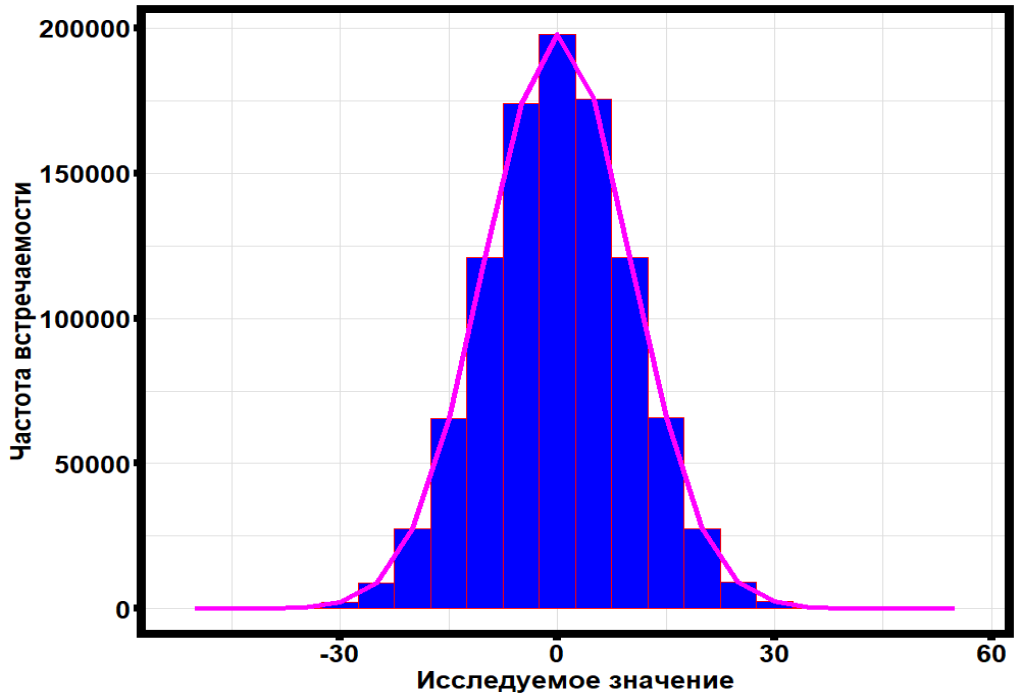


Рисунок 3 – Пример частоты встречаемости наблюдаемого и/или измеренного значения для нормального закона распределения

Опишем нахождение основных метрик и охарактеризуем их для данных, распределенных в соответствии с нормальным законом.

2.1. Базовый статистический анализ количественных данных

Базовый статистический анализ составляет основу любой аналитической работы с данными (независимо от того, какого типа данные анализируются).

В случае, если набор данных содержит в основном качественные данные, то при проведении базового статистического анализа определяют количество данных,

[42] Более подробное построение гистограмм распределения экспериментальных величин рассмотрено в разделе 4.

содержащихся в каждом факторе, количество уровней факторов и количество данных, содержащихся в каждом уровне. На основании этих данных вычисляются:

1. Доля каждого уровня в факторе.
2. Среднее квадратичное отклонение доли в факторе.

В случае, если набор данных содержит в основном количественные данные, то при проведении базового статистического анализа определяют:

1. Выборочное среднее и/или среднее взвешенное значение.
2. Медиану.
3. Среднее квадратическое отклонение.
4. Доверительный интервал выборочного среднего на основании первой и третьей величины.
5. Максимальное значение.
6. Минимальное значение;
7. «Размах» значений на основании максимального и минимального значений.
8. Первый квартиль в распределении данных.
9. Последний квартиль в распределении данных.
10. Межквартильный интервал на основании первого и последнего квартиля в распределении данных.

Перечисленные величины описывают основные статистические свойства исследуемых данных, но не дают ответа на вопрос о наличии различий в группах данных и их взаимном влиянии. Рассмотрим более детально способ вычисления величин, входящих в перечень базовых статистических величин.

2.1.1. Выборочное среднее[43]

При проведении анализа данных первая величина, которую вычисляет исследователь – это выборочное среднее значение. В данных методических рекомендациях не рассматривается вопрос математически строгого представления средних величин, а представлены наиболее часто использующиеся на практике. Выборочная средняя величина вычисляется по уравнению (1):

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad (1)$$

где N – количество исследований, вычисленное по одному параметру в наборе данных; X_i – фактическое значение усредняемой величины.

Она является обобщенной характеристикой в случае однородности данных, описывающей явления, имеющие одну и ту же размерность. Например, если врач проводит исследование веса или роста пациентов, возраст которых составляет 20 полных лет, то ему необходимо описать вес или рост всех пациентов в возрасте 20 полных лет. Для этих целей он будет использовать выборочное среднее значение. В случае данных, распределенных в соответствии с нормальным законом распределения, выборочное среднее значение будет соответствовать максимуму частоты встречаемости исследуемого значения, как представлено на рисунке 4.

[43] Колмогоров А. Н. Избранные труды. Математика и механика. М.: Наука, 1985.

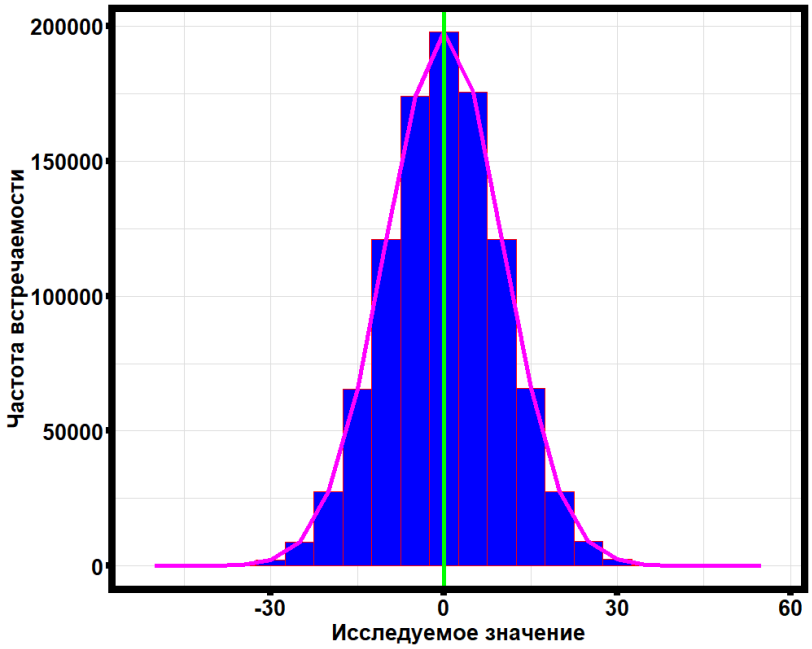


Рисунок 4 – Выборочное среднее (зеленая вертикальная линия) для нормального закона распределения[44]

На языке программирования R выборочное среднее значение вычисляется с помощью функции ***mean()*** из пакета **base.**

В случае, если врачу необходимо более точно описать средний вес пациентов (например) в возрасте 20 лет с учетом их (например) роста, то для этих целей нужно использовать среднее взвешенное значение (2):

$$\bar{X}_{\omega} = \frac{\sum_{i=1}^{N} \omega_i \cdot X_i}{\sum_{i=1}^{N} \omega_i}, \tag{2}$$

где N – количество исследований, вычисленное по одному параметру в наборе данных; X_i – фактическое значение усредняемой величины; ω_i – вес i-го значения усредняемой величины в общей выборке.

В качестве весового коэффициента ω_i в этом случае будет выступать величина, описывающая рост i-го пациента в возрасте 20 лет. Средняя взвешенная величина выступает в качестве характеристики, описывающей совокупность явлений, имеющих одну и туже размерность с учетом влияния сторонних признаков (например, вес пациентов
с учетом их роста).

На языке программирования R выборочное среднее значение вычисляется с помощью функции ***weighted.mean()*** из пакета **base**.

При исследованиях долей возникновения признака может возникнуть необходимость вычисления средней доли признака по отношению ко всем признакам. В этом случае среднее значение доли вычисляется как (3):

$$\bar{X}_e = \frac{p}{p+q}, \tag{3}$$

[44] Совпадает со значением, соответствующим максимуму частоты встречаемости признака или измеряемой величины в случае нормального закона распределения данных.

где p – доля единиц, обладающих исследуемым признаком; q – доля единиц, не обладающих исследуемым признаком, равна $1-p$.

Пример вычисления выборочного среднего и средневзвешенного на языке R

Для примера вычисления средних и средневзвешенных значений используем набор данных **anorexia** из пакета MASS**.** Вычислим среднее и средневзвешенное значение веса пациента. В качестве весового коэффициента (уравнение 2 и текст под ним) будем использовать долю пациентов в каждой группе.

Листинг 3

```
library(MASS)  #[45]Подключаем библиотеку, содержащую набор данных anorexia
#=========================================================================
#Исследуем структуру набора данных
#=========================================================================
str(anorexia)
#=========================================================================
# Выводим структуру данных
#=========================================================================
'data.frame':    72 obs. of 3 variables:
 $ Treat : Factor w/ 3 levels "CBT","Cont","FT": 2 2 2 2 2 2 2 2 2 2 ..
 $ Prewt : num 80.7 89.4 91.8 74 78.1 88.3 87.3 75.1 80.6 78.4 ...
 $ Postwt: num 80.2 80.1 86.4 86.3 76.1 78.1 75.1 86.7 73.5 84.6 ...
#=========================================================================
# Выводим заголовочную часть набора данных
#=========================================================================
head(anorexia)
#=========================================================================
# Выводим заголовочную часть файла
#=========================================================================
 Treat Prewt Postwt
1 Cont  80.7   80.2
2 Cont  89.4   80.1
3 Cont  91.8   86.4
4 Cont  74.0   86.3
5 Cont  78.1   76.1
6 Cont  88.3   78.1
#=========================================================================
#  Вычисляем выборочное среднее в контрольной группе пациентов до
# проведения лечения
#=========================================================================
mean(anorexia[anorexia$Treat=="Cont", ]$Prewt)
#=========================================================================
# результат вычисления среднего веса пациента в контрольной группе
#=========================================================================
[1] 81.55769
#=========================================================================
mean(anorexia[anorexia$Treat=="CBT",]$Prewt)
#=========================================================================
# результат вычисления среднего веса пациента в группе с
# когнитивно-поведенческой терапией
#=========================================================================
[1] 82.68966
#=========================================================================
mean(anorexia[anorexia$Treat=="FT",]$Prewt)
#=========================================================================
# результат вычисления среднего веса пациента в группе с семейной терапией
#=========================================================================
[1] 83.22941
```

[45] Знак экранирования однострочных комментариев.

Продолжение листинга 3

```
#=========================================================================
#Вычисление взвешенного среднего значения
#=========================================================================
# Определяем количество пациентов в каждой группе и общее количество пациентов
numInCont <- length[46](anorexia[anorexia$Treat=="Cont",]$Prewt)

numInCBT <- length(anorexia[anorexia$Treat=="CBT",]$Prewt)

numInFT <- length(anorexia[anorexia$Treat=="FT",]$Prewt)

numAll <- length(anorexia$Prewt)
#=========================================================================
# Формируем вектор весов и вектор данных
# В данном примере в качестве веса выступает доля пациентов в каждой
# группе
#=========================================================================
weightData <- c(numInCont/numAll, numInCBT/numAll, numInFT/numAll )
meanData <- c(mean(anorexia[anorexia$Treat=="Cont",]$Prewt),
 mean(anorexia[anorexia$Treat=="CBT",]$Prewt),
 mean(anorexia[anorexia$Treat=="FT",]$Prewt)
weighted.mean(meanData, weightData)
#=========================================================================
# Результат вычисления средневзвешенного значения веса пациента
#=========================================================================
[1] 82.40833
```

Полученные значения среднего и средневзвешенного веса пациентов отражают типичный уровень веса пациента в исследуемой группе, формирующийся под воздействием доминирующих неслучайных параметров. Установление неслучайных параметров, влияющих (в данном случае на вес пациента) на среднее значение, требует дальнейшего исследования данных с использованием соответствующих статистических критериев, в том числе критериев, позволяющих сравнивать средние величины (например, t-критерий Стьюдента). Оценка среднего отклонения отдельных значений (например, веса пациента) от среднего значения описывается с помощью среднего квадратического отклонения.

2.1.2. Вычисление среднего квадратического отклонения

Среднее квадратическое отклонение является статистической характеристикой, показывающей степень близости отдельных значений к средней величине (чем меньше среднее квадратическое отклонение, тем лучше средняя величина описывает исследуемую переменную). Среднее квадратическое отклонение генеральной совокупности значений вычисляется как (4):

[46] Функция ***length()*** предназначена для определения длины вектора, более подробно см. справку в RStudio - ***?length.***

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\,, \tag{4}$$

где N – полное количество исследований в генеральной совокупности; $\bar{x}$ – среднее значение исследуемой величины; x_i – фактическое значение, полученное при измерении величины.

В большинстве случаев исследователь имеет дело не с генеральной совокупностью (всей популяцией), а с выборочной ее частью. Среднее квадратическое отклонение для выборки из генеральной совокупности вычисляется как (5):

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5}$$

Оценки верны для случаев, когда исследуемые переменные имеют количественную природу, и данные принадлежат к нормальному закону распределения. Иллюстрация среднего квадратического отклонения представлена на рисунке 5.

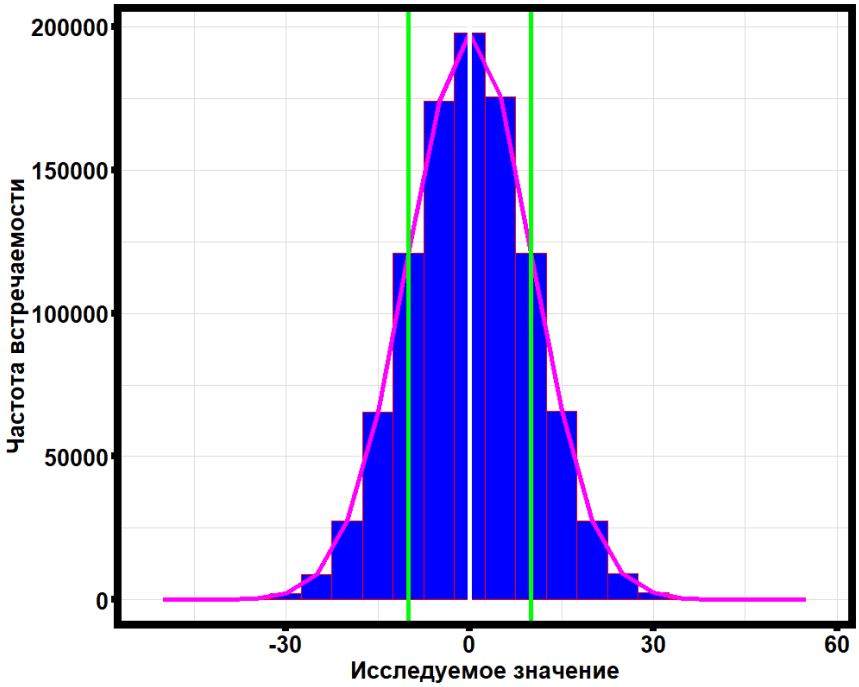


Рисунок 5 – Белой вертикальной линией отмечено выборочное среднее значение данных, распределенных по нормальному (Гауссову) закону. Расстояние от белой вертикальной линии до зеленой (влево или вправо) задается средним квадратическим отклонением

В случае, если исследователю необходимо оценить среднее квадратическое отклонение вероятности возникновения признака, то составляется уравнение (6):

$$S_e = \sqrt{\hat{p} \cdot (1 - \hat{p})}, \tag{6}$$

где $\hat{p}$ – вероятность возникновения исследуемого признака:

$$\hat{p} = \frac{n}{N}, \quad (7)$$

где N – общее количество значений, обладающих различными признаками; n – количество исследований, обладающих анализируемым признаком.

На языке программирования R для вычисления среднего квадратического значения применяется функция ***sd()*** из пакета **stats**. В случае данных, распределенных по нормальному закону, выборочное среднее квадратическое отклонение будет соответствовать рисунку 5.

Пример вычисления среднего квадратического отклонения

Для примера проведения вычислений среднего квадратического отклонения используем набор данных ***anorexia*** из пакета MASS.

Листинг 4

```
library(MASS) #Подключаем пакет MASS, содержащий набор данных anorexia
#=================================================================
#Выделяем из набора данных подгруппы пациентов с разным типом лечения
#=================================================================
contData <- anorexia[anorexia$Treat=="Cont",]$Prewt
cbtData <- anorexia[anorexia$Treat=="CBT",]$Prewt
ftData <- anorexia[anorexia$Treat=="FT",]$Prewt
#=================================================================
#Вычисляем среднее квадратическое отклонение веса пациента до
# прохождения лечения
#=================================================================
sd(contData)
#=================================================================
# Результат вычисления в контрольной группе
#=================================================================
[1] 5.70706
#=================================================================
sd(cbtData)
#=================================================================
# Результат вычисления в группе с когнитивно-поведенческой терапией
#=================================================================
[1] 4.845495
#=================================================================
sd(ftData)
#=================================================================
# Результаты вычислений в группе с семейной терапией
#=================================================================
[1] 5.016693
#=================================================================
#  Вычисляем количество пациентов до прохождения лечения
#=================================================================
lenPrewt <- length(anorexia$Prewt) #Общее количество данных
lenContPrew <- length(anorexia[anorexia$Treat=="Cont",]$Prewt) # Количество
# данных в контрольной группе
lenCBTPrewt <- length(anorexia[anorexia$Treat=="CBT",]$Prewt) # Количество
#  данных в группе с когнитивно-поведенческой терапией
lenBTPrew <- length(anorexia[anorexia$Treat=="FT",]$Prewt) # Количество
#  данных в группе с семейной терапией
#=================================================================
#  Вычисляем доли пациентов в каждой группе
#=================================================================
```

Продолжение листинга 4

```
pCP <- lenContPrew/ lenPrewt # Доля в контрольной группе
pCBTP <- lenCBTPrewt/ lenPrewt # Доля в когнитивно-поведенческой группе
pBTP <- lenBTPrew/lenPrewt # Доля в семейной группе
#=================================================================

sdCP <- sqrt[47](pCP*(1-pCP)) # Среднее квадратичное отклонение долей в
# контрольной группе.
sdCBT <- sqrt(pCBTP*(1- pCBTP))
sdBT <- sqrt(pBTP*(1- pBTP))
#=================================================================
#  Результаты вычисления
#=================================================================
print(sdCP)
#=================================================================
[1] 0.4803227
#=================================================================
print(sdCBT)
#=================================================================
[1] 0.4904568
#=================================================================
print(sdBT)
#=================================================================
[1] 0.4246912
```

2.1.3. Вычисление доверительного интервала

Вычислив значения выборочной средней величины (или средневзвешенной величины)[48] и среднего квадратического отклонения, можно построить доверительный интервал средней величины. Вычисление доверительного интервала для средней величины тесно связано с понятием доверительной вероятности, ошибки средней величины и предельной ошибкой выборки.

Доверительная вероятность определяет степень уверенности того факта, что измеренная величина находится в близи среднего значения. Средняя ошибка выборки показывает объективно возникающее расхождение между характеристиками выборки и генеральной совокупностью, а предельная ошибка выборки – это ошибка выборки, исчисляемая с заданной степенью вероятности. Ошибка средней величины в выборке для количественного признака вычисляется по уравнению (8):

$$\Delta = \frac{S}{\sqrt{N}}, \tag{8}$$

где S – выборочное среднее квадратическое отклонение; N – количество исследований.

Ошибка средней величины для долей вхождения признака в данные также вычисляется по уравнению (8) с той лишь разницей, что вместо выборочного среднего квадратического отклонения в уравнение подставляется среднее квадратическое отклонение долей исследуемого признака (уравнение 5).

Предельная ошибка выборки вычисляется по уравнению (9):

---

[47] Функция **sqrt()** используется для вычисления квадратного корня из количественной величины.

[48] См.: подраздел 2.1.1. Выборочное среднее.

$$\Delta_{lim} = t * \Delta\,, \tag{9}$$

где значение *t* представлено в таблице 4.

Таблица 4 – Значение коэффициента *t* и доверительная вероятность для нормального закона распределения данных

Значение коэффициента *t*	Доверительная вероятность *P(t)*
1	68,3
1,96	95,0
2	95,5
2,58	99,0
3	99,7

В медицинских исследованиях доверительный интервал принимается равным (10)[49]:

$$\bar{X} \mp 2 \cdot \Delta_{lim}\,, \tag{10}$$

где $\bar{X}$ – среднее значение исследуемой величины; Δ_{lim} – предельная ошибка выборки исследуемой величины.

В случае оценки доли фактора и отклонения долей, величина доверительного интервала вычисляется по уравнению (11):

$$\bar{X}_e \mp 2 \cdot \Delta^e_{\text{lim,}} \tag{11}$$

где $\bar{X}_e$ – средняя доля исследуемого признака, определенная по уравнению (3); Δ^e_{lim}– предельная ошибка выборки долей вхождения признака в данные.

На рисунке 6 изображено графическое представление доверительного интервала при условии данных, распределенных в соответствии с нормальным (Гауссовым) законом.

[49] Херцог М., Френсис Г., Кларк А. [и др.]. Статистика и планирование эксперимента для непосвященных: как отучить статистику лгать. М.: ДМК Пресс, 2023. 174 с.

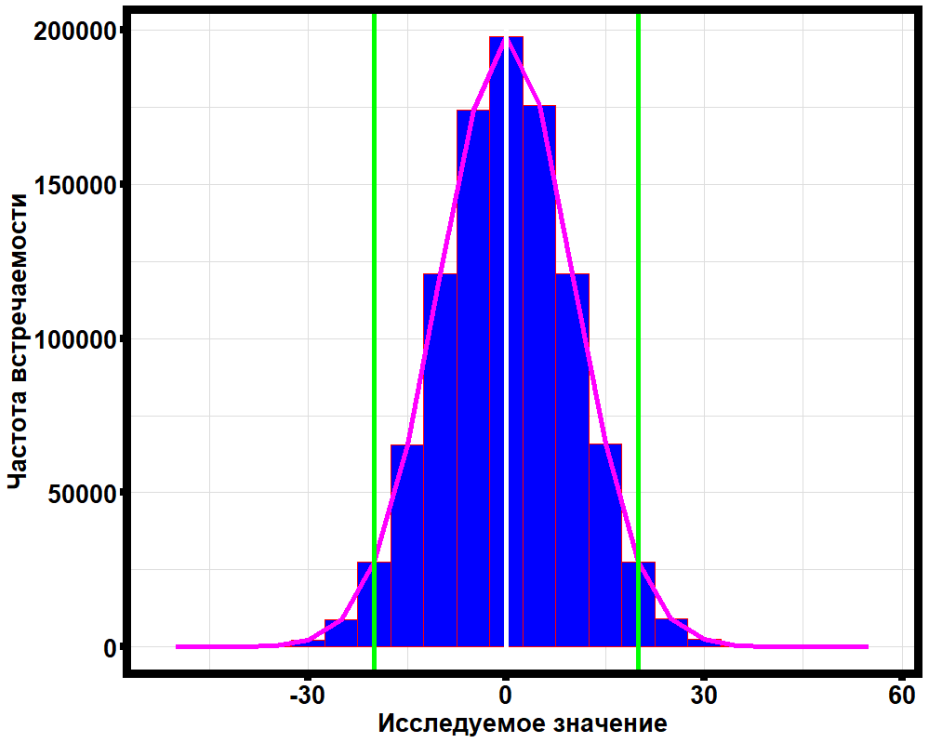


Рисунок 6 – Доверительный интервал (уравнение 10) выборочного среднего значения (расстояние от белой вертикальной линии до зеленой вертикальной линии)

Пример вычисления доверительных интервалов

На примере значений веса пациентов с подтвержденным диагнозом «анорексия» (набор данных ***anorexia*** пакета MASS) проведем анализ доверительных интервалов весов и долей пациентов в контрольной группе, группе с когнитивно-поведенческой терапией и в группе с назначенной семейной терапией.

Листинг 5

```
library(MASS) # Подключаем библиотеку, содержащую набор данных anorexia
library(ggplot2)
#=============================================================
#  Выделяем из набора данных подгруппы пациентов с разным типом лечения
#=============================================================
contData <- anorexia[anorexia$Treat=="Cont",]$Prewt
cbtData <- anorexia[anorexia$Treat=="CBT",]$Prewt
ftData <- anorexia[anorexia$Treat=="FT",]$Prewt
#=============================================================
#  Вычисляем средние значения
#=============================================================
meanCont <- mean(contData) # Средний вес в контрольной группе
meanCBT <- mean(cbtData) # Средний вес в когнитивно-поведенческой группе
meanFT <- mean(ftData) # Средний вес в группе семейной терапии
#=============================================================
# Вычисляем среднее квадратическое отклонение
#=============================================================
sdCont <- sd(contData)/(sqrt(length(contData))) # Среднее квадратическое
# отклонение веса в контрольной группе
sdCBT <- sd(cbtData)/(sqrt(length(cbtData))) # Среднее квадратическое
отклонение
#  веса в когнитивно-поведенческой группе
```

Продолжение листинга 5

```
sdFT <- sd(ftData)/(sqrt(length(ftData))) # Среднее квадратическое отклонение
# веса в группе семейной терапии
#=====================================================================
# Вычисление доверительных интервалов в группах
#=====================================================================
upCont <- meanCont+2*sdCont # Верхняя граница доверительного интервала в
# контрольной группе
downCont <- meanCont-2*sdCont # Нижняя граница доверительного интервала в
# контрольной группе
#=====================================================================
upCBT <- meanCBT+2*sdCBT # Верхняя граница доверительного интервала в
# группе с когнитивно-поведенческой терапией
downCBT <- meanCBT-2*sdCBT # Нижняя граница доверительного интервала в
# группе с когнитивно-поведенческой терапией
#=====================================================================
upFT <- meanFT+2*sdFT # Верхняя граница доверительного интервала в
# группе с семейной терапией
downFT <- meanFT-2*sdFT # Нижняя граница доверительного интервала в
# группе с семейной терапией
#=====================================================================
#  Построение графика средних значений и доверительных интервалов
#=====================================================================
dataAn <- data.frame(Type = c("Cont", "CBT", "FT"), Mean=c(meanCont,
meanCBT, meanFT), Sd = c(sdCont, sdCBT, sdFT))
#=====================================================================
grPP <- ggplot(data = dataAn, mapping = aes(Type, Mean))
grPP <- grPP + geom_point(mapping = aes(Type, Mean),colour="blue",size=5)
grPP <- grPP + geom_errorbar(mapping = aes(ymin=Mean-2*Sd, ymax=Mean+2*Sd),
colour="red",linewidth=1.5,width = 0.5)
grPP <- grPP + theme_light()
grPP <- grPP + theme(panel.border = element_rect(linewidth = 6, colour = "black"),
      axis.title = element_text(size = 20, face = "bold",colour = "black"),
      axis.text.x = element_text(size = 20, face = "bold",colour = "black"),
      axis.text.y = element_text(size = 20, face = "bold",colour = "black"),
      axis.ticks = element_line(linewidth = 3, colour = "black"),
      legend.title = element_text(size = 20, face = "bold", color = "black"),
      legend.text = element_text(size = 20, face = "bold", color = "black"))
grPP <- grPP + labs(x="Therapy type", y="Weight, kg")
print(grPP)
#=====================================================================
#  Вывод графика
#=====================================================================
```

На рисунке 7 изображено графическое представление средних значений веса с доверительными интервалами трех исследуемых групп пациентов до проведения терапии.

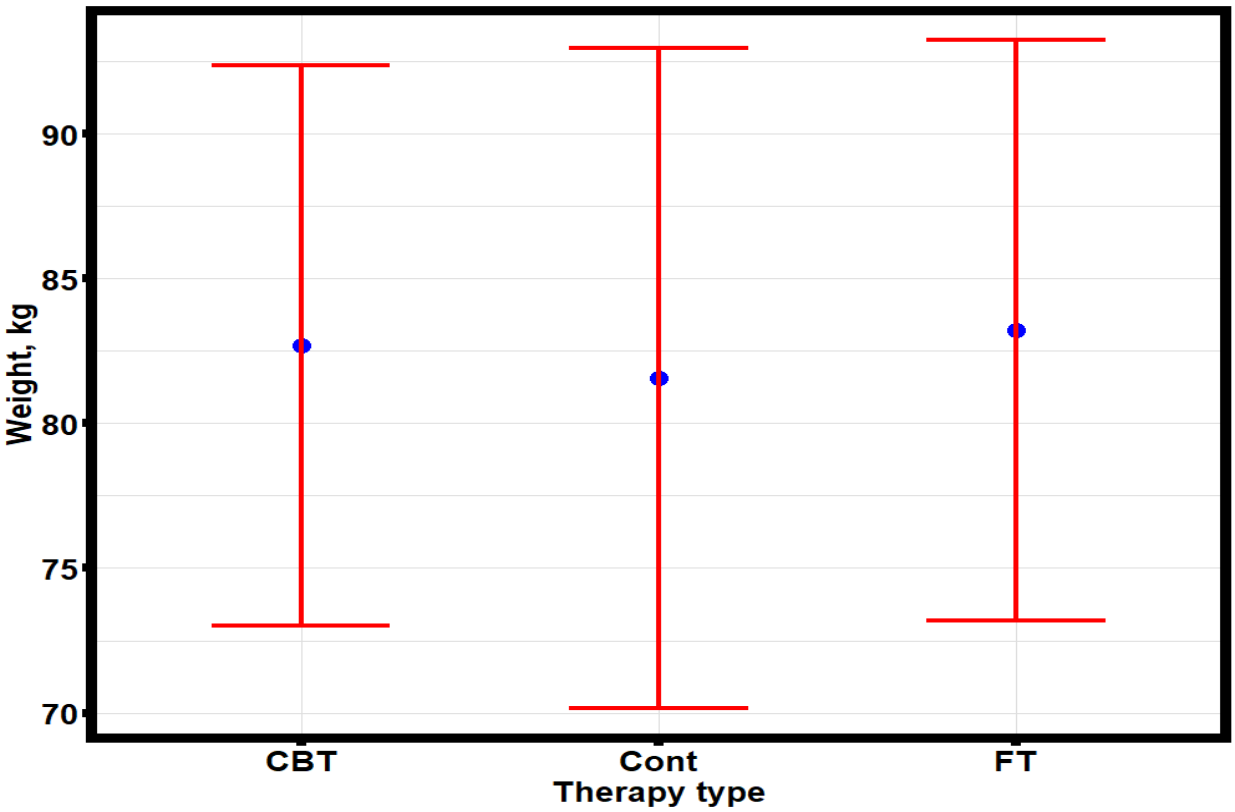


Рисунок 7 – Средние значения (синие кружочки) и доверительный интервал для доверительной вероятности 95,5 % (красные линии, нижние и верхние горизонтальные линии обозначают границы доверительных интервалов) веса пациентов в трех группах (Cont – контрольной группе; CBT – группе с назначенной когнитивно-поведенческой терапией; FT – с назначенной семейной терапией) до проведения терапии

На примере того же набора данных продемонстрируем вычисление вероятности нахождения пациента с определенным весом в определенной группе и вычислим доверительный интервал для этой вероятности.

Листинг 6

```
# Вычисляем количество пациентов до прохождения лечения
#=================================================================
lenPrewt <- length(anorexia$Prewt) #Общее количество данных
lenContPrew <- length(anorexia[anorexia$Treat=="Cont",]$Prewt) # Количество
#  данных в контрольной группе
lenCBTPrewt <- length(anorexia[anorexia$Treat=="CBT",]$Prewt) # Количество
# данных в группе с когнитивно-поведенческой терапией
lenBTPrew <- length(anorexia[anorexia$Treat=="FT",]$Prewt) # Количество
# данных в группе с семейной терапией
#=================================================================
# Вычисляем доли пациентов в каждой группе
#=================================================================
pCP <- lenContPrew/ lenPrewt # Доля в контрольной группе
pCBTP <- lenCBTPrewt/ lenPrewt # Доля в когнитивно-поведенческой группе
pBTP <- lenBTPrew/lenPrewt # Доля в семейной группе
#=================================================================
sdCP <- sqrt (pCP*(1-pCP)) # Среднее квадратичное отклонение долей в
# контрольной группе
sdCBT <- sqrt(pCBTP*(1- pCBTP))
```

Продолжение листинга 6

```
sdBT <- sqrt(pBTP*(1- pBTP))
#=================================================================
# Результаты вычисления
#=================================================================
dataAn <- data.frame(Type = c("Cont", "CBT", "FT"), Mean=c(pCP, pCBTP, pBTP),
        Sd = c(sdCP, sdCBT, sdBT))
#=================================================================
grPP <- ggplot(data = dataAn, mapping = aes(Type, Mean))
grPP <- grPP + geom_point(mapping = aes(Type, Mean),colour="blue",size=5)
grPP <- grPP + geom_errorbar(mapping = aes(ymin=Mean-2*Sd, ymax=Mean+2*Sd),
colour="red",linewidth=1.5,width = 0.5)
grPP <- grPP + theme_light()
grPP <- grPP + theme(panel.border = element_rect(linewidth = 6, colour = "black"),
        axis.title = element_text(size = 20, face = "bold",colour = "black"),
        axis.text.x = element_text(size = 20, face = "bold",colour = "black"),
        axis.text.y = element_text(size = 20, face = "bold",colour = "black"),
        axis.ticks = element_line(linewidth = 3, colour = "black"),
        legend.title = element_text(size = 20, face = "bold", color = "black"),
        legend.text = element_text(size = 20, face = "bold", color = "black"))
grPP <- grPP + labs(x="Therapy type", y="Proportion of patients")
print(grPP)
#=================================================================
#  Вывод графика
#=================================================================
```

На рисунке 8 графически изображены средние значения долей пациентов со средними квадратичными отклонениями в каждой из исследуемых групп пациентов до прохождения лечения.

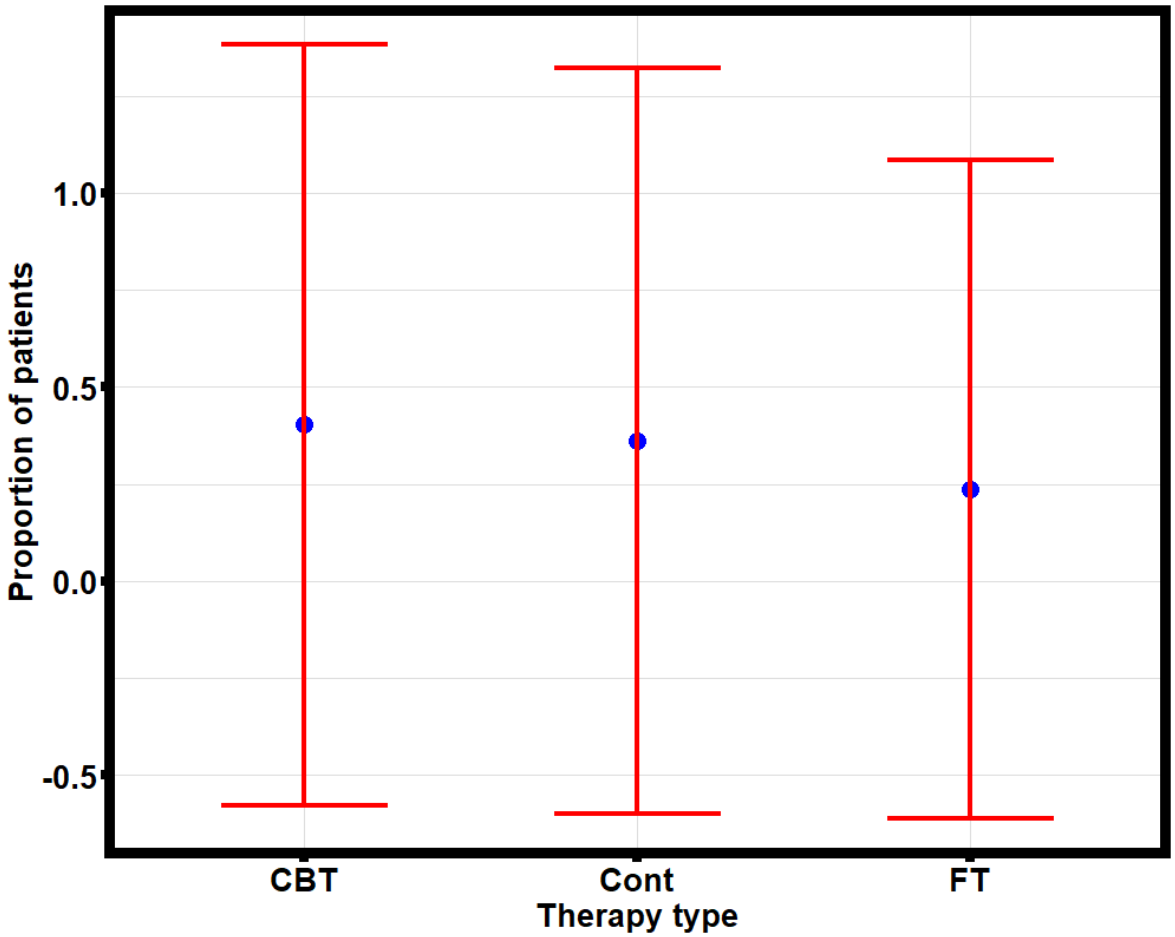


Рисунок 8 – Доля пациентов в каждой группе (синяя точка) и доверительный интервал для 95,5 % доверительной вероятности (красные линии обозначают ширину доверительного интервала, горизонтальными линиями обозначены границы доверительного интервала) доли пациентов в каждой группе пациентов, с назначенными различными типами терапии (Cont – контрольной группе; CBT – группе с назначенной когнитивно-поведенческой терапией; FT – с назначенной семейной терапией)

Сравнение доверительных интервалов и долей показывает, что все три группы достаточно близки друг к другу, что позволяет выдвинуть гипотезу об отсутствии статистически значимых различий. Данная гипотеза требует дальнейшего подтверждения или опровержения посредством статистических критериев.

2.1.4. Поиск максимального, минимального значения и размах

Анализ максимального значения, минимального значения и размаха количественной величины дополняет анализ средних значений и доверительных интервалов в части оценки фактического разброса значений. Рассмотрим оценку максимального, минимального значений и размаха количественной величины на примере анализа веса пациентов с диагнозом «анорексия» в трех группах с различным видом назначенного лечения.

Листинг 7

```
library(MASS) #Подключаем библиотеку, содержащую набор данных anorexia
library(ggplot2) #Подключаем библиотеку построения графиков
#=============================================================
# Выделяем из набора данных подгруппы пациентов с разным типом лечения
#=============================================================
contData <- anorexia[anorexia$Treat=="Cont",]$Prewt
cbtData <- anorexia[anorexia$Treat=="CBT",]$Prewt
ftData <- anorexia[anorexia$Treat=="FT",]$Prewt
#=============================================================
# Определяем максимальные значения в данных
#=============================================================
maxCont <- max[50](contData) # Максимальный вес до проведения лечения
# в контрольной группе пациентов
maxCBT <- max(cbtData) # Максимальный вес до проведения лечения
# в группе с назначенной когнитивно-поведенческой терапией
maxFT <- max(ftData) # Максимальный вес до проведения лечения
# в группе с назначенной семейной терапией
#=============================================================
# Определяем минимальные значения в данных
#=============================================================
minCont <- min[51](contData) # Минимальный вес пациентов до прохождения лечения
# в контрольной группе
minCBT <- min(cbtData) # Минимальный вес пациентов до прохождения лечения
# в группе с назначенной когнитивно-поведенческой терапией
minFT <- min(ftData) # Минимальный вес пациентов до прохождения лечения
# в группе с назначенной семейной терапией
#=============================================================
# Определяем размах значений веса в каждой группе
#=============================================================
deltaCont <- maxCont – minCont # Размах значений веса пациентов до прохождения
# лечения в контрольной группе
deltaCBT <- maxCBT – minCBT # Размах значений веса пациентов до прохождения
#  лечения в группе с назначенной когнитивно-поведенческой терапией
deltaFT <- maxFT – minFT # Размах значений веса пациентов до прохождения
# лечения в группе с назначенной семейной терапией
#=============================================================
#  Группируем результаты вычислений
#=============================================================
dataAn <- data.frame(Type = c("Cont", "CBT", "FT"),
                     Max=c(maxCont, maxCBT,maxFT),
                     Min = c(minCont, minCBT, minFT),
                     Delta = c(deltaCont, deltaCBT, deltaFT))
#=============================================================
grPP <- ggplot(data = dataAn, aes(Type,Max))
grPP <- grPP + geom_point(mapping = aes(Type, Min),colour="magenta", size=7)
grPP <- grPP + geom_point(mapping = aes(Type, Max),colour="blue", size=7)
grPP <- grPP + geom_errorbar(mapping = aes(ymin=Min, ymax=Max),
colour="red",linewidth=1.5,width = 0.5)
grPP <- grPP + theme_light()
grPP <- grPP + theme(panel.border = element_rect(linewidth = 6, colour = "black"),
      axis.title = element_text(size = 20, face = "bold",colour = "black"),
      axis.text.x = element_text(size = 20, face = "bold",colour = "black"),
      axis.text.y = element_text(size = 20, face = "bold",colour = "black"),
```

[50] Функция max() – определение максимального значения в числовом ряду.

[51] Функция min() – определение минимального значения в числовом ряду.

Продолжение листинга 7

```
axis.ticks = element_line(linewidth = 3, colour = "black"),
        legend.title = element_text(size = 20, face = "bold", color = "black"),
        legend.text = element_text(size = 20, face = "bold", color = "black"))
grPP <- grPP + labs(x="Therapy type", y="Value range")
print(grPP)
```

На рисунке 9 графически изображены размах значений массы тела в каждой из исследуемых групп пациентов до прохождения лечения.

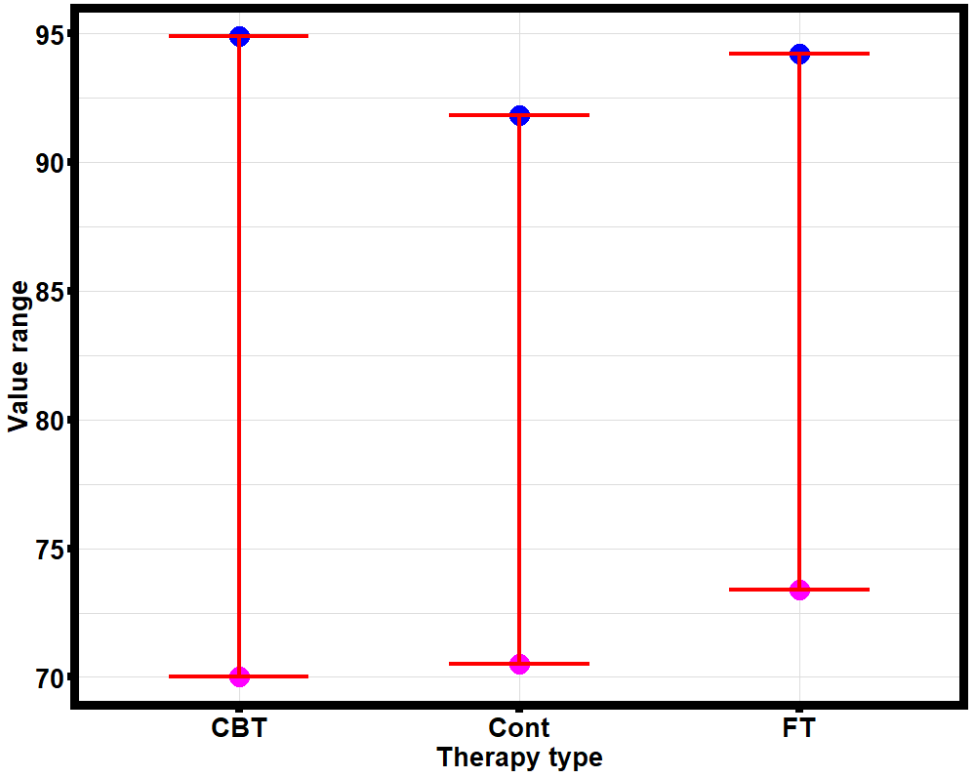


Рисунок 9 – Размах значений веса пациентов (синие закрашенные кружочки – максимальный вес; фиолетовые кружочки – минимальный вес пациентов) до проведения лечения в трех группах (CBT – группа с назначенным когнитивно-поведенческим лечением; Cont – контрольная группа; FT – назначенная семейная терапия)

Сравнение размаха значений и доверительных интервалов позволяет определить наличие выбросов в измерениях. Значения выше или ниже доверительного интервала должны быть проанализированы отдельно от основной выборки и не анализироваться в основной совокупности данных.

2.1.5. Понятие о квантилях, децилях, квартилях распределения

При проведении статистического анализа данных возникает необходимость оценки значения, соответствующего 5 %[52] от общего количества данных или 95 % от общего количества данных. Для этих целей применяются различные уровни квантилей.

Квантили – это значения, которые делят упорядоченную выборку на равные доли.

[52] Значения в процентах выбраны произвольно, чаще это 25 % и 75 %.

Допустим, имеется произвольный ряд десятичных чисел:

```
85.50 99.69 37.19 47.43 40.26 93.58 68.97 19.50 87.39 43.81 42.96
6.985.22 62.67 47.16 76.30 69.56 59.40 92.33 31.22 41.59 33.27 30.25
36.7749.43 33.85 27.44 8.45 99.84 52.81
```

Нам необходимо определить, какое значение соответствует 5 % в представленном ряду значений. Для решения данной задачи необходимо выстроить числа ряда в порядке возрастания (от меньшего к большему):

```
5.21 6.98 8.45 19.50 27.44 30.25 31.22 33.27 33.85 36.77 37.18
40.26 41.59 42.96 43.81 47.16 47.43 49.43 52.81 59.40 62.67 68.97
69.56 76.30 85.50 87.39 92.33 93.58 99.69 99.84
```

В данном ряду содержится 30 значений, номер значения, соответствующего 5 % данного ряда, равен 1.5, т.е. между первым и вторым значением соответственно квантиль будет равен:

$$\frac{5.21 + 6.98}{2} = 6.095$$

Децили – значения, которые делят упорядоченную выборку на десять примерно равных частей. Допустим, необходимо разделить ряд, представленный выше, на 10 равных частей с шагом в 10 %, тогда децили будут равны (процедура поиска значения, соответствующего каждому проценту, такая же, как в описанном выше примере):

10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
8.45	30.25	33.85	40.26	43.81	49.43	62.67	76.30	92.33	99.84

Наиболее часто применяемыми на практике являются **квартили – значения, которые делят упорядоченную выборку на четыре примерно равные части.** Для приведенного ранее ряда значения квартилей равны:

25 %	50 %	75 %	100 %
32.24	43.81	69.27	99.84

На языке программирования R вычисление квартилей проводится с помощью функции ***quantile(),*** входящей в пакет **stats**. На практике наиболее часто применяются первый и третий квартиль, межквартильный интервал представлен на рисунке 10.

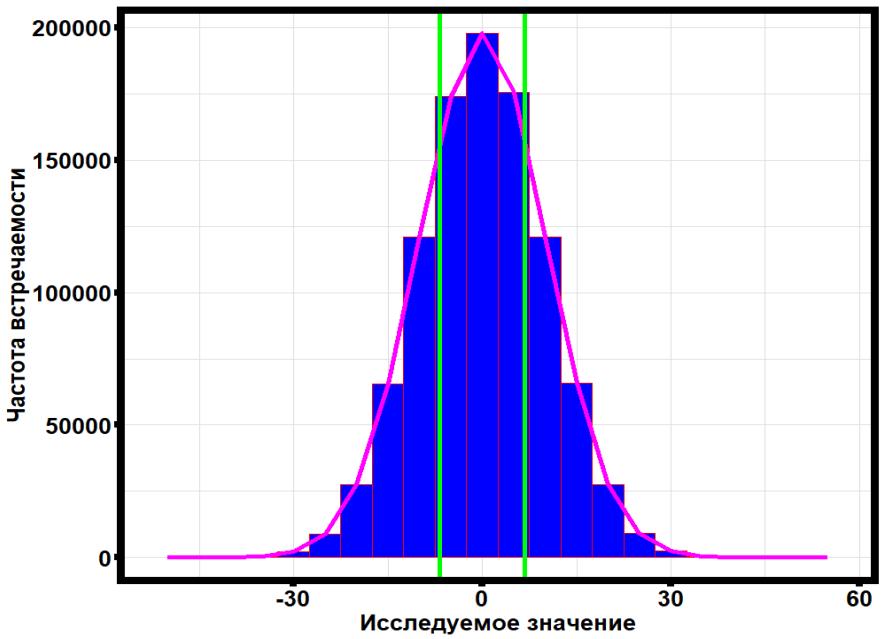


Рисунок 10 – Межквартильный интервал (между первым и третьим квартилем) представлен зелеными вертикальными линиями

В межквартильном интервале находятся 50 % всех встречаемых значений измеряемой величины. Первый квартиль ограничивает слева 25 % (первая зеленая вертикальная линия при просмотре рисунка 10 слева направо) значений, а третий – 75 % (вторая зеленая вертикальная линия при просмотре рисунка 10 слева направо).

Пример вычисления квантилей, децилей и квартилей

На примере набора данных ***anorexia*** из пакета MASS, содержащих информацию о весе пациентов с подтвержденным диагнозом «анорексия», рассмотрим вычисления квантилей, децилей и квартилей массы тела пациентов в контрольной группе, в группе с назначенной когнитивно-поведенческой терапией и семейной терапией.

Листинг 8

```
library(MASS) # Подключаем библиотеку, содержащую набор данных anorexia
library(ggplot2) #Подключаем библиотеку построения графиков
#=============================================================
# Выделяем из набора данных подгруппы пациентов с разным типом
# назначенного лечения
#=============================================================
contData <- anorexia[anorexia$Treat=="Cont",]$Prewt #Вес пациентов в
# контрольной группе до проведения лечения
cbtData <- anorexia[anorexia$Treat=="CBT",]$Prewt # Вес пациентов в группе
# с назначенным когнитивно-поведенческим лечением до проведения лечения
ftData <- anorexia[anorexia$Treat=="FT",]$Prewt # Вес пациентов в группе
# с назначенной семейной терапией
#=============================================================
#Задаем значения интересующих нас квантилей
propQ <- c(0.05, 0.95)
propDec <- c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)
propQvar <- c(0.25, 0.5, 0.75, 1.0)
#=============================================================
```

Продолжение листинга 8

```
contQuan <- quantile(contData, propQ, type = 4)
print(contQuan)
cbtQuan <- quantile(cbtData, propQ, type = 4)
print(cbtQuan)
ftQuan <- quantile(ftData, propQ, type = 4)
print(ftQuan)
```

2.1.6. Вычисление медианы

Выборочное среднее значение при большом объеме данных, распределенных в соответствии с нормальным законом значений, стремится к моде[53] или наиболее вероятному значению в данной выборке. Такая оценка значений не всегда бывает адекватной, особенно в случае наличия аномалий (значений сильно меньше или сильно больше большинства значений из выборочной совокупности). Для более адекватного описания всей совокупности при наличии выбросов или данных, распределенных отлично от нормального закона распределения, лучше использовать медианную оценку или медиану – это значение, которое разделяет ранжированную по возрастанию выборку пополам (50 % квантиль).

Предположим, имеется упорядоченный по возрастанию ряд чисел 1, 4, 6, 9, 11. Центр этого ряда составляет число 6, это и будет медианой данного ряда. В случае, если ряд имеет четное число значений, то медианой будет являться среднее арифметическое значение между двумя центральными значениями. Например, имеется ряд чисел, упорядоченных по возрастанию 1, 4, 6, 9, 11, 12 – двумя центральными значениями будут являться 6 и 9, их среднее арифметическое значение равно 7,5 – оно и будет являться медианой.

Пример вычисления медианы

Рассмотрим вычисление медианы на наборе данных ***anorexia*** из пакета MASS. Рассчитаем медиану веса пациентов до прохождения лечения в контрольной группе, группе пациентов с предписанной когнитивно-поведенческой терапией и в группе с семейной терапией.

Листинг 9

```
library(MASS) # Подключаем пакет, содержащий набор данных anorexia
#Вычисляем медиану в контрольной группе пациентов
median(anorexia[anorexia$Treat=="Cont",]$Prewt)
#================================================================
# Результат вычисления
#================================================================
[1] 80.65
#Вычисляем медиану в группе с когнитивно-поведенческой терапией
median(anorexia[anorexia$Treat=="CBT",]$Prewt)
#================================================================
#Результат вычисления
#================================================================
[1] 82.6
```

[53] Значение исследуемой переменной, которое встречается наиболее часто в исследуемых данных.

Продолжение листинга 9

```
#Вычисляем медиану в группе с семейной терапией
median(anorexia[anorexia$Treat=="FT",]$Prewt)
#=============================================================
#Результат вычисления
#=============================================================
[1] 83.3
```

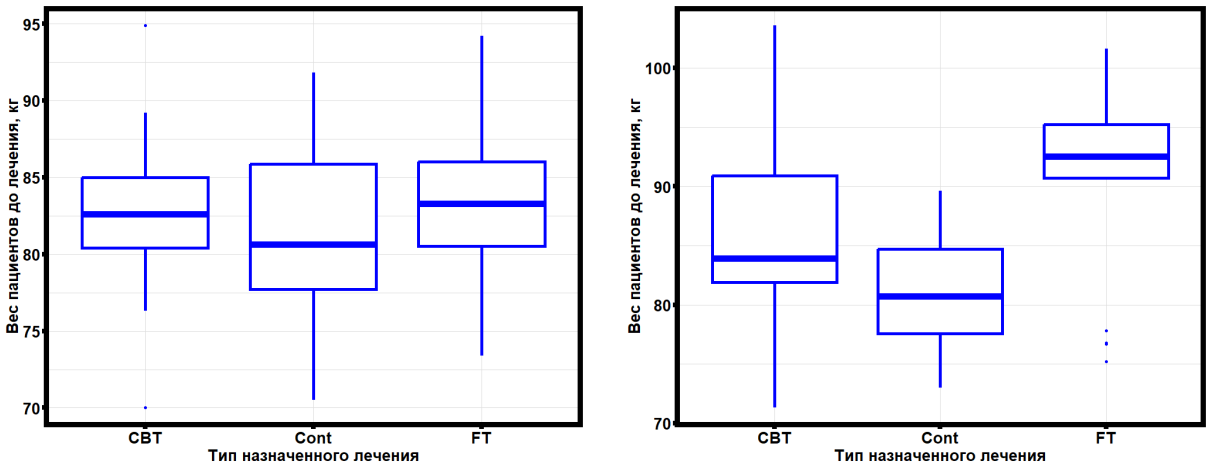
Результаты вычислений показывают, что медианная оценка веса пациентов с подтвержденным диагнозом «анорексия» во всех трех группах приблизительно равна выборочному среднему значению (см. пример в подразделе 2.1.1. «Выборочное среднее»), что согласуется с одним из свойств нормального закона распределения количественной величины – медиана, среднее и мода приблизительно равны между собой.

Все значения проведенного базового статистического анализа обобщаются в виде таблицы значений. В таблице 4 представлен пример сводной таблицы базового статистического анализа, проведенного для пациентов с подтвержденным диагнозом ***anorexia***.

Таблица 4 – Сводная таблица базового статистического анализа веса пациентов с подтвержденным диагнозом «анорексия»

№ п/п	Параметр базовой статистики	Контрольная группа	Группа с когнитивно-поведенческой терапией	Группа с семейной терапией
1	Среднее значение веса до проведения терапии	81,55769	82,68966	83,22941
2	Медиана веса пациента до проведения терапии	80,65	82,60	83,30
3	Среднее квадратическое отклонение веса до проведения терапии	5,707060	4,845495	5,016693
4	Максимальный вес пациента до проведения терапии	91,8	94,9	94,2
5	Минимальный вес пациента до проведения терапии	70,5	70,0	73,4
6	Первый квартиль веса пациента до проведения терапии	77,725	80,400	80,500
7	Третий квартиль веса пациента до проведения терапии	85,875	85,000	86,000
8	Среднее значение веса после проведения терапии	81,10769	85,69655	90,49412
9	Медиана веса пациента после проведения терапии	80,70	83,90	92,50
10	Среднее квадратическое отклонение веса после проведения терапии	4,744253	8,351924	8,475072
11	Максимальный вес пациента после проведения терапии	89,6	103,6	101.6
12	Минимальный вес пациента после проведения терапии	73,0	71,3	75,2
13	Первый квартиль веса пациента после проведения терапии	77,575	81,900	90,700
14	Третий квартиль веса пациента после проведения терапии	84,675	90,900	95,200

Хорошей иллюстрацией для обобщения базового статистического анализа является так называемый ящик с усами – диаграмма размаха, на которой в графическом виде представлен базовый статистический анализ результатов, рассчитанный на основе межквартильного интервала.



а б

Рисунок 11 – Диаграмма размаха «ящик с усами» веса пациентов: а – до прохождения лечения; б – после прохождения лечения

Центральная (жирная линия) ящика означает медиану (рисунок 11), нижняя граница ящика (прямоугольника) означает первый квартиль распределения значений, верхняя граница ящика (прямоугольника) – третий квартиль распределения исследуемой величины (в данном случае веса пациентов), «усы ящика» вычисляются по уравнениям (12, 13):

$$U_{down} = Q_1 - 1.5 \cdot IQR\,, \tag{12}$$

где U_{down} – значение нижней границы «усов ящика»; Q_1 – первый квартиль; IQR – межквартильный интервал.

$$U_{up} = Q_3 + 1.5 \cdot IQR\,, \tag{13}$$

где U_{up} – значение верхней границы усов ящика; Q_3 – третий квартиль; IQR – межквартильный интервал.

Точки, выходящие за пределы длины «усов», могут считаться аномальными (значения, сильно отличающиеся в большую или меньшую сторону от основной совокупности значений).

По результатам проведенного базового анализа выдвигается статистическая гипотеза, которая в результате применения различных статистических тестов и/или критериев принимается или отклоняется.

3. ПОНЯТИЕ О СТАТИСТИЧЕСКОЙ ГИПОТЕЗЕ

Статистическая гипотеза – выдвинутое предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным, содержащимся в исследуемой выборке[54].

Пусть в эксперименте доступна наблюдению случайная величина *X*, распределение которой *P* полностью или частично неизвестно. Тогда любое утверждение относительно *P* называется статистической гипотезой. Выше были представлены примеры базового статистического анализа веса пациентов с подтвержденным диагнозом ***anorexia***, т.е. вес пациентов, участвовавших в эксперименте, представляет собой случайную величину *X*. В разделе 1 было дано понятие о Гауссовом (нормальном) типе распределения данных. Предположение о том, что вес пациентов распределен по закону, близкому к нормальному (близко к рисунку 1), будет являться статистической гипотезой.

Гипотезы различают по виду предположений, содержащихся в них:

- Статистическая гипотеза, однозначно определяющая распределение P, то есть H: $\{P = P_0\}$, где P_0 – какой-то конкретный закон (например, вес пациентов во всех группах подчинен одному нормальному закону распределения), называется простой. H – гипотеза, принято различать нулевую гипотезу H_0 (вес пациентов с подтвержденным диагнозом «анорексия» распределен нормальным образом) и альтернативную гипотезу H_1 (вес пациентов с подтвержденным диагнозом «анорексия» распределен отличным от нормального закона распределения).
- Статистическая гипотеза, утверждающая принадлежность распределения *P* к некоторому семейству распределений, то есть вида $H: \{P \in \mathcal{P}\}$, где $\mathcal{P}$ – семейство распределений, называется сложной[55].

На практике обычно требуется проверить какую-то конкретную и, как правило, простую гипотезу H_0. Такую гипотезу принято называть нулевой. При этом параллельно рассматривается противоречащая ей гипотеза H_1, называемая конкурирующей, или альтернативной (вес пациента принадлежит нормальному закону распределения – нулевая гипотеза, вес пациента не принадлежит нормальному закону распределения – альтернативная гипотеза). В таблице 5 представлены формулировки трех простых нулевых и альтернативных гипотез, которые будут встречаться в тексте методических рекомендаций при решении трех основных задач:

1. Задача принадлежности данных к нормальному закону распределения.
2. Задача сравнения данных, выделенных по какому-либо признаку в различные группы.
3. Задача выявления статистической зависимости между переменными.

[54] Ивановский Р. И. Теория вероятностей и математическая статистика. Основы, прикладные аспекты с примерами и задачами в среде Mathcad: учеб. пос. СПб.: БХВ-Петербург, 2008. 528 с.

[55] В рамках данных методических рекомендаций не рассматриваются варианты возникновения сложных статистических гипотез, материал приведен для справок.

Таблица 5 – Общий вид стандартных формулировок простых нулевых и альтернативных гипотез для трех задач

Тип решаемой задачи	**H_0 – нулевая гипотеза**	**H_1 – альтернативная гипотеза**
Принадлежность данных к нормальному закону распределения	Наблюдаемое распределение данных принадлежит к нормальному (Гауссову) закону распределения	Наблюдаемое распределение данных отлично от нормального (Гауссова) закона распределения
Различие в группах данных, выделенных по какому-либо признаку в разные группы	Две группы данных, выделенные по какому-либо признаку, статистически **не**значимо различаются между собой	Две группы данных, выделенные по какому-либо признаку, статистически значимо различаются между собой
Зависимость между двумя переменными	Зависимость между двумя переменными статистически **не**значима	Наблюдаемая зависимость между переменными статистически значима

Выдвинутая гипотеза нуждается в проверке, которая осуществляется статистическими методами (посредством различных критериев), поэтому гипотезу называют статистической. Для проверки гипотезы используют критерии, позволяющие принять или опровергнуть выдвинутую гипотезу:

- Формулировка основной гипотезы H_0 и конкурирующей гипотезы H_1 (для примера см. таблицу 5).
- Задание уровня значимости α[56], на котором в дальнейшем и будет сделан вывод о справедливости гипотезы. Он равен вероятности допустить ошибку первого рода.
- Расчет статистики φ критерия такой, что[57]:

– ее величина зависит от исходной выборки $X = (X_1 \ldots X_n)$: $\varphi = \varphi(X_1 \ldots X_n)$;

– по ее значению можно делать выводы об истинности гипотезы H_0;

– статистика φ, как функция случайной величины X, также является случайной величиной и подчиняется определенному закону распределения.

- Построение критической области. Из области значений φ выделяется подмножество C таких значений, по которым можно судить о существенных расхождениях с предположением. Его размер выбирается таким образом, чтобы выполнялось равенство $P(\varphi \in C)=\alpha$. Это множество C и называется критической областью, α – ошибка первого рода.
- Вывод об истинности гипотезы. Наблюдаемые значения выборки подставляются в статистику φ и по попаданию (или непопаданию) в критическую область C выносится решение об отвержении (или принятии) выдвинутой гипотезы H_0.

В качестве примера рассмотрим нулевую гипотезу о принадлежности распределения данных нормальному закону. На рисунке 12 представлена гистограмма распределения случайной величины – вес пациента с подтвержденным диагнозом «анорексия» (вес пациента является исходной выборкой X, а предполагаемый закон распределения – φ).

[56] Более подробно уровень статистической значимости, ошибки первого и второго рода рассмотрены в разделе 5 настоящих методических рекомендаций.

[57] Данный материал приведен как справочный и редко применяется на практике, только в том случае, когда вычисления по критерию проводятся посредством самостоятельно реализованного программного кода.

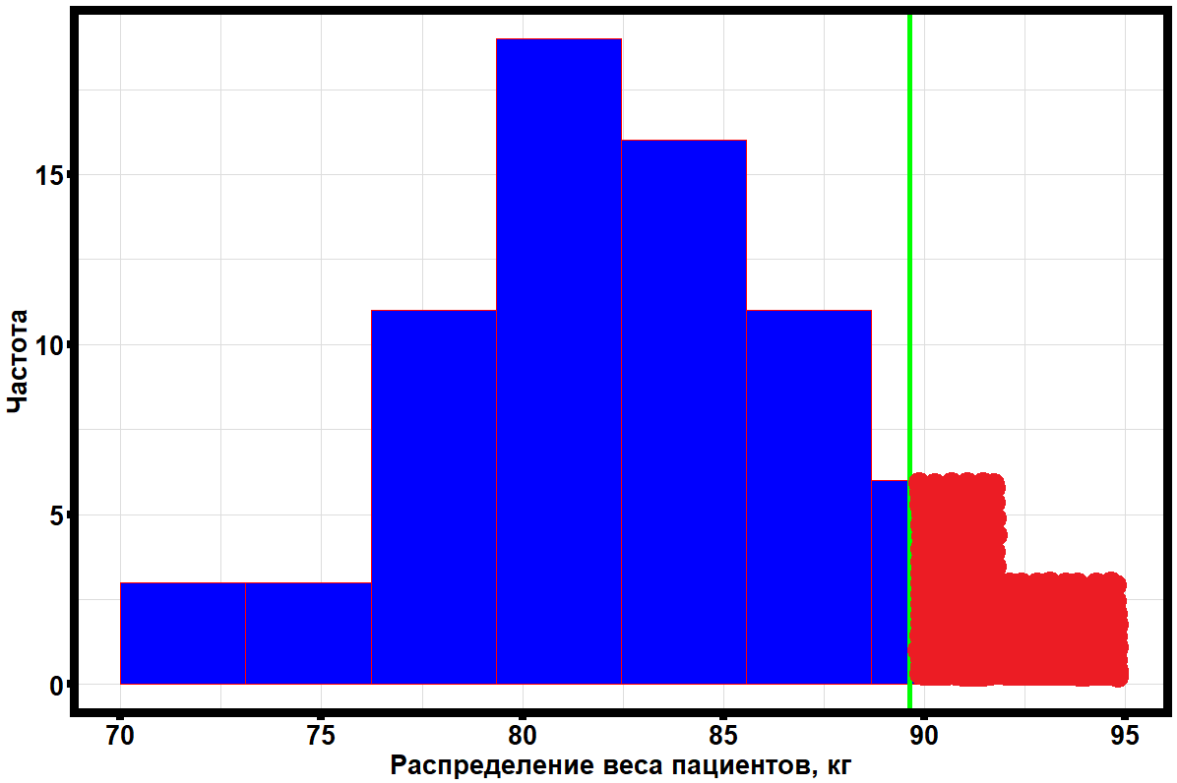


Рисунок 12 – Распределение веса пациентов с подтвержденным диагнозом «анорексия». Вертикальная зеленая линия соответствует 95 % распределения случайной величины. Красным цветом закрашена критическая область $\alpha = 5$ %

Соответственно, если $P(\varphi \in C) > 0{,}05$, то расхождения считаются статистически незначимыми, а если $P(\varphi \in C) \leq 0{,}05$, то расхождения статистически значимы, и необходимо отвергнуть нулевую гипотезу и принять альтернативную.

В большинстве случаев статистические критерии основаны на случайной выборке (X_1, X_2, …, X_n) фиксированного объема $n \geq 1$ для распределения P. В последовательном анализе выборка формируется в ходе самого эксперимента, и потому ее размер является случайной величиной[58].

[58] См.: https://ru.wikipedia.org/wiki/Проверка_статистических_гипотез.

4. ФОРМУЛИРОВКА НУЛЕВОЙ ГИПОТЕЗЫ

Одним из подходов предварительной формулировки нулевой гипотезы является построение гистограмм распределения случайной величины. На рисунке 13 представлена гистограмма распределения абстрактной случайной величины X.

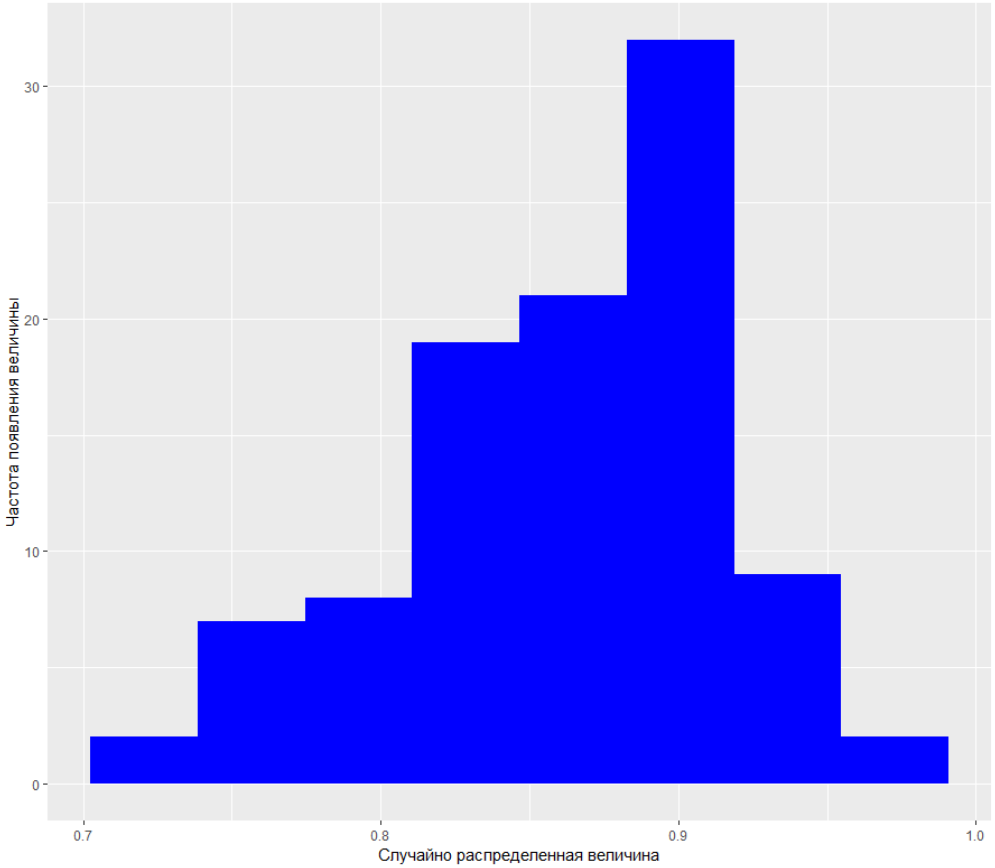


Рисунок 13 – Гистограмма распределения абстрактной случайной величины X

По виду диаграммы можно сделать предположение о близости типа распределения случайной величины к нормальному закону. Однако не стоит забывать о том, что данное предположение будет являться нулевой гипотезой, которая требует дальнейшей проверки[59].

Требование проверки нулевой гипотезы связано с особенностью построения гистограммы распределения случайной величины, а именно с выбором ширины интервалов, в которых подсчитывается число попаданий случайных значений. На рисунке 14 представлены изменения гистограммы случайной величины X (рисунок 13).

[59] Кендалл М., Стьюарт А. Теория распределений. М.: Наука, 1966. 588 с.

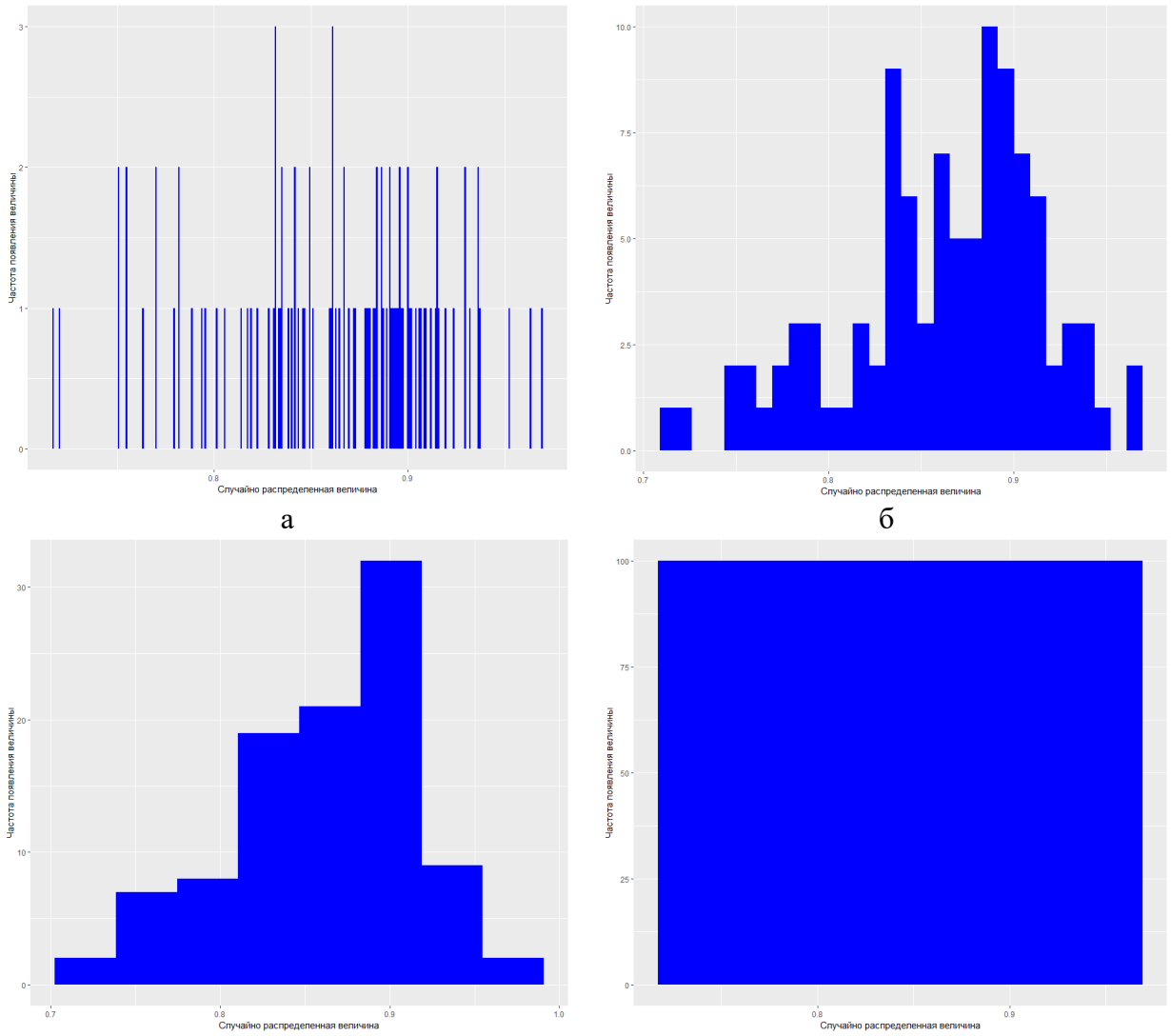


Рисунок 14 – Изменение вида гистограммы распределения случайной величины X в зависимости от выбора ширины интервалов, в которых подсчитывается количество случайных величин, попавших в данный интервал: а– Зауженный интервал; б – интервал увеличенной ширины; в – интервал средней ширины; г – широкий интервал

В результате предварительного анализа гистограммы, представленной на рисунке 14а, можно выдвинуть нулевую гипотезу о принадлежности распределения случайной величины X близко к дискретному типу распределения.

Результатом визуального анализа гистограммы, представленной на рисунке 14б, может служить постановка нулевой гипотезы о мультимодальности распределения случайной величины X (в этом случае необходимо проверить сложную статистическую гипотезу; рассмотрение анализа сложных статистических гипотез выходит за рамки данных методических рекомендаций).

Анализ гистограммы, представленной на рисунке 14в, может стать основой для формулирования нулевой гипотезы о принадлежности распределения случайной величины к нормальному закону распределения.

Таким образом, гистограмма является точкой начала отсчета для формулировки нескольких нулевых гипотез, которые должны быть проверены различными статистическими критериями.

4.1. Расчет оптимального интервала на гистограмме

Расчет оптимального количества интервалов и диапазона значений, попадающих в интервал гистограммы распределения, является важной составляющей базовой статистики и может служить основой для выделения групп равномерно распределенных данных и выработки классификации на их основе. В данных методических рекомендациях рассматривается несколько простых подходов расчета оптимального количества интервалов и диапазона изменения значений. Первый подход основан на эвристической формуле Стерджесса (14)[60]:

$$n = \log_2 N + 1 = 3.33 \cdot \log_{10} N + 1, \quad (14)$$

где N – объем анализируемой выборки (количество значений); n – количество интервалов разбиения данных.

В работе[61], описывающей применение статистического анализа при контроле качества продукции, применяют уравнение Брукса–Каррузера (15):

$$n = 5 \cdot \log_{10} N \quad (15)$$

Наиболее простым методом оценки количества интервалов является (16)[62]:

$$n = \sqrt{N} \quad (16)$$

При больших объемах данных рекомендуется использовать уравнение (17)[63]:

$$n = \sqrt[3]{N} \quad (17)$$

После расчета количества интервалов разбиения гистограммы распределения проводят вычисления ширины интервала по уравнению (18):

$$\Delta X_{hist} = \frac{\max(X) - \min(X)}{n}, \quad (18)$$

где $\max(X)$ – максимальное значение исследуемой переменной X; $\min(X)$ – минимальное значение исследуемой переменной X; n – количество интервалов на гистограмме.

Соответственно, с использованием уравнения (18) вычисляются первый и последующие интервалы по уравнению (19):

$$X_{i+1} = X_i + \Delta X_{hist,} \quad (19)$$

где X_{i+1} – правая граница i-го интервала, не превышающая $\max(X)$; X_i – левая граница i-го интервала, начало отсчета которой равно $\min(X)$.

---

[60] Herbert A. S. The choice of a class interval // Journal of the American statistical association. 1926. Vol. 21, №153. P. 65–66.

[61] Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. М.: Мир, 1970. 368 с.

[62] Heinhold I., Gaede K. W. Ingeniur statistic. München; Wien, Springler Verlag, 1964. 352 p.

[63] Ченцов Н. Н. Статистические решающие правила и оптимальные выводы. М.: Наука, 1972. 520 с.

Высота столбца диаграммы определяется по уравнению (20):

$$H_i = \frac{N_i}{N \cdot \Delta X_i} \quad (20)$$

где N – общее количество исследований; N_i – количество исследований в i-м интервале; H_i – высота интервала.

Более точная (и она же более сложная) оценка размера интервала может быть получена при оценках статистической мощности[64], данный подход не рассматривается в настоящих методических рекомендациях.

4.2. Построение гистограмм распределения на языке R

В языке программирования R существует несколько способов построения гистограмм распределения случайной величины:

– с применением функции ***hist ()***, входящей в пакет **graphics**;

– посредством функции ***geom_histogram()***, входящей в состав пакета **ggplot2**;

– с помощью столбчатой диаграммы с предварительным расчетом частот и интервалов распределения величин. Построение столбчатых диаграмм можно произвести с помощью:

– функции ***barplot()***, входящей в пакет **graphics**;

– функции ***geom_bar()***, входящей в пакет **ggplot2**.

В примерах данного раздела приведены все перечисленные способы построения гистограммы распределений.

Примеры построения гистограмм распределений на языке R

Для построения диаграмм распределения использовались данные по выживаемости заболевших вирусом иммунодефицита, собранных в Австралии после 1 июля 1991 года. Все данные содержатся в пакете MASS, наборе данных ***Aids2***.

Листинг 10

```
#Построение гистограммы распределения с помощью функции hist () пакета graphics
library("MASS") #Подключаем пакет MASS, содержащий набор данных Aids2
N <- length(Aids2$Age) #Вычисляем количество пациентов в наборе данных
dBreak <-5*log10(N) #Определяем количество интервалов для гистограммы
dBreak <- round(dBreak) #Округляем значение до целых чисел
dX <- (max(Aids2$Age)-min(Aids2$Age))/ dBreak #Рассчитываем шаг
step <- seq(min(Aids2$Age),max(Aids2$Age), dX) # Вектор интервалов
hist (x = Aids2$age, break = step, xlab= "Возраст выявления, полных лет",
   ylab = "Количество, чел", main = "Выживаемость, заболевших СПИДом",
   col = "blue") #Строим гистограмму распределения возрастов пациентов с
                 # с выявленным вирусом иммунодефицита
# x – вектор значений исследуемых распределений, полученный из Data Frame
#;
# break – количество участков на гистограмме;
```

[64] Лемешко Б. Ю., Чимитова Е. В. О выборе числа интервалов в критериях согласия типа с2 // Заводская лаборатория. Диагностика материалов. 2003. Т. 69, №1. С. 61–67.

Продолжение листинга 10

xlab – подпись оси OX;
ylab – подпись оси OY;
main – название гистограммы;
col – задание цвета гистограммы распределения.

Результат выполнения данной команды представлен на рисунке 15.

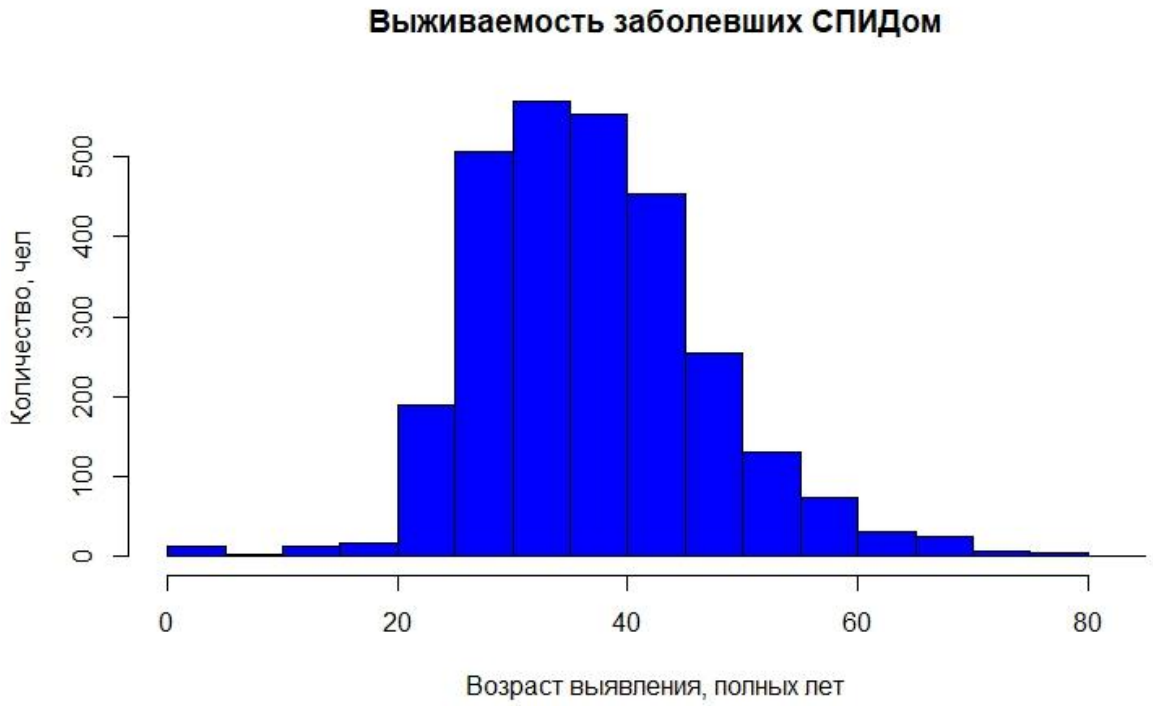


*Рисунок 15 – Результат применения функции **hist ()** из пакета **graphics***

Функция ***hist ()*** хороша для применения при необходимости быстро построить гистограмму и посмотреть на распределение данных, однако более гибким инструментом для построения и дальнейшей публикации гистограмм является функция ***geom_histogram(),*** входящая в пакет **ggplot2**.

Листинг 11

```
# Построение гистограммы распределения с помощью функции
# geom_histogram () пакета ggplot2
library("MASS") # Подключаем пакет MASS, содержащий набор данных Aids2
library("ggplot2")  # Подключаем пакет ggplot2, содержащий функцию
        # geom_histogram ()
p <- ggplot() #Создаем объект p, содержащий[65] слои графика
p<- p + geom_histogram (mapping = aes (x= Aids2$age), fill= "blue", binwith=30)
p<- p+labs (x = "Возраст выявления, полных лет",
   y= "Количество, чел",
    title = "Выживаемость, заболевших СПИДом")
print(p)
```

[65] Более подробно структура и применение пакета ggplot2 представлена в работе: Мастицкий С. Э. Визуализация данных с помощью ggplot2. М.: ДМК Пресс, 2017. 222 с.

Результат выполнения кода листинга 11 представлен на рисунке 16.

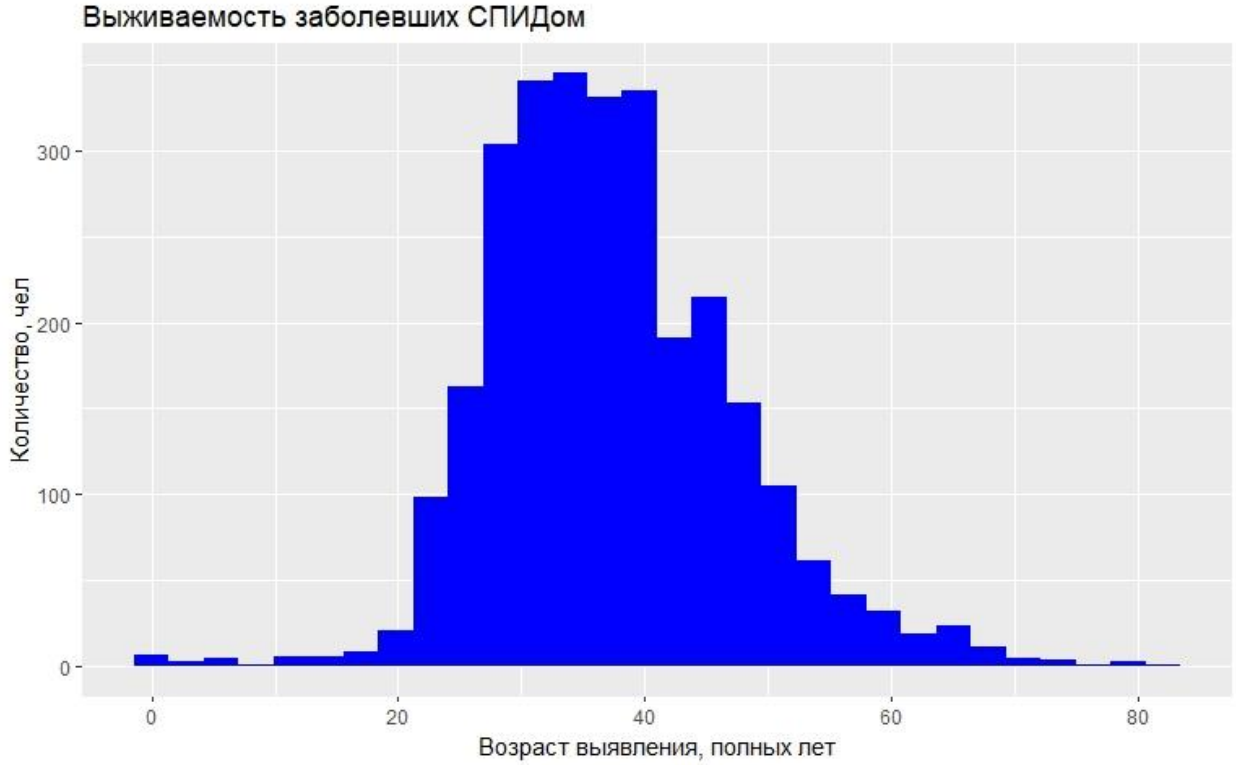


Рисунок 16 – Гистограмма распределения по возрасту пациентов с подтвержденным диагнозом иммунодефицита

Библиотека ***ggplot2*** содержит большое количество функций, позволяющих строить различные типы графиков и производить их тонкую настройку, однако требует и большего количества строк кода. Описание всех возможностей библиотеки выходит за рамки данных методических рекомендаций.

Классический способ построения гистограмм распределения основан на вычислении размаха данных, вычислении интервалов и построении столбчатой диаграммы. При таком подходе выбирается количество интервалов, равномерно распределенных по данным, и подсчитывается количество вхождений в каждый интервал исследуемых значений, а на последнем этапе строится столбчатая диаграмма количества вхождений в интервал.

Листинг 12

```
library(“MASS”) #Подключаем пакет, содержащий исследуемые данные
N <- length(Aids2$age) # Выделяем количество пациентов
dBreak <-5*log10(N) #Определяем количество интервалов для гистограммы
dBreak <- round(dBreak) #Округляем значение до целых чисел
interval <- cut (Aids2$age, breaks = dBreak,) #Разбиваем данные на 12 интервалов
freqData <- summary(interval) #Вычисляем количество вхождений в каждый
  # интервал
barplot(freqData, xlab = "Возраст выявления, полных лет",
ylab = "Количество, чел",
main = "Выживаемость, заболевших СПИДом", col = "blue")
```

Результат выполнения данного кода представлен на рисунке 17.

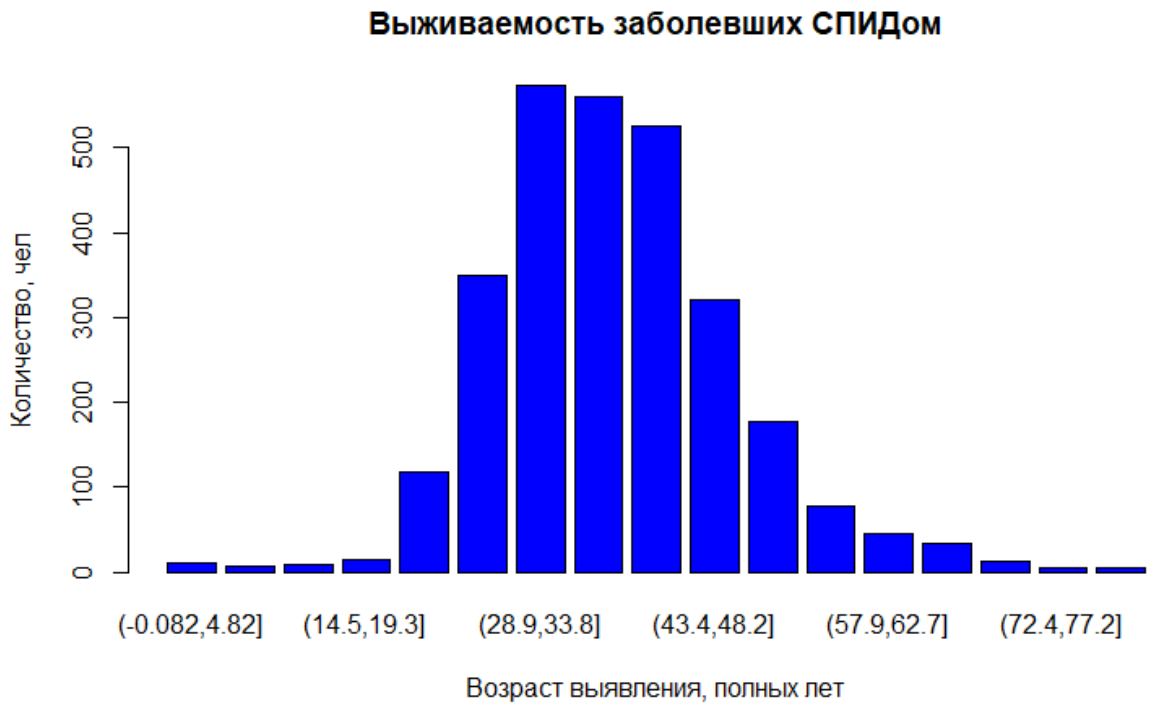


Рисунок 17 – Гистограмма распределения по возрасту пациентов с подтвержденным диагнозом иммунодефицита

В листинге 13 представлен пример построения гистограммы распределения с помощью ***geom_bar ()***, входящей в пакет **ggplot2**.

Листинг 13

```
library("MASS") #Подключение пакета, содержащего исследуемые данные
library("ggplot2") #Подключаем пакет ggplot2, содержащий geom_bar ()
dBreak <-5*log10(N) #Определяем количество интервалов для гистограммы
dBreak <- round(dBreak) #Округляем значение до целых чисел
interval <- cut (Aids2$age, breaks = dBreak) # Разбиваем данные на 12 интервалов
interval <- as. factor(interval) #Преобразуем интервалы в факторы
p<- ggplot()
p<- p+geom_bar(mapping = aes(x=interval), fill = "blue", stat= "count")
p<- p+labs (x= "Возраст выявления, полных лет",
   y = "Количество, чел", title = "Выживаемость, заболевших СПИДом")
print(p)
```

На рисунке 18 представлен результат выполнения листинга 13.

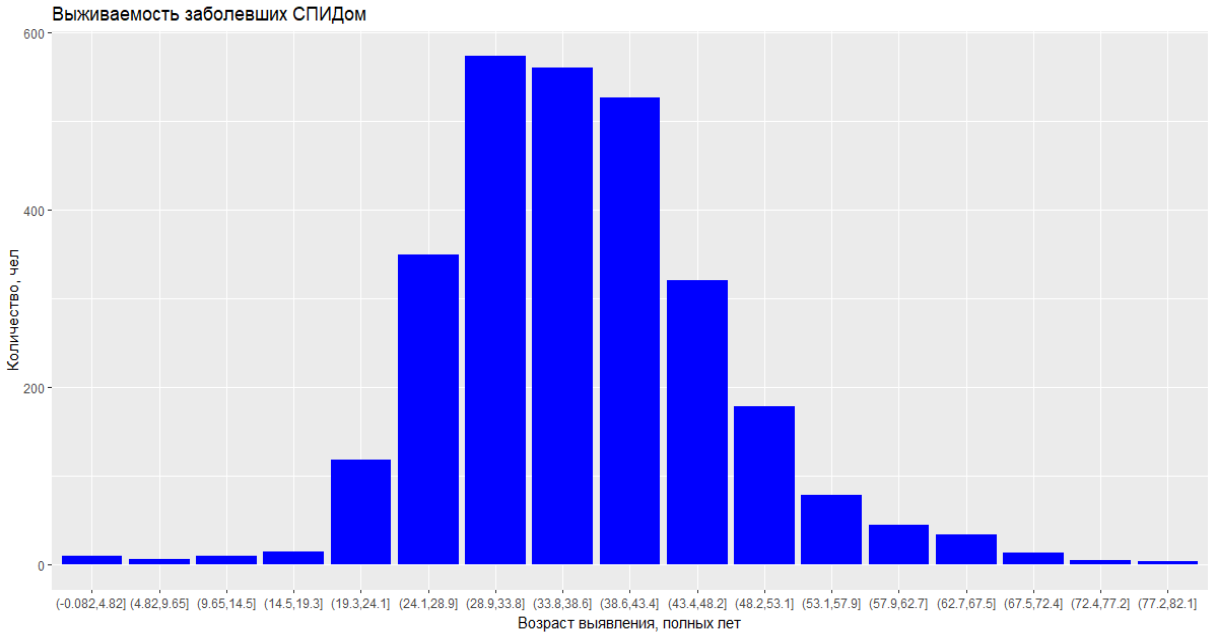


Рисунок 18 – Гистограмма распределения по возрасту пациентов с подтвержденным диагнозом иммунодефицита

Стоит отметить, что при применении функции ***geom_bar ()***, входящей в пакет **ggplot2**, не требуется отдельного подсчета количества вхождений в каждый интервал, функция делает это автоматически.

5. ЗАДАНИЕ УРОВНЯ СТАТИСТИЧЕСКОЙ ЗНАЧИМОСТИ

В статистике величину (значение) переменной называют статистически значимой, если мала вероятность случайного возникновения этой или еще более крайних величин. Здесь под крайностью понимается степень отклонения тестовой статистики от нулевой гипотезы.

Разница между двумя выборками называется статистически значимой, если появление имеющихся данных было бы маловероятно при предположении, что эта разница отсутствует.

Популярными уровнями значимости (α-ошибка) являются 10 %, 5 %, 1 %, и 0,1 %.

Меньшие α-уровни дают большую уверенность в том, что уже установленная альтернативная гипотеза значима, но при этом есть больший риск не отвергнуть ложную нулевую (или отвергнуть истинную альтернативную) гипотезу (ошибка второго рода, или «ложноотрицательное решение», или β-ошибка), и таким образом меньшая статистическая мощность.

Ошибкой первого рода (α-ошибка – уровень значимости, ложноположительное заключение) – называют ситуацию, когда отвергнута верная нулевая гипотеза (об отсутствии связи между явлениями, группами данных или принадлежностью данных к какому-либо типу распределений).

Вычисление уровня значимости при моделировании и анализе данных может быть выполнено по уравнению (21)[66]:

$$\alpha = \sum_{i=k}^{n} C_n^i * p_1^i * (1 - p_1)^{n-i}, \tag{21}$$

где k – минимальное количество измерений, выходящих за заданный доверительный интервал; n – общее число измерений исследуемой величины; p_1 – вероятность возникновения измерения за пределами доверительного интервала.

Коэффициент C_n^i – коэффициент биноминального распределения, рассчитывается, как (22):

$$C_n^i = \frac{n!}{(n-i)!\,i!}, \tag{22}$$

где i – моделируемое количество измерений, выходящих за заданный доверительный интервал; n – общее количество измерений в данных.

На практике существуют две основные проблемы при вычислении ошибки первого рода:

1. Вероятность возникновения измерения за пределами доверительного интервала не известна и должна быть исследована отдельно от основной исследуемой величины.
2. Определение доверительного интервала при заранее неизвестных типах распределений также не известно.

Пути решения данных проблем будут рассмотрены далее.

Ошибка второго рода (β-ошибка, ложноотрицательное заключение) – ситуация, когда принята неверная нулевая гипотеза[67]. Популярными уровнями ошибки второго рода являются 30 %, 20 % и 10 %.

---

[66] Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Методы обработки данных / пер. с англ. М.: Мир, 1980.

[67] См.: https://ru.wikipedia.org/wiki/Ошибки_первого_и_второго_рода.

Вероятность ошибки второго рода определяется по уравнению (23):

$$\beta = \sum_{i=n-(k-1)}^{n} C_n^i * p_2^i * (1 - p_2)^{n-i}, \quad (23)$$

где p_2 – вероятность возникновения выброса внутри доверительного интервала.

Основные проблемы, возникающие при вычислении ошибки второго рода, такие же, как и в случае определения ошибки первого рода. Выбор α-уровня неизбежно требует компромисса между значимостью и мощностью и, следовательно, между вероятностями ошибок первого и второго рода.

Общая картина проблемы такова: даны выборка X из некоторого пространства Ω элементарных событий (например, список пациентов, прошедших обследование на выявление некоторого заболевания) $X: \{X \in \Omega\}$ и возможные значения в этой выборке некоторых переменных (функций от $\omega \in \Omega$, например – возраст пациента, интенсивность курения, количество часов физических упражнений и т.п.). Вероятностное распределение случайной величины X на Ω неизвестно и является главным объектом поиска.

Различные гипотезы соответствуют различным возможным вероятностным распределениям на пространстве Ω. Точный смысл термина «статистическая гипотеза» – набор утверждений, который содержит полное описание некоторого вероятностного распределения[68].

Пример моделирования ошибки первого и второго рода на языке R

Для демонстрации понятий ошибки первого и второго рода проведем моделирование поведения распределений плотностей вероятностей. Предположим (это только предположение и является сугубо учебным примером, реальное распределение признака может иметь любой закон[69]), что распределение какого-то количественного показателя (признака) пациентов с положительным тестом на СПИД подчиняется нормальному закону распределения. Также предположим, что распределение какого-то количественного показателя (признака) пациентов старше 52 лет отлично от нормального закона распределения (предположим, что это отрицательное биноминальное распределение)[70].

Листинг 14

```
library("MASS") #Подключаем библиотеку, содержащую исследуемые данные
library("fitdistrplus") #Подключаем библиотеку вычисления параметров
# распределений по методу максимального правдоподобия
library("ggplot2") #Подключаем библиотеку графического представления данных
denNorm <- fitdist (Aids2$age, "norm") #Вычисляем параметры нормального
  # закона распределения
denNBinom <- fitdist (Aids2[Aids2$age>52,] $age, "nbinom") #Вычисляем
# параметры отрицательного биноминального распределения
normDens <- dnorm (Aids2$age, mean = denNorm$estimate[1], sd = denNorm$estimate[2]) # Вычисляем функцию плотности
```

[68] См.: https://ru.wikipedia.org/wiki/Статистическая_значимость.

[69] Рассмотрение методов определения наиболее близкого теоретического распределения выходит за рамки данных методических рекомендаций.

[70] Здесь и далее все графические построения будут проводиться с применением пакета ggplot2.

Продолжение листинга 14

```
        # вероятности нормального распределения
nbinDens <- dnbinom (Aids2$age, size = denNBinom$estimate[1],
mu = denNBinom$estimate [2]) # вычисляем плотность вероятности
#отрицательного биноминального распределения
qvanNorm <- quantile (Aids2$age, probs = seq (0.9,1, 0.05)) # Определяем значения
  # квантилей
qvanNBin <- quantile (Aids2$age, probs = seq (0,1,0.2))
p <- ggplot()
p <- p + geom_line (mapping = aes (Aids2$age, normDens), colour = "magenta",
size = 2)
p <- p + geom_line(mapping = aes (Aids2$age, nbinDens), colour = "green", size=2)
p <- p + geom_vline(xintercept = qvanNBin[2], colour = "green", size=2)
p <- p + geom_vline(xintercept = qvanNorm[2], colour = "magenta", size = 2)
p <- p + labs (x = "Возраст выявления, полных лет", y = "Количество, чел",
       title = "Выживаемость заболевших СПИДом")
print(p)
```

На рисунке 19 представлены результаты работы программы листинга 14.

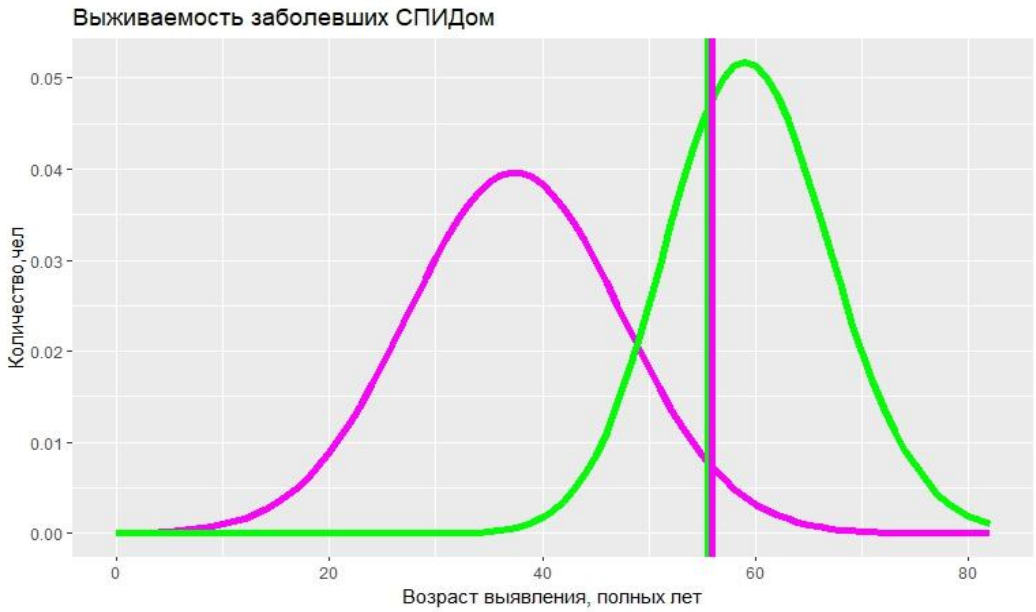


Рисунок 19 – Демонстрация ошибок первого и второго рода

Распределение пациентов левее зеленой вертикальной линии и до нуля зеленой функции распределения может быть ошибочно отнесено к фиолетовой функции распределения, тогда как исследования, находящиеся правее фиолетовой вертикальной линии, могут быть ошибочно отнесены к зеленой функции распределения. В первом случае такие исследования составляют 20 %, а во втором – 5 %.

6. ПРОВЕРКА ДАННЫХ НА ПРИНАДЛЕЖНОСТЬ К НОРМАЛЬНОМУ ЗАКОНУ РАСПРЕДЕЛЕНИЯ

Проверка данных на принадлежность к нормальному закону распределения является одним из ключевых шагов при проведении статистического анализа и дальнейшего построения моделей.

В настоящее время существует достаточно большое количество критериев, позволяющих сделать вывод о принадлежности распределения той или иной количественной переменной к нормальному (Гауссову) закону распределения. Рекомендованными метрологическими стандартами и работами ряда авторов являются[71],[72]:

- критерии асимметрии и эксцесса;
- критерий Жарка-Бера;
- критерий Дэ'Агустино;
- критерий Шапиро-Уилка;
- критерий Эппса-Палли.

Рассмотрим каждый из них подробнее.

6.1. Критерии асимметрии и эксцесса

Критерий асимметрии предназначен для проверки гипотезы о симметричности наблюдаемого закона. Статистический критерий имеет вид (24):

$$\hat{\beta}_1 = \frac{\widehat{\mu_3}}{\widehat{\sigma^3}}, \tag{24}$$

где $\widehat{\sigma^3}$ – среднее квадратическое отклонение в третьей степени; $\hat{\beta}_1$ – коэффициент асимметрии; $\widehat{\mu_3}$ – третий момент исследуемой случайной величины

Общее описание моментов представляется, как (25):

$$\widehat{\mu_J} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^J, \tag{25}$$

где J – равно трем для третьего момента и равно двум для второго момента; $\bar{X}$ – определяется по уравнению (26):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{26}$$

Нулевая гипотеза в данном случае выглядит как H_0: $\beta_1 = 0$. Альтернативная гипотеза H_1: $\beta_1 > 0$ (положительная асимметрия) или $\beta_1 < 0$ (отрицательная асимметрия).

Критерий проверки на эксцесс вычисляется по уравнению (27):

$$\widehat{\beta_2} = \frac{\widehat{\mu_4}}{\hat{\sigma}^4} \tag{27}$$

[71] ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятности от нормального распределения.

[72] Лемешко Б. Ю., Лемешко С. Б. Сравнительный анализ критериев проверки отклонения распределения от нормального закона // Метрология. 2005. № 2. С. 3–23

Проверке подвергается нулевая гипотеза H_0: $\hat{\beta}_2 = 3$; альтернативная гипотеза H_1: $\hat{\beta}_1 > 3$ (больший эксцесс) или $\hat{\beta}_1 < 3$ (меньший эксцесс). Оба критерия применяются при условии, что количество наблюдений составляет значение от 8 до 5000.

На языке программирования R для фактических численных данных можно вычислить значение критерия асимметрии и эксцесса для реальных данных. С помощью функции ***skewness(x)*** библиотеки ***moments*** вычисляется значение критерия асимметрии, ***kurtosis(x)*** – функция для вычисления эксцесса, где ***x*** – вектор данных, для которых вычисляется критерий.

Пример расчета асимметрии и эксцесса

На примере анализа распределения возраста пациентов с положительным результатом теста на иммунодефицит рассмотрим вычисление эксцесса и асимметрии, а также интерпретацию полученных результатов.

Листинг 15

```
library("MASS") # Подключаем библиотеку, содержащую данные пациентов
install.packages("moments") # Устанавливаем библиотеку, содержащую функции вычисления эксцесса и асимметрии
library("moments") # Подключаем библиотеку, содержащую функции вычисления
    # эксцесса и асимметрии
aidData <- Aids2$age # Создаем переменную <aidData>, содержащую возраст
 # пациентов
assimData <- skewness(aidData) # Вычисляем асимметрию распределения
eksData <- kurtosis(aidData) # Вычисляем эксцесс распределения
hist (x = Aids2$age, break = 12, xlab= "Возраст выявления, полных лет",
   ylab = "Количество, чел", main = "Выживаемость, заболевших СПИДом",
   col = "blue") #Строим гистограмму распределения возраста пациентов с
                 # с выявленным вирусом иммунодефицита
```

Результаты выполнения кода в листинге 15 представлены на рисунке 20 и в тексте под рисунком.

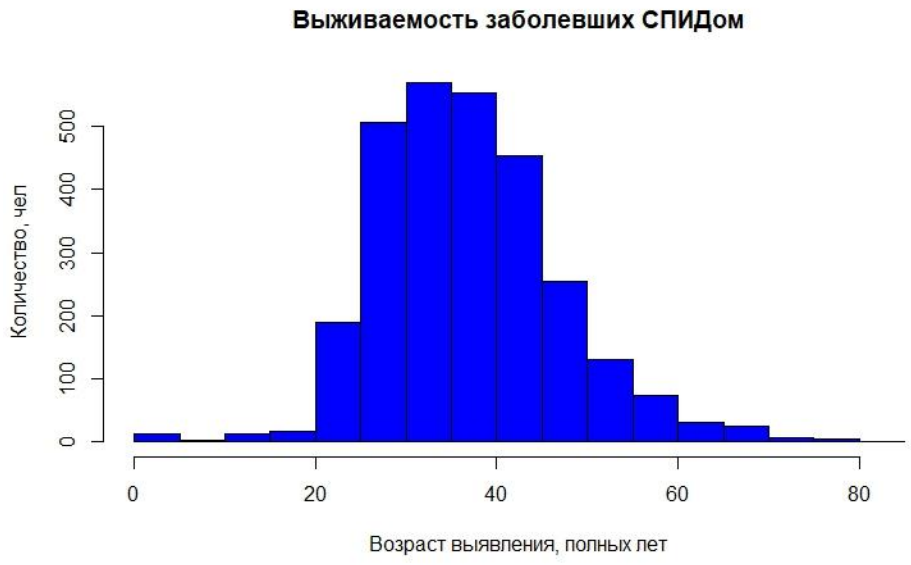


Рисунок 20 – Гистограмма распределения возраста пациентов с положительным результатом теста на вирус иммунодефицита

Продолжение листинга 15

```
> skewness(aidData) # результат вычисления коэффициента асимметрии
[1] 0.487489
> kurtosis(aidData) # результат вычисления коэффициента эксцесса
[1] 4.279771
```

Значение эксцесса превышает значение, равное 3, что свидетельствует о том, что распределение не подчиняется нормальному закону распределения, а положительное значение коэффициента асимметрии говорит о том, что центр симметрии смещен в сторону увеличения возраста пациента.

6.2. Критерий Жарка–Бера[73]

Критерий Жарка–Бера наравне с другими критериями применяется для проверки принадлежности распределения данных к нормальному закону. Статистический критерий имеет вид (28):

$$LM_N = N[\frac{(\sqrt{\widehat{\beta_1}})^2}{6} + \frac{(\hat{\beta}_2 - 3)^2}{24}], \quad (28)$$

где $\hat{\beta}_1$ – значение коэффициента асимметрии; $\widehat{\beta_2}$ – значение коэффициента эксцесса.

При интерпретации результатов применения теста необходимо иметь в виду, что в качестве нулевой гипотезы выступает H_0-предположение о том, что асимметрия и эксцесс соответствуют нормальному закону распределения, а альтернативной гипотезой H_1 является то, что асимметрия и эксцесс не соответствуют нормальному закону распределения. Количество наблюдений должно составлять значение от 8 до 5000.

На языке программирования R в библиотеке ***moments*** присутствует функция ***jarque.test(x)***, где ***x*** – вектор значений, распределение которых проверяется на принадлежность к нормальному закону критерием Жарка–Бера.

Пример применения критерия Жарка–Бера

Рассмотрим применение теста Жарка–Бера на примере данных по анорексии из пакета MASS и интерпретацию результата. Используем данные по весу пациента до и после обследования.

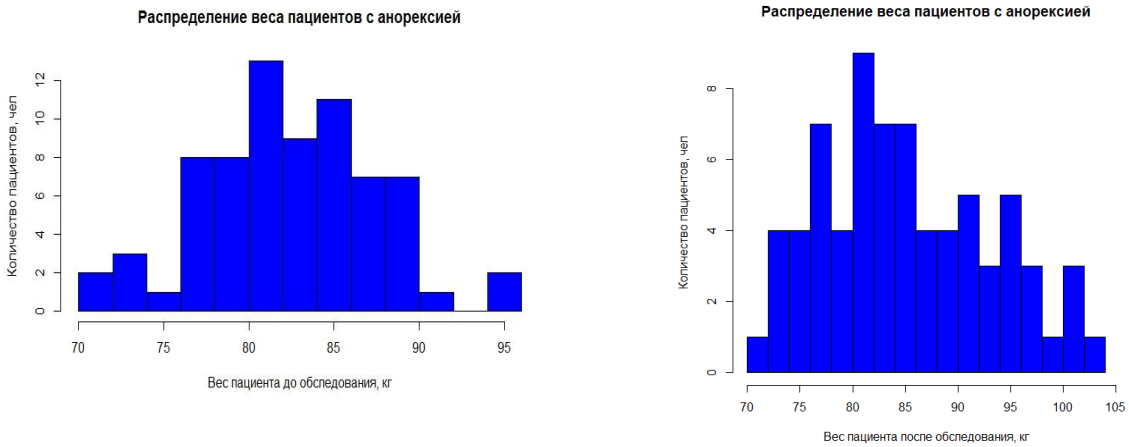
Листинг 16

```
library("MASS") #Подключаем библиотеку, содержащую данные по пациентам
  # с анорексией
library("MOMENTS") # Подключаем библиотеку, содержащую тест Жарка–
Бера
dataBSP <- anorexia$Prewt # Создаем переменную, содержащую вектор с весом
# пациента до обследования
dataASP <- anorexia$Postwt # Создаем переменную, содержащую вектор с весом
                    # пациента после обследования
jarque.test(dataBSP) # проводим тест Жарка–Бера для веса пациента до
```

[73] Jarque C. M., Bera, A. K. A test for normality of observations and regression residuals // International Statistical Review. 1987. № 55. P. 163–172

Продолжение листинга 16

```
# обследования
Jarque-Bera Normality Test

data: dataBSP
JB = 0.04897, p-value = 0.9758
alternative hypothesis: greater
jarque.test(dataASP) # проводим тест Жарка–Бера для веса пациента после
# обследования
Jarque-Bera Normality Test

data: dataASP
JB = 3.3656, p-value = 0.1859
alternative hypothesis: greater

hist(dataBSP, breaks = 12, xlab = "Вес пациента до обследования, кг",
        ylab = "Количество пациентов, чел",
        main = "Распределение веса пациентов с анорексией", col="blue")

hist(dataASP, breaks = 12, xlab = "Вес пациента после обследования, кг",
        ylab = "Количество пациентов, чел",
        main = "Распределение веса пациентов с анорексией", col="blue")
```



а б

Рисунок 21 – Гистограммы распределений веса пациентов до и после обследования

Из гистограмм распределения веса пациентов визуально можем сделать предположение, что вес пациентов до и после обследования распределен по закону, отличному от нормального. Однако на практике визуальная и точная оценки зачастую не совпадают, поэтому необходимо посмотреть на уровень значимости, полученный при проведении критерия Жарка–Бера, и сделать вывод о принадлежности распределения веса пациентов с анорексией к нормальному закону.

Для пациентов, страдающих анорексией, до обследования получено значение статистического критерия (28), равное 0,04897, и уровень значимости, равный 0,9758, при пороговом уровне значимости, равным 0,05. Из полученных результатов следует, что асимметрия и эксцесс распределения значений веса пациентов до обследования соответствуют нормальному закону распределения.

Для пациентов, страдающих анорексией, после обследования получено значение статистического критерия (28), равное 3,3656, и уровень значимости, равный 0,1859, при пороговом уровне значимости, равном 0,05. Из полученных результатов следует,

что асимметрия и эксцесс распределения значений веса пациентов после обследования соответствуют нормальному закону распределения.

6.3. Критерий Дэ'Агустино

Критерий Дэ'Агустино базируется на анализе отклонений средних величин в упорядоченной выборке $X=x_1, x_2, \ldots, x_n$ и вычисляется по уравнению (29):

$$D_{ag} = \frac{\sum_{i=1}^{n} i * x_{(i)} - \left(\frac{1}{2}\right) * (n+1) * \sum_{i=1}^{n} x_{(i)}}{\sqrt[n]{n * \sum_{i=1}^{n} (x_i - \bar{x})^2}}, \quad (29)$$

где n – число значений в упорядоченной выборке; i – порядковый номер элемента в выборке; $\bar{x}$ – среднее арифметическое значений в выборке.

На языке R критерий Дэ'Агустино вычисляется с помощью функции ***agostino.test(x)*** из библиотеки ***moments***, где ***x*** – упорядоченная выборка значений.

Пример применения критерия Дэ'Агустино

Рассмотрим применение критерия Дэ'Агустино для проверки на нормальность возраста пациентов с выявленной меланомой.

Листинг 17

```
library("MASS") # Подключаем пакет MASS, содержащий данные Melanoma
library("moments") # Подключаем пакет moments, содержащий функцию теста
  # Дэ'Агустино
dataAge <- Melanoma$age #Вектор, содержащий вес пациентов
agostino.test(dataAge)
D'Agostino skewness test

data: dataAge
skew = -0.29798, z = -1.76638, p-value = 0.07733
alternative hypothesis: data have a skewness
hist(dataAge, breaks = 12, xlab = "Возраст пациентов с меланомой, лет",
          ylab = "Количество пациентов, чел",
             main = "Распределение возраста пациентов с меланомой",
col="blue")
```

На рисунке 22 представлена гистограмма распределения возраста пациентов с подтвержденным диагнозом «меланома».

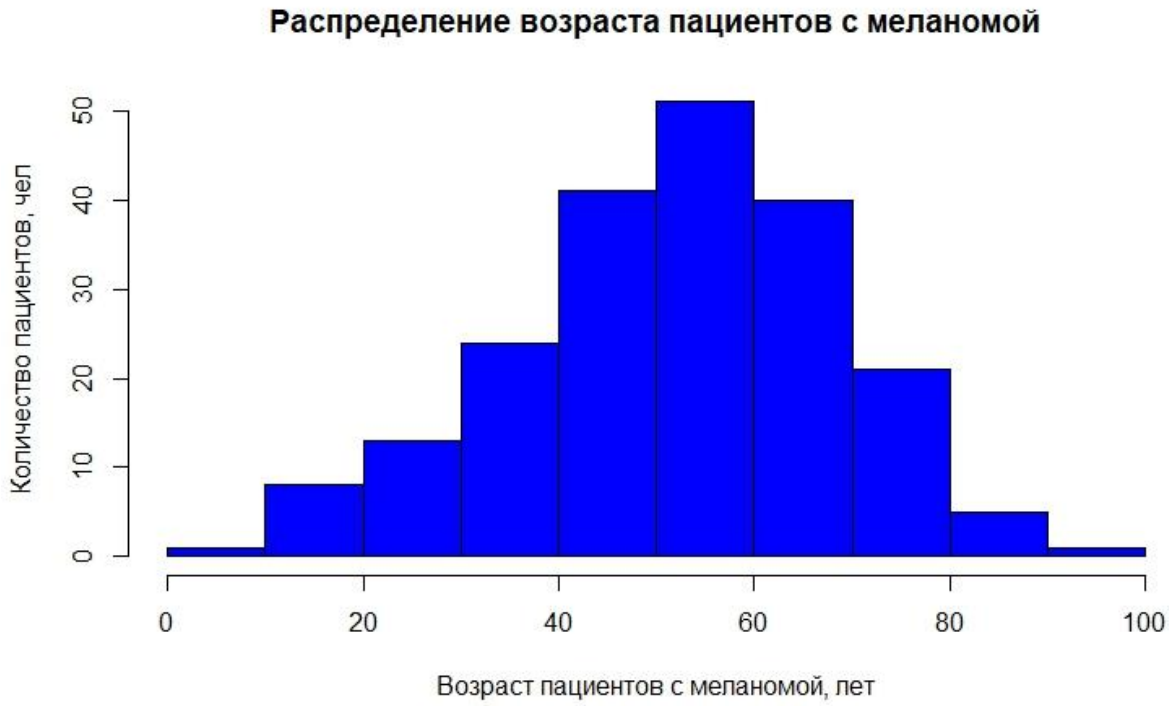


Рисунок 22 – Гистограмма распределения возраста пациентов с подтвержденным диагнозом «меланома»

Анализ результатов применения критерия Дэ'Агустино показывает: p-value = 0,07733> 0,05, что говорит о том, что распределение возрастов пациентов с подтвержденным диагнозом «меланома» подчиняется нормальному закону распределения. Визуальный анализ гистограммы распределения также свидетельствует об этом.

6.4. Критерий Шапиро–Уилка

Критерий Шапиро–Уилка[74,75] базируется на анализе линейной комбинации разностей порядковых статистик (под порядковой статистикой понимается упорядоченная по возрастанию выборка одинаково распределенных независимых случайных величин, и ее элементы занимают строго определенное место в ранжированной (упорядоченной по возрастанию) совокупности).

При построении статистики для вариационного ряда $X_1 \le X_2 \le \cdots \le X_n$, полученного по наблюдаемой выборке $X_1, X_2, \dots, X_n$, вычисляется величина (30)[76]:

$$S = \sum_k a_k \left[X_{(n-k+1)} - X_{(k)} \right], \qquad (30)$$

[74] Shapiro S. S., Wilk M. B. An analysis of variance test for normality (complete samples) // Biometrika. 1965. № 52. P. 591–611.

[75] Shapiro S. S., Francia R. S. An appriximate analysis of variance test fo normality // J. Amer. Statist. Assoc. 1972. №337. P. 215–216.

[76] Более подробную информацию по выводу уравнений смотрите в оригинальных публикациях, приведенных выше.

где k – изменяется от 1 до $\frac{n}{2}$, если n – четное число, или от 1 до $\frac{(n-1)}{2}$, если n – нечетное число; a_k – поправочный коэффициент[74].

Статистический критерий вычисляется как (31):

$$W = \frac{S^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (31)$$

На языке программирования R данный тест реализуется посредством команды ***shapiro.test(x)*** (функция, выполняющая тест Шапиро–Уилка, содержится в пакете **stats** и подключается автоматически при загрузке RStudio), где ***x*** – вектор числовых данных, в отношении которых высказывается гипотеза о нормальном законе распределения числовой переменной; количество исследуемых значений в ряду *n* должно быть больше 3 и меньше или равно 5000.

В соответствии с требованиями стандарта[77] применение данного критерия рекомендуется в случае, когда распределение является симметричным, т.е. критерий проверки на симметричность $|\beta_1| < \frac{1}{2}$ и критерий проверки на эксцесс $\beta_2 < 3$ или ассиметричное распределение $|\beta_1| > \frac{1}{2}$. В противном случае рекомендуется применение критерия Эппса–Палли.

Пример применения критерия Шапиро–Уилка

Рассмотрим применение критерия Шапиро–Уилка для проверки принадлежности данных к нормальному закону распределения. Для примера используем данные по возрасту пациентов с подтвержденным диагнозом «меланома», содержащиеся в наборе данных ***Melanoma*** пакета MASS.

Листинг 18

```
library("MASS") # Подключаем пакет MASS, содержащий данные Melanoma
dataMale <- Melanoma[Melanoma$sex==1,]$age # Создаем вектор, содержащий
# возраст пациентов мужского пола
dataFemale <- Melanoma[Melanoma$sex==0,]$age # Создаем вектор, содержащий
# возраст пациентов женского пола
hist(dataMale, breaks=12 xlab = "Возраст пациентов с меланомой, лет",
        ylab = "Количество пациентов, чел",
           main = "Распределение возраста мужчин с диагнозом «меланома»",
col="blue")
hist(dataFemale, breaks=12 xlab = "Возраст пациентов с меланомой, лет",
        ylab = "Количество пациентов, чел",
           main = "Распределение возраста женщин с диагнозом «меланома»",
col="blue")
shapiro.test(dataMale) # Применяем тест Шапиро–Уилка для анализа
 # распределения возраста мужчин с диагнозом «меланома»
```

```
  Shapiro-Wilk normality test

data: dataMale
```

[77] ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятности от нормального распределения.

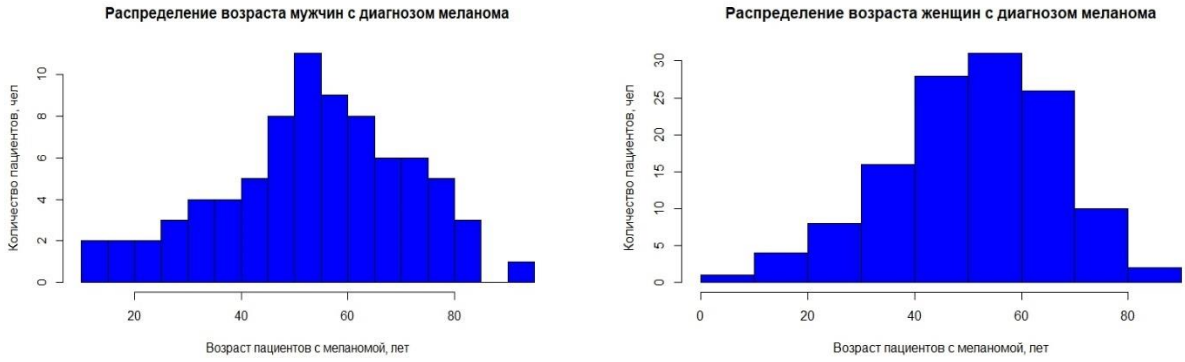
Продолжение листинга 18

```
W = 0.98757, p-value = 0.6442
```

```
shapiro.test(dataFemale) # Применение теста Шапиро–Уилка для анализа
                         # распределения возраста женщин с диагнозом «меланома»
```

```
  Shapiro-Wilk normality test

data: dataFemale
W = 0.98663, p-value = 0.2553
```

На рисунке 23 представлены гистограммы распределения возраста пациентов мужского пола с подтвержденным диагнозом «меланома» и возраста пациентов женского пола с подтвержденным диагнозом «меланома».



а б

Рисунок 23 – Гистограммы распределений возраста пациентов: а – мужского пола, с подтвержденным диагнозом «меланома»; б – женского пола, с подтвержденным диагнозом «меланома»

Анализ результатов теста Шапиро–Уилка показывает, что статистическая гипотеза о нормальном законе распределения возраста пациентов мужского пола выполняется, т.к. p-value = 0,6442 (при уровне статистической значимости, равном 0,05). Распределение возраста пациентов женского пола (рисунок 23б) также соответствует закону, близкому к нормальному, т. к. p-value = 0,2553 теста Шапиро–Уилка при уровне статистической значимости, равном 0,05.

6.5. Критерий Эппса–Палли

В соответствии с рекомендациями стандарта[78] применение данного критерия возможно при условии, если объем выборки составляет не менее 8 исследований. Для выборки $x = (x_1, x_2, \ldots, x_n)$ вычисляются следующие значения (32, 33):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (32)$$

и

[78] ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятности от нормального распределения.

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,, \tag{33}$$

где $\bar{x}$ – среднее арифметическое; m_2 – выборочный центральный момент второго порядка; n – объем выборки.

Статистический критерий T_{EP} Эппса–Палли вычисляют по формуле (34):

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n}\sum_{k=2}^{n}\sum_{i=1}^{k-1}\exp\left\{\frac{-(x_j - x_k)^2}{2m_2}\right\} - \sqrt{2}\sum_{j=1}^{n}\exp\left\{\frac{-(x_j - \bar{x})^2}{4m_2}\right\} \tag{34}$$

На языке программирования R функция проверки данных на принадлежность к нормальному закону распределения по критерию Эппса–Палли осуществляется с использованием функции ***epps.test(),*** находится в пакете ***nortsTest***.

Пример применения критерия Эппса–Палли

Рассмотрим применение критерия Эппса–Палли для проверки гипотезы о нормальном законе распределения выборки на примере распределения концентрации гормона прегнанетриола в сыворотке крови у пациентов с подтвержденным диагнозом синдрома Кушинга. Используем данные ***Cushings*** из пакета ***MASS***.

Листинг 19

```
library("MASS") # Подключаем пакет MASS, содержащий данные Cushings
install.packages("nortsTest") # Устанавливаем библиотеку, содержащую функции epps.test()
library("nortsTest") # Подключаем пакет nortsTest, содержащий функцию
# epps.test()
dataPT <- Cushings$Pregnanetriol
epps.test(dataPT) # Применение теста Эппса–Палли
Epps test

data: dataPT
epps = 5.641, df = 2, p-value = 0.05958
alternative hypothesis: dataPT does not follow a Gaussian Process
hist(dataPT, breaks=12, xlab = "Концентрация прегнанетриола, мг/24 ч ",
        ylab = "Количество пациентов, чел ",
    main = "Распределение концентрации прегнанетриола", col="blue")
```

На рисунке 24 представлена гистограмма распределения, концентрации прегнанетриола в сыворотке крови пациентов с подтвержденным диагнозом синдрома Кушинга.

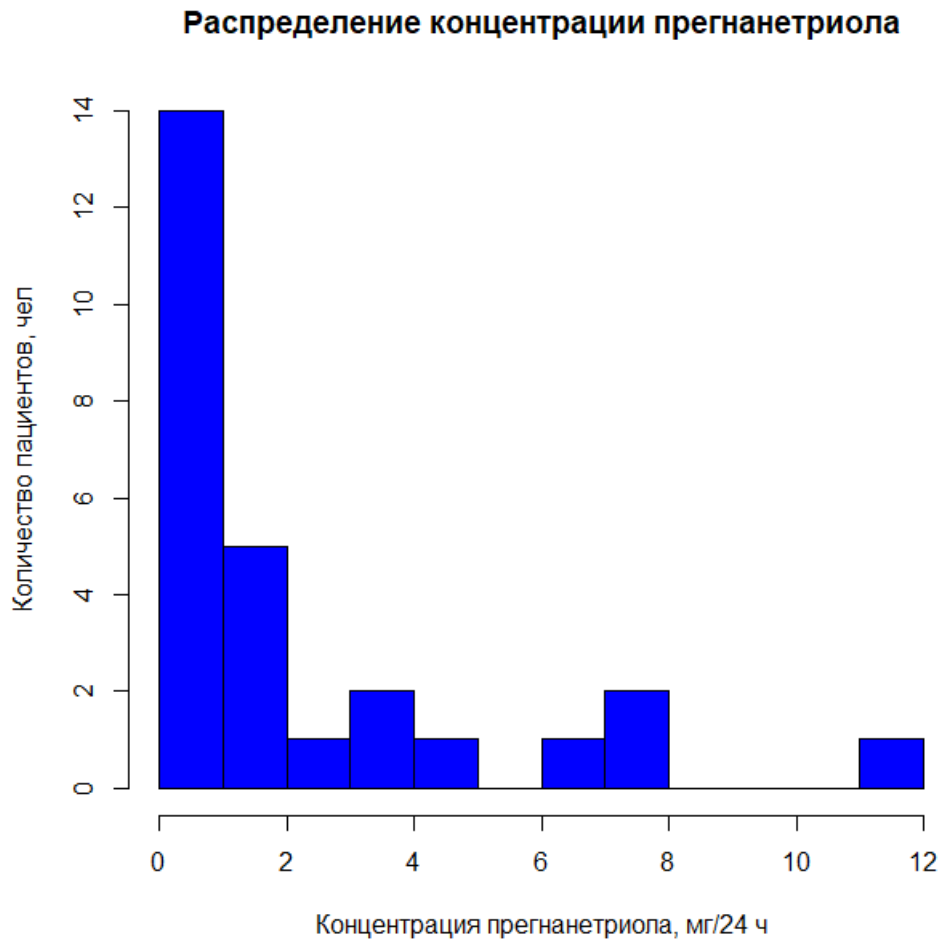


Рисунок 24 – Гистограмма распределения концентрации прегнанетриола в сыворотке крови пациентов с подтвержденным диагнозом синдрома Кушинга

Результаты теста Эппса–Палли показывают, что распределение концентрации прегнанетриола в сыворотке крови пациентов с подтвержденным диагнозом синдрома Кушинга не отлично от нормального закона p-value = 0,05958, что превышает уровень статистической значимости гипотезы.

Необходимо обратить внимание, что визуально гистограмма распределения исследуемой случайной величины не соответствует гистограмме нормального распределения, и скорее всего наблюдается низкая мощность критерия. Понятие мощности статистического критерия рассматривается в следующем подразделе.

6.6. Мощность параметрических статистических критериев

Статистической мощностью называют вероятность отклонения основной (или нулевой) гипотезы H_0 при проверке статистических гипотез в случае, когда конкурирующая (или альтернативная) гипотеза H_1 верна. Чем выше мощность статистического теста, тем меньше вероятность совершить ошибку второго рода. Величина мощности также используется для вычисления размера выборки, необходимой для подтверждения гипотезы с необходимой силой эффекта.

В большинстве случаев, для вычисления статистической мощности применяются методы моделирования, основанные на алгоритме Монте-Карло, далее приводится пример результатов сравнения мощности четырех критериев проверки на нормальность:

– критерия Шапиро–Уилка;
– критерия Дэ'Агустино;
– критерия Эппса–Палли;
– критерия Жарка–Бера.

В качестве исходных данных использовались значения диагностической точности 100 врачей при просмотре 100 исследований пациентов на предмет наличия нормы и патологии на рентгенологических изображениях органов грудной клетки. За нулевую гипотезу было принято предположение о нормальном законе распределения диагностической точности врачей. В качестве альтернативных гипотез были выбраны распределения: логарифмически нормальное, логистическое, Вейбулла и Коши. Для каждого из альтернативных распределений рассчитывались параметры распределений по методу максимального правдоподобия. На рисунке 25 представлены результаты вычисления мощности критериев в зависимости от количества исследований, для нормального закона распределения значений диагностической точности 100 врачей.

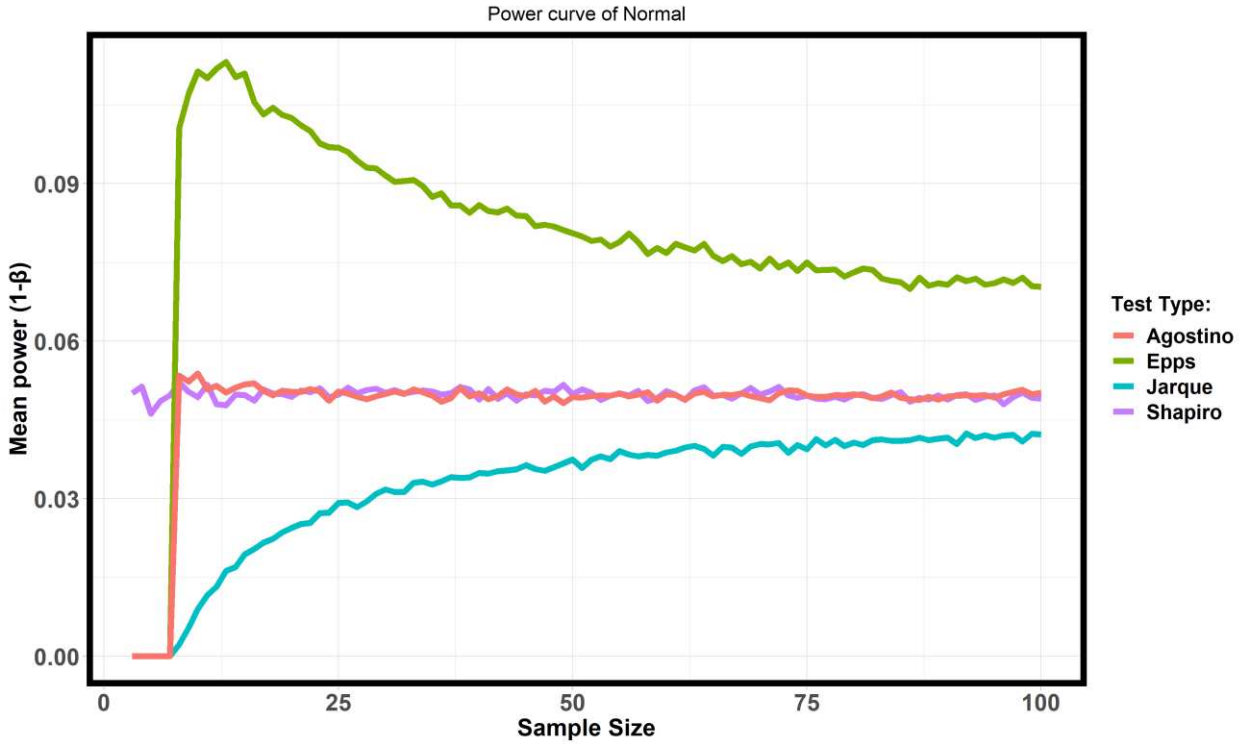


Рисунок 25 – Вычисление мощности тестов на нормальность методом Монте-Карло, в зависимости от объема выборки

Результаты проведенного моделирования показывают, что для критериев Шапиро–Уилка и Дэ'Агустино мощность колеблется вблизи уровня значимости, равного 0,05 для всех объемов выборок. Критерий Жарка–Бера обладает минимальной мощностью среди рассмотренных критериев и также асимптотически стремится к значению 0,04 с увеличением объема выборки. Критерий Эппса–Палли обладает максимальной мощностью среди всех рассмотренных критериев и также асимптотически стремится к значению 0,07 мощности с увеличением объема выборки. Однако в тех случаях, когда данные распределены отличным от нормального закона распределения, картина поведения мощности меняется.

На рисунке 26 представлены результаты моделирования мощности методом Монте-Карло четырех статистических тестов в зависимости от объема выборки, проведенного для четырех типов распределений.

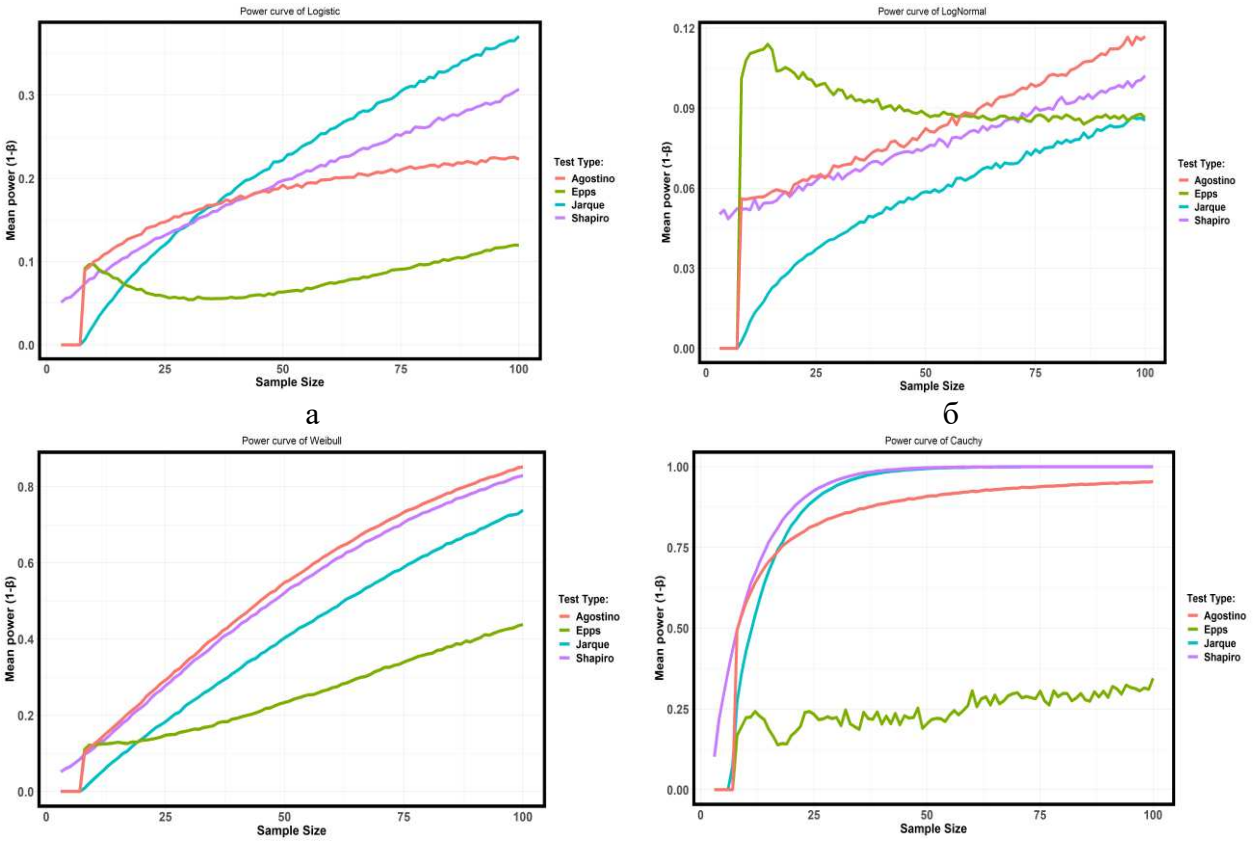


Рисунок 26 – Вычисление мощности тестов на нормальность методом Монте-Карло, в зависимости от объема выборки: а – логистическое распределение; б – логарифмически нормальное распределение; в – двухпараметрическое распределение Вейбулла; г – распределение Коши

Представленные результаты моделирования показывают, что критерий Эппса–Палли обладает максимальной мощностью по сравнению с остальными критериями, только если данные распределены логарифмически нормально, и объем выборки составляет меньше 50 исследований. В остальных случаях данный критерий проигрывает по мощности другим специализированным критериям проверки на нормальность. Для всех типов распределений, и в случае количества исследований больше 100, основными являются критерии Жарка–Бера, Шапиро–Уилка и Дэ'Агустино. Данные критерии имеют устойчивую тенденцию к повышению мощности с увеличением числа исследований, и как следствие, – к снижению вероятности совершить ошибку второго рода. При этом с использованием критериев Жарка–Бера и Шапиро–Уилка существует минимальная вероятность совершить ошибку второго рода при количестве исследований свыше 50 и распределении данных по закону, близкому к распределению Коши.

По совокупности результатов моделирования для проверки выборки на принадлежность к нормальному закону распределения при количестве исследований больше 8 и меньше 5000 рекомендуется использовать критерий Жарка–Бера, Шапиро–Уилка и Дэ'Агустино. Критерий Эппса–Палли не рекомендуется к использованию из-за высокой вероятности получения ошибки второго рода.

Кроме описанных выше параметрических критериев проверки закона распределения случайных величин на нормальность, существуют непараметрические критерии (Колмогорова–Смирнова, Крамера–фон Мизеса, Андерсона–Дарлинга), которые предназначены для проверки простых гипотез[79] и не имеют привязки к нормальному закону распределения.

[79] Понятие простой статистической гипотезы рассматривается в разделе 3.

Пример вычисления мощности критерия методом Монте-Карло

Одним из способов определения наиболее подходящего для анализа данных критерия является расчет его мощности в зависимости от типа распределения и количества данных, которые есть в наличии. На основании распределения концентрации индометацина в сыворотке крови демонстрируется алгоритм расчета средней мощности статистического критерия методом Монте-Карло. В качестве тестового распределения использовано логистическое распределение, количество шагов расчета Монте-Карло составляет 100 000, выборка содержится в наборе данных ***Indometh*** пакета ***MASS***, а средняя мощность вычисляется для критерия Шапиро–Уилка.

Листинг 20

```
library("MASS") # Подключаем пакет, содержащий набор данных Indometh
install.packages("fitdistrplus") # Устанавливаем библиотеку, содержащую пакет "fitdistrplus"
library("fitdistrplus") # Подключаем пакет, содержащий функции вычисления
# параметров распределения методом максимального
# правдоподобия
# Создаем функцию, возвращающую среднюю мощность критерия
#
power_Shapiro_Logis <- function(resp, alpha, loc, sc, sample){
 power <- c() # Создаем пустой вектор, содержащий среднюю мощность
 num <- c() # Создаем пустой вектор, содержащий количество исследований
for (i in 3:sample) { # Цикл, проходящий по всем исследованиям
# Вычисление средней мощности критерия
  test <- mean(replicate(resp,(shapiro.test(rlogis(i, loc, sc))$`p.value`<alpha)))
  power<-c(power,test) # Запись средней мощности критерия для количества
  # исследований
  num <- c(num,i) # количество исследований, для которых рассчитывается
    # средняя мощность
 }
 power <- data.frame(power, samples=num) # Формируем фрейм данных из
  # результатов вычислений
 return(power) # Возвращаем результаты расчета внутри функции
}
repl <- 100000 # Количество повторений метода Монте-Карло
alpha <- 0.05 # Уровень статистической значимости
concIndomet <- Indometh$conc # Создаем вектор, содержащий концентрацию
   # индометацина
samplData <- length(concIndomet) # Определяем количество измерений
# концентрации
  # индометацина
paramDistrib <- fitdist(concIndomet, "logis") # Вычисляем параметры
                    # распределения методом максимального правдоподобия
locationPar <- paramDistrib$estimate[1] # Параметр расположения
    # логистического распределения
scalePar <- paramDistrib$estimate[2] # Параметр ширины логистического
  # распределения
# Результаты вычислений средней мощности теста Шапиро-Уилка для
# фактических данных
        powerSHTest <- power_Shapiro_Logis(resp = repl, alpha = alpha,
loc = locationPar, sc = scalePar, sample = samplData)
```

Продолжение листинга 20

```
# Построение графика зависимости средней мощности
# критерия Шапиро-Уилка от количества исследований
plot(powerSHTest$samples, powerSHTest$power, type = "l",
xlab = "Количество исследований, шт",
ylab = "Средняя мощность критерия",
main = "Зависимость средней мощности критерия Шапиро–Уилка от количества исследований", col = "blue", lwd = 4)
```

На рисунке 27 представлены результаты вычислений средней мощности критерия Шапиро–Уилка в зависимости от количества исследований.

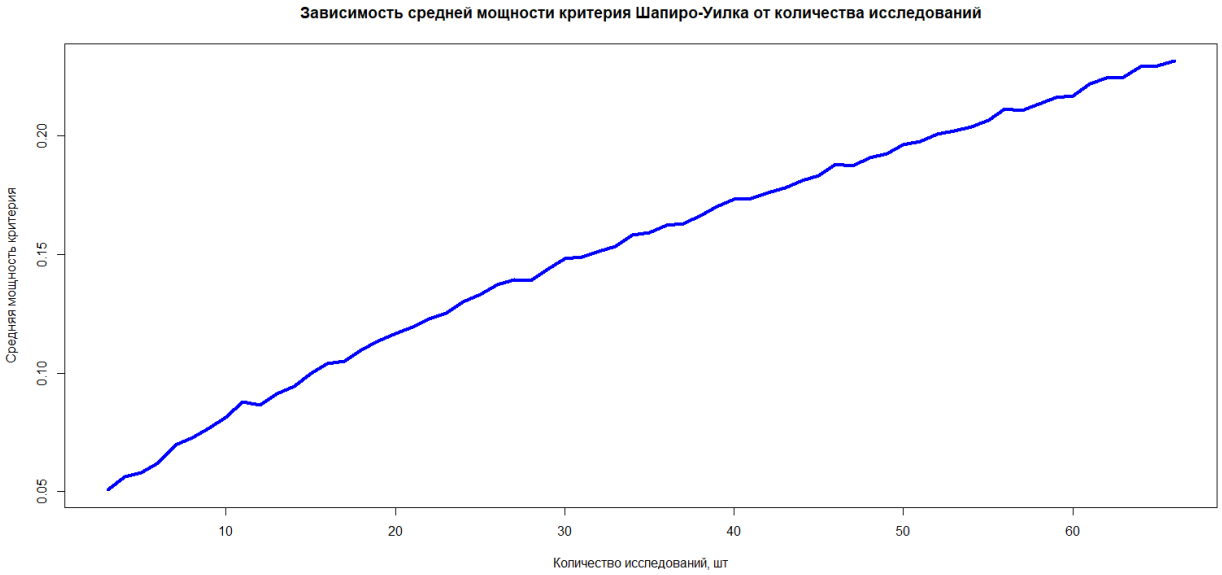


Рисунок 27 – Зависимость средней мощности критерия Шапиро–Уилка от количества исследований

6.7. Непараметрические критерии проверки нулевой гипотезы

Специализированные критерии проверки на принадлежность распределения данных к нормальному закону распределения хорошо применимы тогда, когда у исследователя есть определенная степень уверенности в отсутствии «выбросов» в данных, но на практике такое встречается крайне редко. Более устойчивыми к выбросам являются непараметрические критерии проверки принадлежности к нормальному закону распределения данных[80,81]:

– Колмогорова–Смирнова;
– Крамера–фон Мизеса;
– Андерсона–Дарлинга.

6.7.1. Критерий Колмогорова–Смирнова

Тест Колмогорова–Смирнова относится к категории независящих от параметров распределения тестов для проверки равенства двух непрерывных или дискретных типов одномерных распределений и может быть использован для сравнения выборки с эталонным распределением вероятности. В большинстве случаев в качестве эталонной функции распределения при определении закона распределения выборки используется нормальное распределение, а эмпирическая функция распределения вычисляется по уравнению (35):

$$F_n(X) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) \tag{35}$$

где $1_{(-\infty,x]}(X_i)$ – индикаторная функция равна 1, если $X_i \leq x$ и равна 0 в противном случае.

В случае определения отклонения эмпирической функции распределения от эталонной критерий Колмогорова–Смирнова записывается как (36):

$$D_n = \sup_x |F_n(x) - F(x)| \tag{36}$$

где $F_n(x)$ – эмпирическая функция распределения (35); $F(x)$ – эталонная функция распределения.

Необходимо отметить, что минимальное количество исследований для применения теста Колмогорова–Смирнова составляет 3, а верхняя граница не определена (некоторые авторы рекомендуют применять данный тест при «больших» объемах выборки – не менее 50 исследований)[82]. Также тест Колмогорова–Смирнова позволяет устанавливать наличие статистической значимости в различиях двух случайных величин, и часто используется

[80] Huber P. J. Robust Statistics. New York: John Wiley & Sons, 1981. P. 317.

[81] Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть 2. Непараметрические критерии.

[82] Ядгаров М. Я., Кузовлев А. Н., Берикашвили Л. Б. [и др.]. Важность оценки закона распределения данных: теория и практическое руководство // Анестезиология и реаниматология. 2021. № 2. С. 136–142. DOI: 10.17116/anaesthesiology2021021136.

наравне с тестом Манна–Уитни[83]. На языке R тест Колмогорова–Смирнова может быть проведен с помощью функции ***ks.test()*** из пакета **stats**.

Пример применения критерия Колмогорова–Смирнова

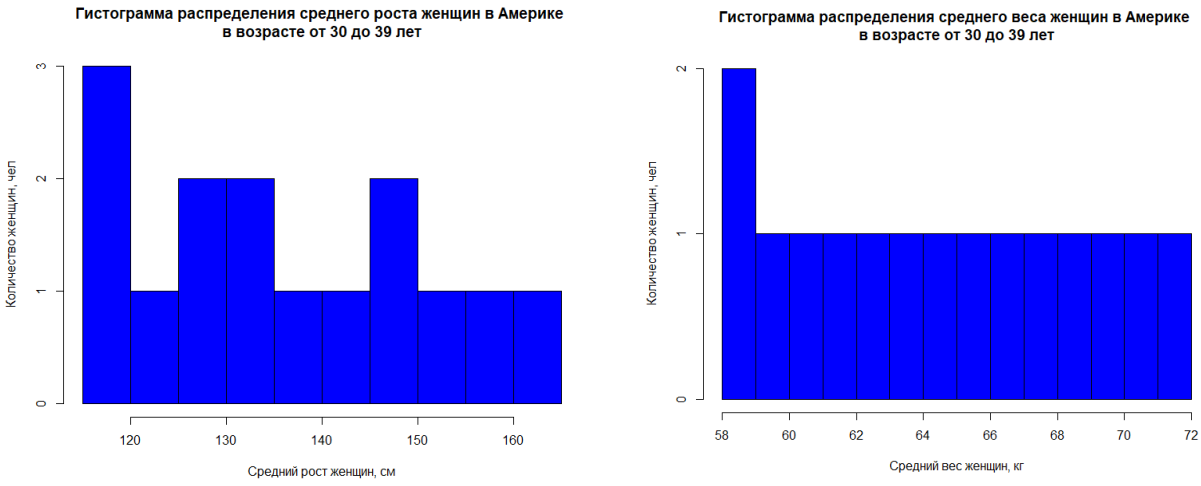
Для демонстрации применения теста Колмогорова–Смирнова используем данные о среднем росте и весе женщин в Америке в возрасте от 30 до 39 лет, содержащиеся в наборе данных ***women*** пакета ***MASS***.

Листинг 21

```
library("MASS") #Подключаем пакет MASS, содержащий набор данных women
heightWomen <- women$height # Вектор, содержащий средний рост женщин
weightWomen <- women$weight # Вектор, содержащий средний вес женщин
ks.test(heightWomen, "pnorm") # Проверяем на принадлежность к нормальному
  # закону распределения средний рост женщин
Exact one-sample Kolmogorov-Smirnov test

data: heightWomen
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

ks.test(weightWomen, "pnorm") # Проверяем на принадлежность к нормальному
                              # закону распределения, средний вес женщин
Exact one-sample Kolmogorov-Smirnov test

data: weightWomen
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

hist(heightWomen, breaks = 12 ,xlab = "Средний рост женщин, см",
ylab = "Количество женщин, чел",
main = "Гистограмма распределения среднего роста женщин в Америке
в возрасте от 30 до 39 лет",
col = "blue")
hist(weightWomen, breaks = 12 ,xlab = "Средний вес женщин, кг",
ylab = "Количество женщин, чел",
main = "Гистограмма распределения среднего веса женщин в Америке
в возрасте от 30 до 39 лет",
col = "blue")
```

На рисунке 28 представлены гистограммы распределения среднего веса и роста женщин в возрасте от 30 до 39 лет в Америке.

[83] Семке В. Я., Харитонов С. В. Сравнительная оценка эффективности когнитивно-поведенческой, рациональной и комбинированной (когнитивно-поведенческой и рациональной) психотерапии у больных личностными расстройствами // Сибирский вестник психиатрии и наркологии. 2011. № 2 (65).



а б

Рисунок 28 – Гистограмма распределения среднего роста (а) и веса (б) женщин в возрасте от 30 до 39 лет в Америке

Результаты теста Колмогорова–Смирнова показывают, что распределение среднего роста и веса женщин в возрасте от 30 до 39 лет в Америке не подчиняется нормальному закону (p-value < 2.2e-16). Гистограммы распределения визуально демонстрируют схожий результат.

6.7.2. Критерий Крамера–фон Мизеса

Критерий Крамера–фон Мизеса является основой для целого семейства критериев, в частности Критерия Крамера–Мизеса–Смирнова, критерия Андерсона–Дарлинга и критерия Фронцини[84]. В основе данных критериев лежит оценка соответствия между эмпирической и теоретической функциями распределения случайной величины (37):

$$\omega^2 = \int_{-\infty}^{\infty} |F_n(x) - F^*(x)|^2 dF^*(x), \tag{37}$$

где $F^*(x)$ – эмпирическая функция распределения случайной величины, вычисленная по уравнению (36); $F(x)$ – теоретическая функция распределения случайной величины.

Для упорядоченной выборки случайных величин статистика Крамера–фон Мизеса записывается, как (38):

$$T = n\omega^2 = \frac{1}{12n} \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2 \tag{38}$$

Превышение точности критерия Крамера–фон Мизеса по сравнению с критерием Крамера–Мизеса–Смирнова обсуждается, в работе Андерсона[85]; критерий Андерсона–

[84] Иванов А. И., Малыгин А. Ю., Полковникова С. А. Удвоение числа статистических критериев семейства Крамера–фон Мизеса дифференцированием малых выборок с нормальным и равномерным распределением биометрических данных // Известия высших учебных заведений. Поволжский регион. Технические науки. 2022. № 1. С. 53–61.

[85] Anderson T. W. On the Distribution of the Two-Sample Cramer-von Mises Criterion // Ann. Math. Statist. 1962. Vol. 33, №3. P. 1148–1159. DOI: 10.1214/aoms/1177704477.

Дарлинга будет рассмотрен в подразделе 7.3. Критерий Фронцини подробно рассмотрен в работе Д. А. Огурцова и С. В. Ушанова[86]. Минимальное количество исследований для корректного использования критерия Крамера–фон Мизеса составляет 8, ограничений по верхнему значению нет. Тест Крамера–фон Мизеса на языке R можно провести, используя функцию ***cvm.test()*** из библиотеки **goftest**.

Пример применения критерия Крамера–фон Мизеса

Демонстрация применения критерия Крамера–фон Мизеса осуществляется с применением набора данных ***Cushings*** из пакета ***MASS***, содержащих скорость выведения с мочой тетрагидрокортизона у пациентов с подтвержденным диагнозом синдрома Кушинга.

Листинг 22

```
library("MASS") # Подключаем пакет, содержащий набор данных Cushings
install.packages("goftest") # Устанавливаем библиотеку, содержащую пакет "goftest"
library("goftest") # Подключаем пакет, содержащий функцию
  # теста Крамера–фон Мизеса
tetraData <- Cushings$Tetrahydrocortisone
cvm.test(tetraData) # Применяем тест Крамера–фон Мизеса
Cramer-von Mises test of goodness-of-fit
Null hypothesis: uniform distribution
Parameters assumed to be fixed

data: tetraData
omega2 = 9, p-value < 2.2e-16
hist(tetraData, breaks = 12 , xlab = "Скорость выведения тетрагидрокортизона, мг/сут",
ylab = "Количество пациентов, чел",
main = "Гистограмма распределения скорости выведения тетрагидрокортизона",
col = "blue")
```

На рисунке 29 представлена гистограмма распределения скорости выведения тетрагидрокортизона у пациентов с подтвержденным диагнозом синдрома Кушинга.

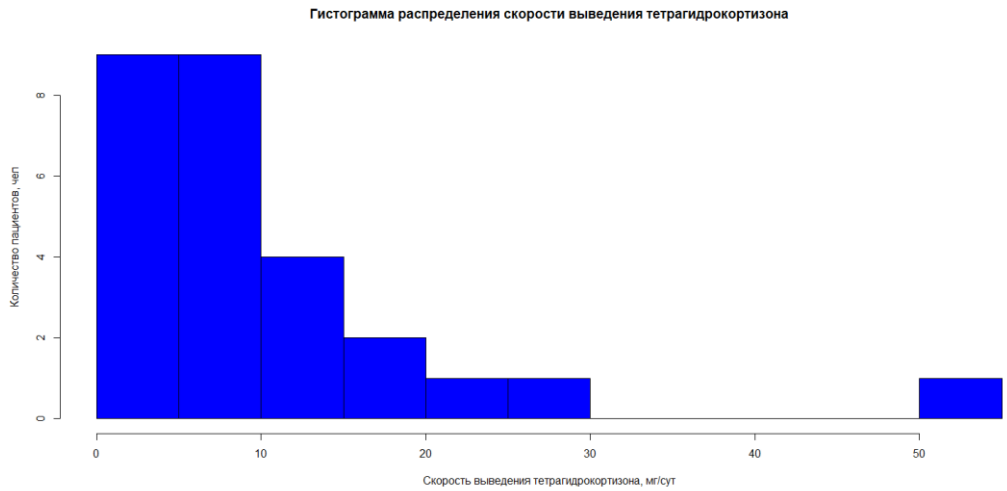


Рисунок 29 – Гистограмма распределения скорости выведения тетрагидрокортизона у пациентов с подтвержденным диагнозом синдрома Кушинга

[86] Огурцов Д. А., Ушанов С. В. Проверка по критерию Фронцини гипотезы о нормальном распределении случайных величин, полученных с округлением // Актуальные проблемы авиации и космонавтики. 2019. Т. 2. С. 283–285.

Визуально по гистограмме распределения скорости выведения тетрагидрокортизона у пациентов с подтвержденным диагнозом синдрома Кушинга можно выдвинуть гипотезу о том, что данные распределены по закону, отличному от нормального, и применение теста Крамера–фон Мизеса подтверждает данную гипотезу (p-value < 2.2e-16). Стоит отметить, что критерий Крамера–фон Мизеса устойчив к наличию в данных совпадающих значений в отличие от критерия Колмогорова–Смирнова.

6.7.3. Критерий Андерсона–Дарлинга

Как было отмечено ранее, критерий Андерсена–Дарлинга является развитием идеи критерия Крамера–фон Мизеса. В основе критерия лежит интегральная оценка соответствия между эмпирической и теоретической функциями распределения случайной величины с учетом весовых коэффициентов (39):

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 * \rho(x) dF(x), \quad (39)$$

где $F_n(x)$ – эмпирическая функция случайной величины; $F(x)$ – теоретическая функция случайной величины; $\rho(x)$ – весовая функция.

Для упорядоченной выборки случайных величин статистика Андерсена–Дарлинга записывается как (40):

$$A^2 = n * \omega^2 = n * \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x) * \big(1 - F(x)\big)} dF(x), \quad (40)$$

где $\rho(x) = \frac{1}{F(x)(1-F(x))}$ – весовая функция.

Для проведения проверки данных на принадлежность к нормальному закону распределения с помощью критерия Андерсена–Дарлинга на языке R предусмотрена функция ***ad.test()*** из пакета **nortest**.

Пример применения критерия Андерсена–Дарлинга

Рассмотрим пример применения критерия Андерсона–Дарлинга на наборе данных ***GAGurine*** из пакета ***MASS***, содержащий концентрацию гликозаминогликанов (GAG) в моче у детей в возрасте от 0 до 17 лет.

Листинг 23

```
library("MASS") # Подключаем пакет, содержащий набор
  # данных GAGurine
library("nortest") # Подключаем пакет, содержащий функцию
  # теста Андерсона–Дарлинга
gagData <- GAGurine$GAG # Создаем вектор, содержащий данные
            # концентрации GAG в моче
ad.test(gagData) #Применяем тест Андерсона–Дарлинга к концентрации GAG
Anderson-Darling normality test

data: gagData
A = 10.396, p-value < 2.2e-16
```

Продолжение листинга 23

```
hist(gagData, breaks = 12, xlab = "Концентрация гликозаминогликанов в моче, мг/мл",
ylab = "Частота встречаемости значений",
main = "Гистограмма распределения концентрации гликозаминогликанов в моче детей в возрасте от 0 до 17
лет", col = "blue")
```

На рисунке 30 представлена гистограмма распределения GAG в моче детей в возрасте от 0 до 17 лет.

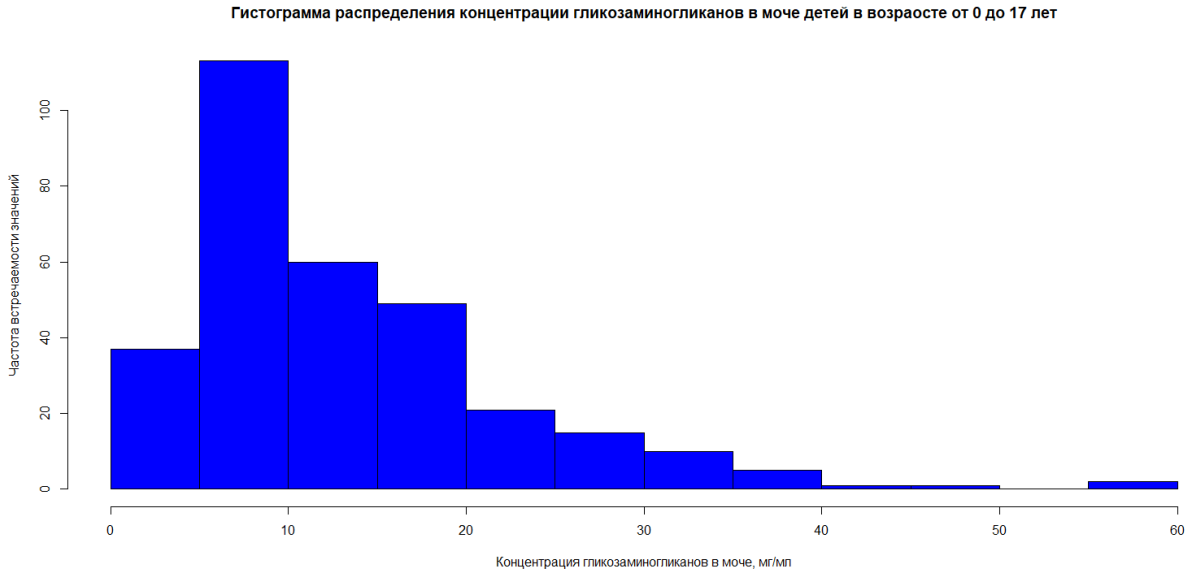


Рисунок 30 – Гистограмма распределения концентрации GAG в моче детей в возрасте от 0 до 17 лет

Визуальный анализ гистограммы распределения концентрации GAG в моче детей в возрасте от 0 до 17 лет позволяет выдвинуть нулевую гипотезу об отличии распределения данных от нормального закона. Применение теста Андерсона–Дарлинга подтверждает данную гипотезу (p-value < 2.2e-16).

Выбор применяемого для анализа близости распределения данных к нормальному закону распределения осуществляется на основании вычисления мощности критерия в зависимости от количества исследований. Для непараметрических критериев проверки на соответствие данных нормальному закону распределения были рассчитаны зависимости мощности критерия от числа исследований для метрик диагностической точности 100 врачей.

6.7.4. Мощность непараметрических статистических критериев

Так же, как и в случае специализированных критериев проверки на принадлежность данных к нормальному закону распределения для критериев Колмогорова–Смирнова, Крамера–фон Мизеса и Андерсона–Дарлинга, с помощью алгоритма Монте-Карло было проведено моделирование зависимости мощности критерия от числа исследований.

В качестве исходных данных использовались значения диагностической точности 100 врачей при просмотре 100 исследований на предмет наличия нормы и патологии. За нулевую гипотезу было принято предположение о нормальном распределении диагностической точности врачей. На рисунке 31 представлены результаты моделирования.

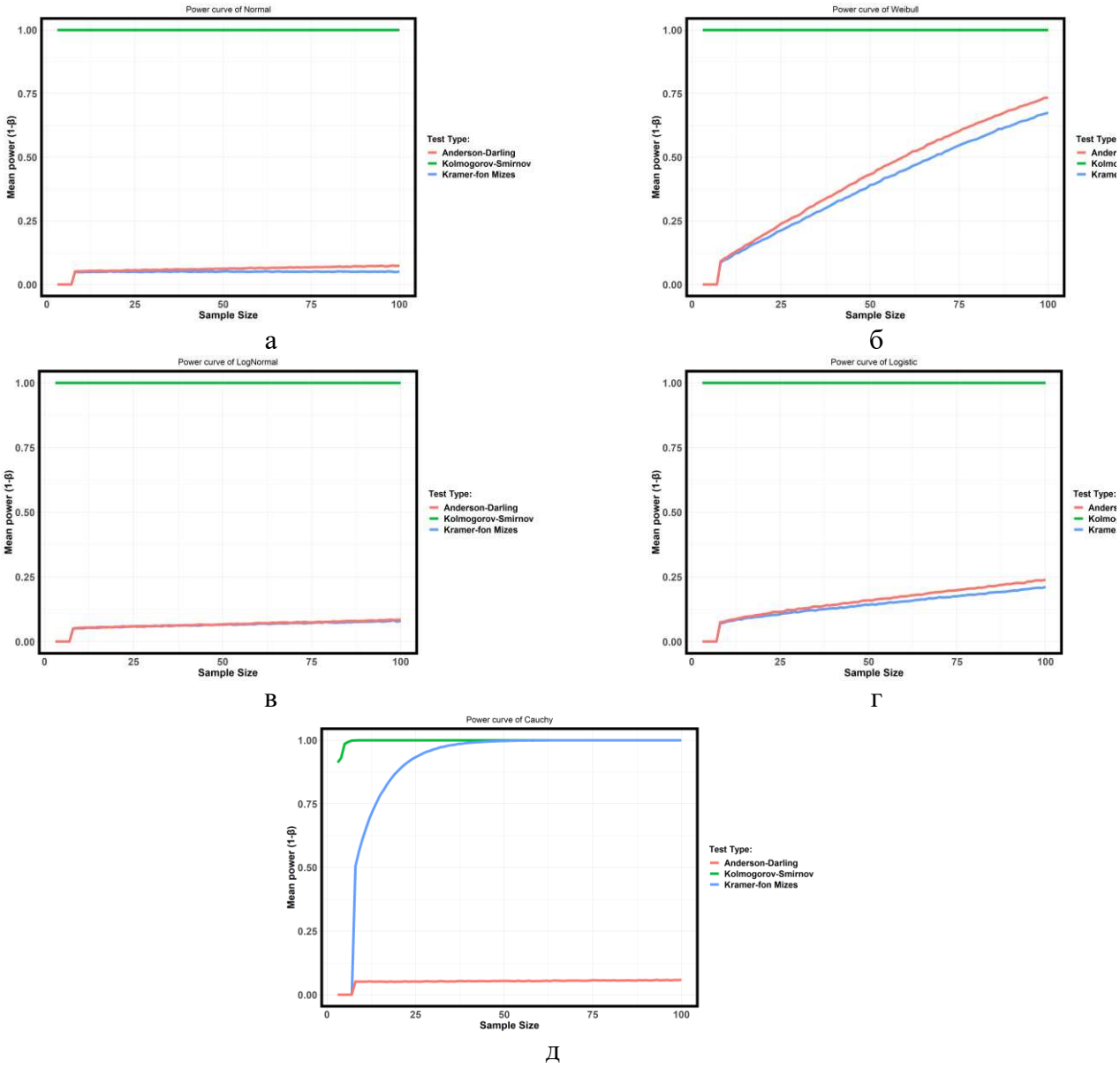


Рисунок 31 – Моделирование мощности методом Монте-Карло трех непараметрических критериев Колмогорова–Смирнова (зеленая кривая), Крамера–фон Мизеса (голубая кривая) и Андерсона–Дарлинга (красная линия) в зависимости от количества исследований для пяти типов входных распределений: а – нормальное; б – Вейбулла; в – логарифмически нормальное; г – логистическое; д – Коши

Результаты моделирования показывают, что для всех исследованных типов распределений, поданных на вход критериям Колмогорова–Смирнова, Крамера–фон Мизеса и Андерсона–Дарлинга, максимальной мощностью обладает критерий Колмогорова–Смирнова. Исключение составляет случай, когда данные имеют распределение Коши, и количество исследований превышает число 30. В этом случае мощность критериев Колмогорова–Смирнова и Крамера–фон Мизеса имеет близкие

к единице значение. Стоит отметить, что при малом количестве исследований (до 10) критерии обладают мощностью меньше 1.

Моделирование мощности различных критериев в зависимости от количества исследований метрик диагностической точности 100 врачей показало, что мощность критерия Колмогорова–Смирнова максимальна, и вероятность совершить ошибку второго рода при его применении на исследуемых данных практически отсутствует. Представленный результат также показывает, что мощность теста Колмогорова–Смирнова не меняется в зависимости от типа распределения и количества исследований свыше 10.

Применение критерия Колмогорова–Смирнова к метрикам диагностической точности 100 врачей с нулевой гипотезой – данные распределены по нормальному закону и альтернативной гипотезой – данные распределены отлично от нормального закона показывает значения D = 0,76331 и p-value = 2,26e-16. Этот результат теста Колмогорова–Смирнова свидетельствует о том, что нулевая гипотеза неверна, и данные распределены по закону, отличному от нормального.

Дальнейшая работа по анализу данных может быть выстроена по нескольким сценариям в зависимости от наличия или отсутствия дополнительных признаков:

1. Проведение сравнительного анализа между группами, выделенными по дополнительным признакам, на предмет наличия или отсутствия статистически значимых различий между этими группами.
2. Установление наличия или отсутствия статистически значимой связи между двумя и более численными выборками (поиск корреляций).
3. Построение эмпирической модели, описывающей выявленные закономерности.

Пример расчета мощности непараметрических критериев проверки принадлежности данных к нормальному распределению

На примере набора данных ***GAGurine*** из пакета ***MASS***, содержащего информацию о концентрации гликозаминогликанов (GAG) в моче у детей в возрасте от 0 до 17 лет, рассмотрим вычисление средней мощности непараметрического статистического критерия Крамера–фон Мизеса в зависимости от количества исследований. В качестве модельного будет использовано распределение Вейбулла.

Листинг 24

```
library("MASS") # Подключаем пакет, содержащий набор данных Indometh
library("fitdistrplus") # Подключаем пакет, содержащий функции вычисления
# параметров распределения методом максимального
# правдоподобия
library("goftest") # Подключаем библиотеку, содержащую функцию теста
   # Крамера–фон Мизеса
# Создаем функцию, возвращающую среднюю мощность критерия
power_KFM_Weibull <- function(resp, alpha, dataAnaliz, sample){
 power <- c() # Создаем пустой вектор, содержащий среднюю мощность
 num <- c() # Создаем пустой вектор, содержащий количество исследований
for (i in 3:sample) { # Цикл, проходящий по всем исследованиям
paramDistrib <- fitdist(dataAnaliz[1:i], "weibull") # Вычисляем параметры
     # распределения методом максимального правдоподобия
loc <- paramDistrib$estimate[1] # Параметр расположения
    # логистического распределения
sc <- paramDistrib$estimate[2] # Параметр ширины логистического
  # распределения
  # Вычисление средней мощности критерия
```

Продолжение листинга 24

```
 test <- mean(replicate(resp,(cvm.test(rweibull(i, loc, sc))$`p.value`<alpha)))
  power<-c(power,test) # Запись средней мощности критерия для количества
  # исследований
  num <- c(num,i) # количество исследований, для которых рассчитывается
    # средняя мощность
 }
 power <- data.frame(power, samples=num) # Формируем фрейм данных из
  # результатов вычислений
 return(power) # Возвращаем результаты расчета внутри функции
}
repl <- 100000 # Количество повторений метода Монте-Карло
alpha <- 0.05 # Уровень статистической значимости
gagData <- GAGurine$GAG # Создаем вектор, содержащий концентрацию
   # GAG
samplData <- length(gagData) # Определяем количество измерений
# концентрации GAG
# Результаты вычислений средней мощности теста Крамера–фон Мизеса для
# фактических данных
        powerKFMtest <- power_KFM_Weibull(resp = repl, alpha = alpha,
dataAnaliz = gagData, sample = samplData)
# Построение графика зависимости средней мощности
# критерия Крамера–фон Мизеса от количества исследований
plot(powerKFMtest$samples, powerKFMtest$power, type = "l",
xlab = "Количество исследований, шт",
ylab = "Средняя мощность критерия",
main = "Зависимость средней мощности критерия Крамера–фон Мизеса от
количества исследований", col = "blue", lwd = 4)
```

Результат расчета средней мощности критерия Крамера–фон Мизеса в зависимости от количества исследований представлен на рисунке 32.

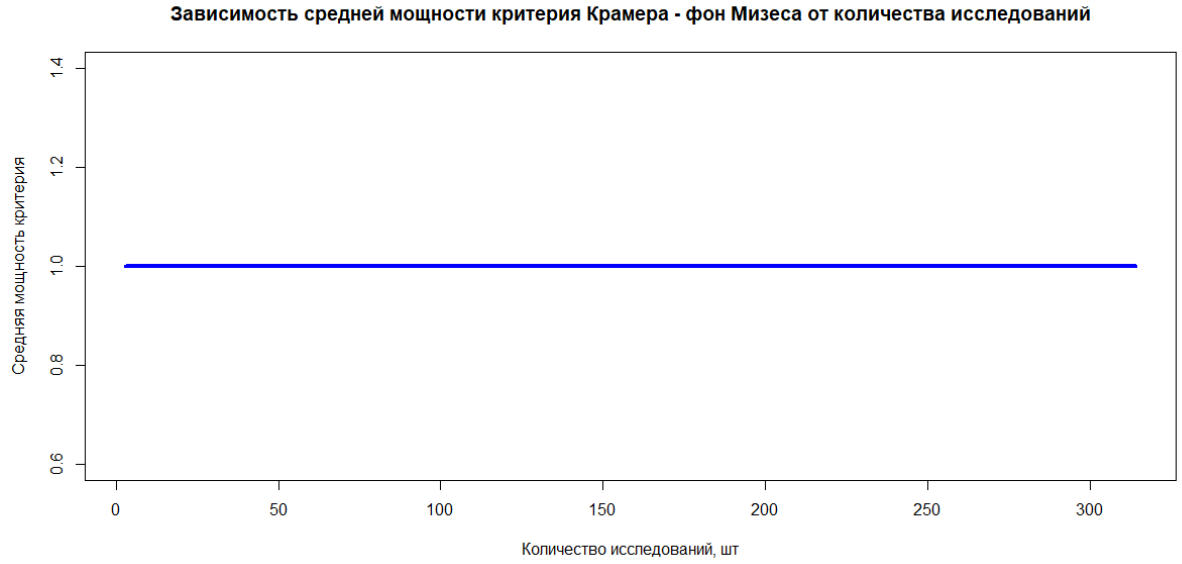


Рисунок 32 – Результаты вычислений средней мощности теста Крамера–фон Мизеса в зависимости от количества исследований, содержащихся в наборе данных ***GAGurine*** *из пакета MASS*

Анализ результатов вычисления мощности теста Крамера–фон Мизеса на наборе данных GAGurine из пакета MASS показывает, что данный тест обладает 100-процентной мощностью при проведении исследований на нормальность.

ЗАКЛЮЧЕНИЕ

В первой части методических рекомендаций представлены основные понятия, принятые в статистическом анализе данных. Также описаны основные величины и методы их вычисления, принятые в описательной (базовой) статистике. Кроме того, рассмотрены критерии проверки данных на соответствие нормальному закону распределения, представлен метод выбора критерия проверки на соответствие нормальному закону распределения на основе метода Монте-Карло.

Способы вычисления всех рассматриваемых в первой части статистических критериев и величин демонстрируются с использованием медико-биологических наборов данных и программного кода, реализованного на языке программирования R.

Представленный в методических рекомендациях материал может быть использован как для проведения занятий по основам статистического анализа медицинских данных на языке программирования R, так и для проведения научно-исследовательских работ по различным направлениям и специализациям.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Мастицкий С. Э., Шитиков В. К. Статистический анализ и визуализация данных с помощью R. М.: ДМК Пресс, 2015. 496 с.

2. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. М.: ДМК Пресс, 2016. 450 с.

3. Мэтлофф Н. Искусство программирования на R. Погружение в большие данные. СПб.: Питер, 2019. 416 с.

4. Кабаков Р. И. R в действии. Анализ и визуализация данных с использованием R и Tidyverse. 3-е изд. М.: ДМК Пресс, 2023. 768 с.

Серия «Лучшие практики лучевой и инструментальной диагностики»

Выпуск 139

Авторы-составители:

Васильев Юрий Александрович
Никитин Никита Юрьевич
Будыкина Анна Владимировна
Памова Анастасия Петровна
Бобровская Татьяна Михайловна
Арзамасов Кирилл Михайлович

ПРОВЕДЕНИЕ СТАТИСТИЧЕСКОГО АНАЛИЗА НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ R В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Часть 1

Методические рекомендации

Отдел координации научной деятельности ГБУЗ «НПКЦ ДиТ ДЗМ»
Технический редактор
Компьютерная верстка

ГБУЗ «НПКЦ ДиТ ДЗМ»
127051, г. Москва, ул. Петровка, д. 24, стр. 1