

**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



«23» октября 2024 г.

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

Часть 2

Методические рекомендации 59

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 2 / авт.-сост. Ю. А. Васильев, Н. Ю. Никитин, А. В. Будыкина [и др.]. – Вып. 140. – М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2024. – 85 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.

Во второй части рассмотрен корреляционный анализ как один из основных способов установления наличия взаимосвязи между переменными, присутствующими в исследуемых наборах данных. Изложены методы сравнительного анализа данных в группах, а также рассмотрены методы построения моделей на основе логистической регрессии и их оценка. Все примеры, приведенные в тексте издания, реализованы на языке программирования R на открытых наборах данных, входящих в пакеты этого языка программирования.

Вторая 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»

Данный документ является собственностью Департамента здравоохранения города Москвы, не подлежит тиражированию и распространению без соответствующего разрешения

ISBN

©Департамент здравоохранения города Москвы, 2024
© ГБУЗ «НПКЦ ДиТ ДЗМ», 2024
© Коллектив авторов, 2024

СОДЕРЖАНИЕ

| | |
|--|----|
| НОРМАТИВНЫЕ ССЫЛКИ..... | 4 |
| ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ..... | 5 |
| ВВЕДЕНИЕ..... | 6 |
| 1. СТАТИСТИЧЕСКАЯ СВЯЗЬ МЕЖДУ ДАННЫМИ..... | 7 |
| 1.1. Корреляционный анализ..... | 7 |
| 1.1.1. Корреляция по Пирсону..... | 8 |
| 1.1.2. Корреляция по Спирмену..... | 9 |
| 1.1.3. Корреляция по Кендаллу..... | 13 |
| 1.1.4. Шкалы «силы» корреляции..... | 17 |
| 1.1.5. Замечания к результатам корреляционного анализа..... | 18 |
| 2. СРАВНЕНИЕ ДАННЫХ В ГРУППАХ..... | 23 |
| 2.1. Дисперсионный анализ ANOVA..... | 23 |
| 2.1.1. Частный случай дисперсионного анализа. Метод MRMC..... | 27 |
| 2.2. t-критерий Стьюдента..... | 30 |
| 2.3. F-критерий Фишера..... | 37 |
| 2.4. Критерий Крускала–Уоллиса..... | 39 |
| 2.5. Критерий Данна..... | 41 |
| 2.6. Критерий Манна–Уитни..... | 43 |
| 2.7. Поправки на множественную проверку гипотез..... | 45 |
| 2.8. Источники наборов данных для проведения самостоятельного статистического анализа..... | 54 |
| 3. СТАТИСТИЧЕСКИЕ КРИТЕРИИ СОГЛАСИЯ..... | 55 |
| 3.1. Критерий согласия X^2 Пирсона..... | 55 |
| 3.2. Графический метод определения согласия Бланда–Альтманна..... | 57 |
| 4. РЕГРЕССИОННЫЙ АНАЛИЗ. ЛИНЕЙНАЯ РЕГРЕССИЯ..... | 62 |
| 5. РЕГРЕССИОННЫЙ АНАЛИЗ. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ..... | 67 |
| 5.1. Оценка моделей, построенных на основе логистической регрессии..... | 70 |
| 5.2. Оценка корректности классификации..... | 71 |
| 5.3. Тест Макнемара..... | 74 |
| 5.4. Индекс Юдена..... | 77 |
| 5.5. Метрика диагностической точности ROC-AUC..... | 80 |
| ЗАКЛЮЧЕНИЕ..... | 83 |
| СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ..... | 84 |

НОРМАТИВНЫЕ ССЫЛКИ

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

– Рекомендации по стандартизации Р 50.1.033-2001. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть 1. Критерий Хи-квадрат.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ГОСТ – государственный стандарт.

Листинг – распечатка программного кода, написанного на языке R.

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.

$\{\dots\}$ – в математических выражениях в фигурных скобках обозначены «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...».

“ \in ” – данный знак читается как «принадлежит». Например, $X:\{P \in \mathcal{P}\}$ читается как: «величина X определена на множестве P , принадлежащего к \mathcal{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).

ВВЕДЕНИЕ

Вторая часть методических рекомендаций по проведению статистического анализа на языке программирования R в медико-биологических исследованиях является логическим продолжением первой части методических рекомендаций.

В данной части рассматриваются особенности проведения корреляционного анализа, обозначены нюансы интерпретации результатов корреляционного анализа медико-биологических данных.

Одной из особенностей анализа данных, содержащих результаты исследований пациентов, дифференцированных по разным группам, является сравнение мер центральной тенденции (средних, медиан) величин проявления эффекта в группе. Этому вопросу посвящен отдельный раздел второй части методических рекомендаций. Рассматриваются разные критерии сравнения групп по количественным переменным.

В конце издания описываются два базовых метода построения предсказательных и классифицирующих моделей, а также приводятся способы их оценки, настройки и сравнения результатов. Так же, как и в первой части методических рекомендаций, все рассмотренные критерии и методы иллюстрируются на открытых медико-биологических данных и посредством программного кода, реализованного на языке программирования R.

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

1. СТАТИСТИЧЕСКАЯ СВЯЗЬ МЕЖДУ ДАННЫМИ

Установление наличия статистической взаимосвязи между исследуемыми данными может выражаться в виде нескольких различных задач:

1. Поиск статистической зависимости между количественными данными.
2. Поиск взаимосвязи статистически независимых групп данных.

Для решения первой задачи наиболее часто применяют методы корреляционного анализа, которые можно разделить (так же, как и критерии проверки на нормальность) на две группы:

1. Параметрические
2. Непараметрические.

Далее каждая из них будет рассмотрена более подробно.

Решение второй задачи – процесс творческий и зависит от того, присутствует

ли в данных дополнительная информация (возможно ли выделение групп за счет наличия уровней в факторах, или нет). Если дополнительная информация присутствует,

то возможно выделение групп по дополнительным признакам с последующим установлением наличия взаимосвязи между ними по статистическим критериям.

Из непараметрических критериев наиболее частое употребление имеют:

- 1) критерий Колмогорова–Смирнова;
- 2) критерий Манна–Уитни;
- 3) критерий Краскелла–Уоллиса;

Из параметрических критериев:

- 1) дисперсионный анализ;
- 2) критерий Стьюдента;
- 3) критерий Фишера.

Рассмотрим более подробно решение каждой из представленных задач.

1.1. Корреляционный анализ

При проведении корреляционного анализа необходимо понимать, соответствуют ли исследуемые данные нормальному закону распределению или нет. Если данные нормально распределены, то исследования корреляции ведутся по методу Пирсона, в противном случае используется корреляция по Кендаллу или Спирмену.

Для оценки силы корреляции (взаимосвязи) существует ряд общепризнанных шкал^{1, 2}.

Для оценки доли влияния одной группы наблюдений на другую применяют коэффициент детерминации, равный квадрату коэффициента корреляции и выражаемый в процентах.

¹ Баврина А. П., Борисов И. Б. Современные правила применения корреляционного анализа // Медицинский альманах. 2021. №3(68).

² Котеров А. Н., Ушенкова Л. Н., Зубенкова Э. С. [и др.]. Сила связи. Сообщение 2. Градации величины корреляции // Медицинская радиология и радиационная безопасность. 2019. Т. 64, № 6. С. 12–24.

1.1.1. Корреляция по Пирсону

Вычисление коэффициентов корреляции по Пирсону производится для двух случайных величин, каждая из которых содержит группу наблюдений, т.е. $X = (x_1, x_2, \dots, x_n)$ и $Y = (y_1, y_2, \dots, y_n)$ и является нормально распределенной величиной. Коэффициент корреляции вычисляется по уравнению (1):

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{s_x s_y}, \quad (1)$$

где $\text{cov}(x, y)$ – матрица ковариации; s_x – среднее квадратичное отклонение в наблюдениях X; s_y – среднее квадратичное отклонение в наблюдениях Y.

Матрица ковариации вычисляется по уравнению (2):

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (2)$$

Среднее квадратичное отклонение в наблюдениях вычисляется по уравнениям (3, 4):

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (3)$$

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}} \quad (4)$$

На языке программирования R коэффициент корреляции вычисляется с помощью команды `cor(x, y, method="pearson")`, проверка статистической значимости проводится командой `cor.test(x, y, method="pearson")`.

Данные по метрикам диагностической точности 100 врачей содержат время, за которое каждый врач прошел тест. Необходимо установить, есть ли между временем прохождения теста и метриками диагностической точности корреляция или нет. Коэффициент корреляции по Пирсону между метриками диагностической точности и временем прохождения теста равен 0,2720272, уровень статистической значимости $p = 0,006183$, что составляет значение меньше 0,05, а коэффициент детерминации составляет 7,4 %. Таким образом, вывод из результатов корреляционного анализа по Пирсону следующий: выявлена слабая статистически значимая корреляция между метриками диагностической точности и временем прохождения тестирования, при этом время тестирования влияет на метрики диагностической точности в 7,4 % случаев. Однако этот вывод не может считаться достоверным, т.к. данные не распределены по нормальному закону.

Пример корреляционного анализа по Пирсону

Используя набор данных **Cushings** из пакета MASS, содержащий информацию о скорости выведения с мочой тетрагидрокортизона и прегнантриола у пациентов с подтвержденным диагнозом синдрома Кушинга, исследуем корреляцию по Пирсону.

Листинг 1

```
library("MASS") # Подключаем пакет, содержащий набор данных
Cushings
dataTetra <- Cushings$Tetrahydrocortisone # Создаем вектор, содержащий
# скорость выведения тетрагидрокортизона
dataPregna <- Cushings$Pregnanetriol # Создаем вектор, содержащий
скорость
# выведения прегнанетриола
cor.test(dataTetra, dataPregna, method="pearson") # Проводим
корреляционный
# анализ по Пирсону
Pearson's product-moment correlation
data: dataTetra and dataPregna
t = 0.31517, df = 25, p-value = 0.7552
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3248706 0.4325823
sample estimates:
 cor
0.0629099
```

Результаты корреляционного анализа показывают наличие статистически незначимой ($p\text{-value} = 0,7552$) очень слабой корреляции ($\text{cor} = 0,0629099$). Это свидетельствует о независимости скорости выведения тетрагидрокортизона от скорости выведения прегнанетриола.

1.1.2. Корреляция по Спирмену

Коэффициент корреляции Спирмена относится к непараметрическим критериям проверки корреляционной связи между переменными, измеряемыми по ранговой шкале. В основе вычисления коэффициента корреляции по Спирмену отсутствуют предположения о характере распределения, что показывает тесноту связи между ранжируемыми величинами [1]. Коэффициент корреляции по Спирмену вычисляется как (5):

$$\rho = 1 - \frac{6 \sum_{i=1}^n (D^2)}{n(n^2 - 1)}, \quad (5)$$

где n – количество ранжируемых значений; D – разница между рангами двух переменных, вычисляется как (6):

$$D = R(X) - R(Y) \quad (6)$$

На языке программирования R коэффициент корреляции по Спирмену вычисляется с помощью команды `cor(x, y, method="spearman")`, проверка статистической значимости проводится командой `cor.test(x, y, method="spearman")`.

Проведем вычисления коэффициента корреляции по Спирмену для метрик диагностической точности 100 врачей, оценим уровень статистической значимости и коэффициент детерминации. Коэффициент корреляции составил 0,1969877, уровень статистической значимости $p = 0,04948$, а коэффициент детерминации – 3,9 %. Вывод из вычисления коэффициента корреляции по Спирмену: между

метриками диагностической точности и скоростью прохождения врачом теста существует слабая линейная связь, проявляющаяся в 3,9 % случаев.

Пример проведения корреляционного анализа по Спирмену

Применение непараметрических критериев тесно связано с оценкой типа распределения, которому подчиняются данные. При проведении анализа необходимо понять, подчиняются ли анализируемые данные нормальному закону распределения или нет. После проверки данных на принадлежность нормальному закону распределения принимается решение о применении корреляционного анализа по Пирсону или Спирмену.

На основе данных, содержащихся в *Cushings* из пакета MASS, анализируется скорость выведения с мочой тетрагидрокортизона и прегнанетриола у пациентов с подтвержденным диагнозом синдрома Кушинга.

Листинг 2

```
library("MASS") # Подключаем пакет, содержащий набор данных  
Cushings  
library("nortest") # Подключаем пакет, содержащий функцию  
# теста Андерсона–Дарлинга  
dataTetra <- Cushings$Tetrahydrocortisone # Создаем вектор, содержащий  
# скорость выведения тетрагидрокортизона  
dataPregna <- Cushings$Pregnanetriol # Создаем вектор, содержащий  
# выведения прегнанетриола  
# скорость  
hist(dataTetra, breaks=12,  
      xlab = "Скорость выведения тетрагидрокортизона, мг/сут",  
      ylab = "Количество пациентов, чел",  
      main = "Гистограмма распределения скорости выведения  
тетрагидрокортизона",  
      col = "blue")  
# На рисунке 1 представлена гистограмма распределения скорости  
выведения  
# тетрагидрокортизона
```

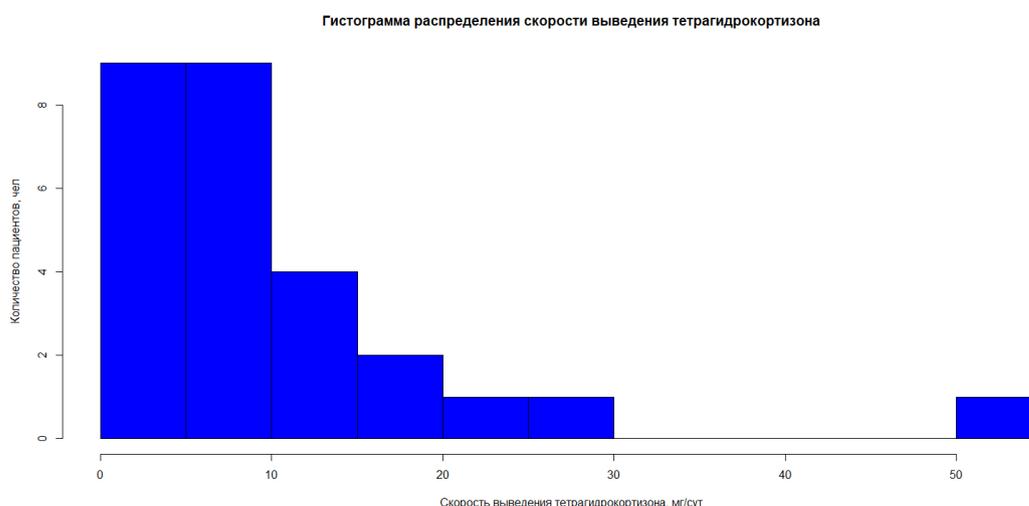


Рисунок 1 – Гистограмма распределения скорости выведения тетрагидрокортизона у пациентов с подтвержденным диагнозом синдрома Кушинга

Продолжение листинга 2

```
ad.test(dataTetra) # Проверяем данные о скорости выведения с мочой
тетрагидрокортизона на
# принадлежность к нормальному закону распределения
# тестом Андерсона–Дарлинга
# Результаты применения теста Андерсона–Дарлинга к данным о
скорости
# выведения тетрагидрокортизона

Anderson-Darling normality test
data: dataTetra
A = 2.3876, p-value = 3.222e-06
# Построение гистограммы распределения скорости выведения
прегнанетриола
hist(dataPregna, breaks=12,
      xlab = "Скорость выведения прегнанетриола, мг/сут",
      ylab= "Количество пациентов, чел",
      main= "Гистограмма распределения скорости выведения
прегнанетриола",
      col = "blue")
# На рисунке 2 представлена гистограмма распределения скорости
выведения
# прегнанетриола
```

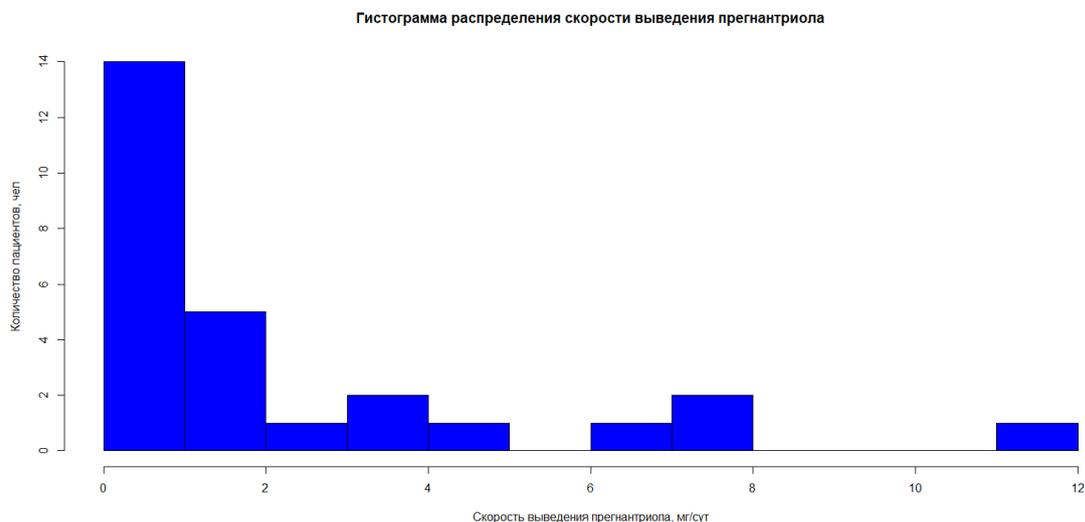


Рисунок 2 – Гистограмма распределения скорости выведения прегнанетриола пациентов с подтвержденным диагнозом синдрома Кушинга

```
ad.test(dataPregna) # Проверяем данные о прегнанетриоле на
принадлежность
# к нормальному закону распределения тестом Андерсона–Дарлинга
# Результаты применения теста Андерсона–Дарлинга к данным о
скорости
# выведения прегнанетриола
```

```
Anderson-Darling normality test
data: dataPenta
A = 2.8542, p-value = 2.171e-07
```

Продолжение листинга 2

```
#=====
=====
#По результатам теста Андерсона–Дарлинга данные распределены
отличным
# от нормального закона распределения, поэтому для оценки корреляции
применяем метод
# Спирмена
#=====
=====
plot(dataPregna, dataTetra,
xlab= "Скорость выведения прегнанетриола, мг/сут",
ylab= "Скорость выведения тетрагидрокортизона, мг/сут",
main= "Диаграмма рассеяния исследуемых на корреляцию величин",
col = "blue", pch = 21, cex = 1.2, bg = "blue") # Строим диаграмму
рассеяния данных
# На рисунке 3 представлена диаграмма рассеяния скорости выведения
# тетрагидрокортизона в зависимости от скорости выведения
прегнанетриола
```

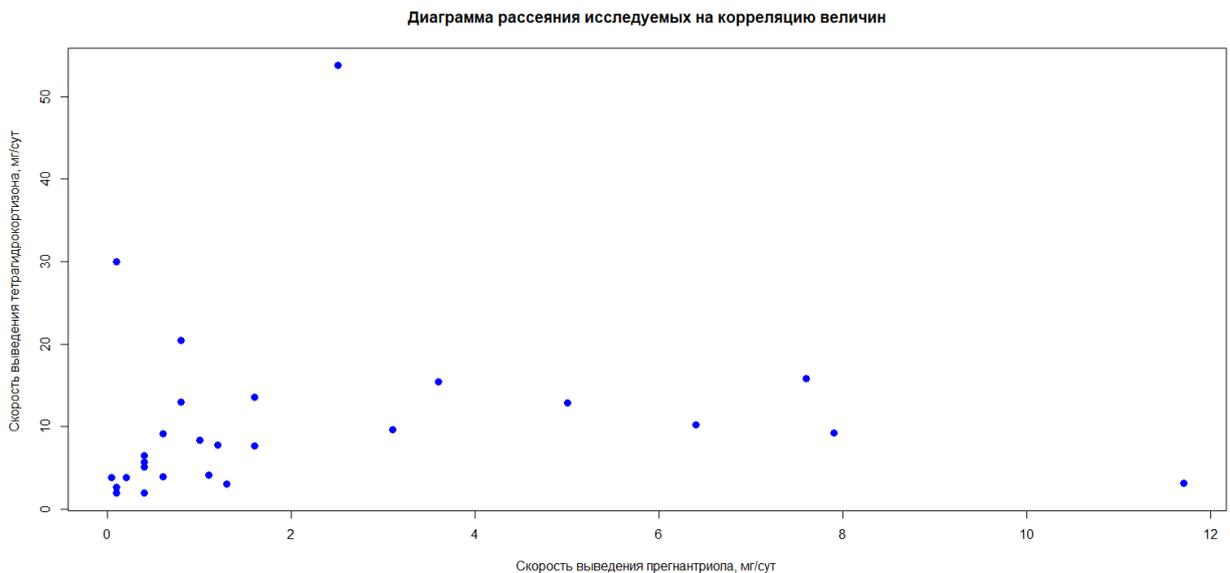


Рисунок 3 – Диаграмма рассеяния скорости выведения тетрагидрокортизона в зависимости от скорости выведения прегнанетриола

```
resCoreS<-cor.test(dataPregna, dataTetra, method= "spearman")
resCoreS#Результаты проведения корреляционного анализа по Спирмену
скорости
# выведения тетрагидрокортизона в зависимости от скорости
выведения
# прегнанетриола
Spearman's rank correlation rho
data: dataTetra and dataPenta
S = 1820.3, p-value = 0.02022
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.4443606
```

Продолжение листинга 2

```
coefDeterm <- resCoreS$estimate^2 # Вычисляем коэффициент  
детерминации  
print(coefDeterm) # Выводим на печать коэффициент детерминации  
0.1974563
```

Результаты проведенного анализа показывают, что скорость выведения тетрагидрокортизона и скорость выведения прегнанетриола у пациентов с подтвержденным диагнозом синдрома Кушинга не подчиняются нормальному закону распределения и обладают слабой статистической значимой корреляционной связью, проявляющейся в 20 % случаев.

1.1.3. Корреляция по Кендаллу

Коэффициент корреляции по Кендаллу относится к числу непараметрических критериев и не зависит от того, какому типу распределения подчиняются данные. Предназначен для работы с данными, полученными по ранговой шкале, в этом отношении он сходен с коэффициентом корреляции по Спирмену. Коэффициент корреляции по Кендаллу вычисляется как (7):

$$\tau = \frac{2 * (P - Q)}{N(N - 1)}, \quad (7)$$

где P – число совпадений; Q – число инверсий; N – число ранжируемых признаков (черт).

На языке программирования R коэффициент корреляции по Кендаллу вычисляется с помощью команды `cor(x, y, method="kendall")`, проверка статистической значимости проводится командой `cor.test(x, y, method="kendall")`.

Проведем вычисления коэффициента корреляции по Кендаллу для метрик диагностической точности 100 врачей, оценим уровень статистической значимости и коэффициент детерминации. Коэффициент корреляции составил 0,1412835, уровень статистической значимости $p = 0,03737$, а коэффициент детерминации – 2%. Вывод из вычисления коэффициента корреляции по Кендаллу: «Между метриками диагностической точности и скоростью прохождения врачом теста существует очень слабая линейная связь, проявляющаяся в 2 % случаев».

Пример проведения корреляционного анализа по Кендаллу

Основным различием между оценками корреляции по Кендаллу и по Спирмену является учет при анализе по Кендаллу наличия совпадающих значений в данных. Алгоритм проведения анализа остается таким же, как и при проведении анализа корреляции по Спирмену. В данном примере рассматривается исследование наличия корреляции между концентрацией раствора хлорида кальция и изменением длины полосы сокращения сердечной мышцы. Используется набор данных `muscle` из пакета `MASS`. Проверка данных на принадлежность к нормальному закону распределения будет выполняться с помощью теста Колмогорова–Смирнова.

Листинг 3

```
library("MASS") # Подключаем пакет, содержащий набор данных muscle
dataCalcium <- muscle$Conc # Создаем вектор, содержащий данные по
# концентрации хлорида кальция
dataLength <- muscle$Length # Вектор, содержащий данные об изменении длины
# полосы сокращения сердечной мышцы
# Строим гистограмму распределения концентрации хлорида кальция
hist(dataCalcium, breaks=12,
xlab= "Концентрация хлорида кальция",
ylab= "Количество значений",
main = "Гистограмма распределения концентрации хлорида кальция",
col = "blue")
# На рисунке 4 представлена гистограмма распределения концентрации
хлорида кальция при проведении исследований, содержащихся в наборе данных
muscle.
```

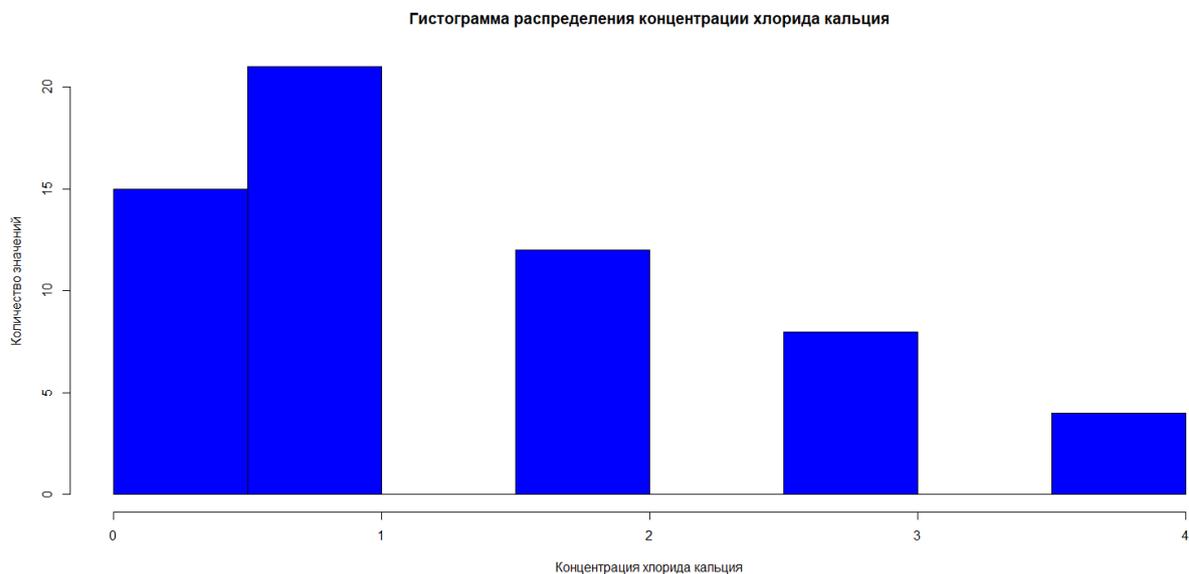


Рисунок 4 – Гистограмма распределения концентрации хлорида кальция в наборе данных *muscle* пакета *MASS*

```
ks.test(dataCalcium, "pnorm") # Проводим тест Колмогорова–Смирнова
на выяснение принадлежности данных к нормальному закону распределения
# Результаты применения теста Колмогорова-Смирнова к данным по
# концентрации хлорида кальция
```

Asymptotic one-sample Kolmogorov-Smirnov test

```
data: dataCalcium
D = 0.59871, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
# Построение гистограммы распределения изменения длины полосы
# сокращения сердечной мышцы
hist(dataLength, breaks=12,
xlab= "Длина полосы сокращения сердечной мышцы",
ylab= "Количество значений",
```

Продолжение листинга 3

```
main = "Гистограмма распределения длины полосы сокращения  
сердечной мышцы",  
col = "blue")
```

```
# На рисунке 5 представлена гистограмма распределения длины полосы  
# сокращения сердечной мышцы при проведении исследований,  
# содержащихся в наборе данных muscle
```

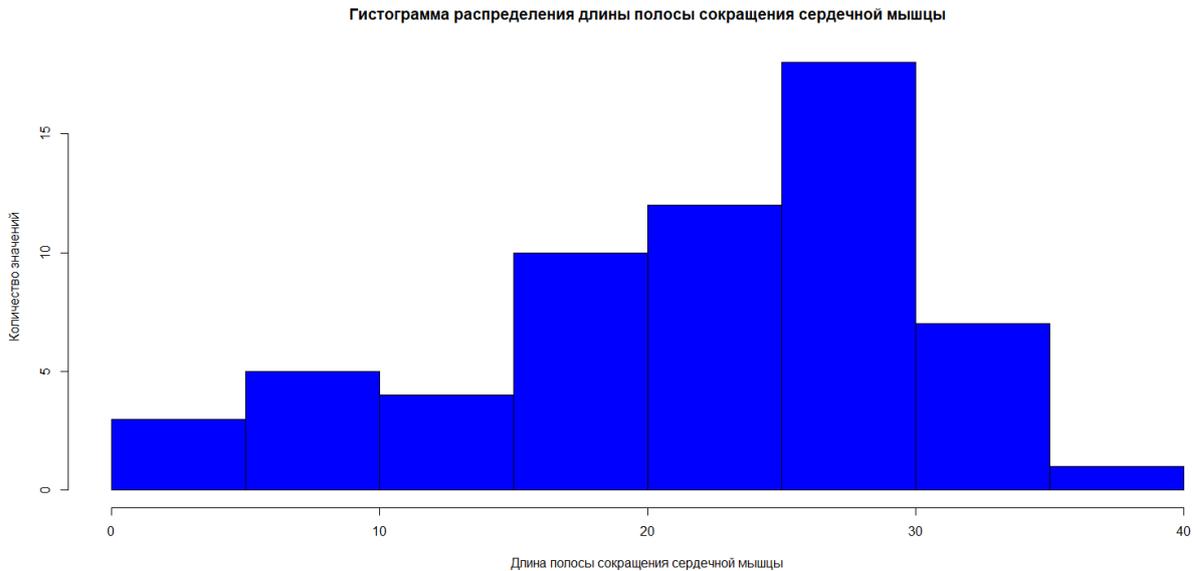


Рисунок 5 – Гистограмма распределения длины полосы сокращения сердечной мышцы в наборе данных *muscle* пакета *MASS*

```
ks.test(dataLength, "pnorm") # Проводим тест Колмогорова- Смирнова  
# на выяснение принадлежности данных к нормальному  
закону распределения  
# Результаты применения теста Колмогорова–Смирнова к данным по  
# длине полосы сокращения сердечной мышцы
```

Asymptotic one-sample kolmogorov-smirnov test

```
data: dataLength  
D = 0.97725, p-value < 2.2e-16  
alternative hypothesis: two-sided  
#=====
```

=====

```
# По результатам теста Колмогорова–Смирнова данные распределены  
# отличным от нормального закона распределения, поэтому для оценки  
корреляции  
# применяем корреляцию по Кендаллу  
#=====
```

=====

```
plot(dataCalcium, dataLength,  
xlab= "Концентрация хлорида кальция",  
ylab= "Длина полосы сокращения сердечной мышцы",  
main= "Диаграмма рассеяния исследуемых на корреляцию величин",
```

Продолжение листинга 3

```
col = "blue", pch = 21, cex = 1.2, bg = "blue") # Строим диаграмму
рассеяния данных
# На рисунке 6 представлена диаграмма рассеяния длины полосы
сокращения
# сердечной мышцы в зависимости от концентрации хлорида кальция
```

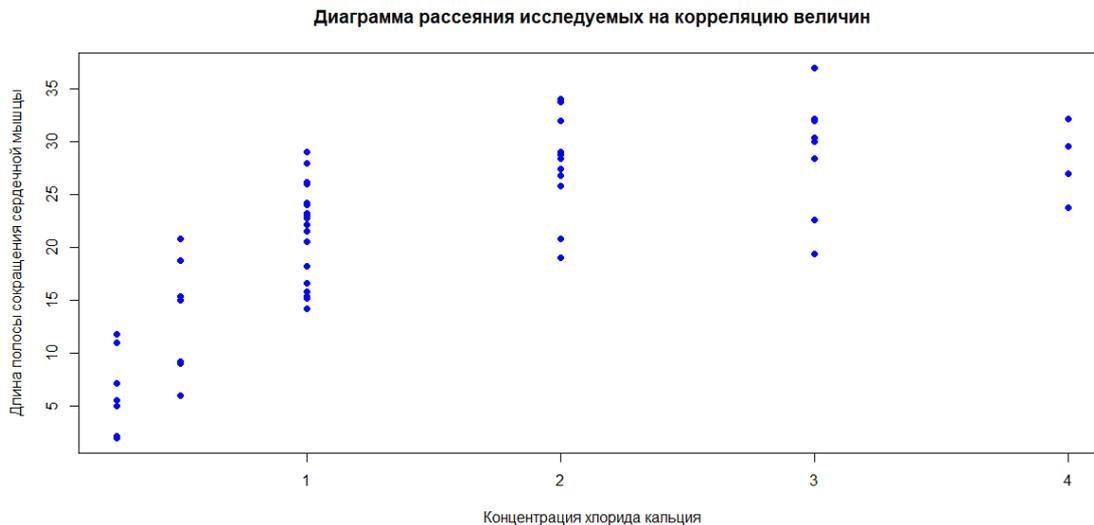


Рисунок 6 – Диаграмма рассеяния длины полосы сокращения сердечной мышцы в зависимости от концентрации хлорида кальция в наборе данных *muscle* пакета *MASS*

```
resCor<-cor.test(dataCalcium, dataLength, method = "kendall")
print(resCor) # Выводим на печать результаты корреляции по Кендаллу

kendall's rank correlation tau

data: dataCalcium and dataLength
z = 6.5831, p-value = 4.609e-11
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6354742
coefDeterm <- resCor$estimate^2 # Вычисляем коэффициент
детерминации
print(coefDeterm) # Выводим на печать коэффициент детерминации
> 40.38275
```

Результаты проведенного анализа показывают, что значения длины сокращения сердечной мышцы и концентрации хлорида кальция распределены отличным от нормального закона. Проведенный корреляционный анализ по Кендаллу показывает наличие сильной взаимосвязи между анализируемыми признаками, проявляющейся в 40 % случаев (коэффициент детерминации равен 40 %). Предварительный анализ диаграммы рассеяния показывает, что зависимость длины сокращения сердечной мышцы от концентрации хлорида кальция имеет нелинейный характер (данная часть анализа крайне важна для дальнейшего построения регрессионной модели).

1.1.4. Шкалы «силы» корреляции

Оценка «силы» корреляции между исследуемыми величинами может быть классифицирована по разным шкалам. Для медицинских приложений рекомендуют использовать «шкалу Чеддока»³ для исследования силы корреляции по Пирсону. В таблице 1 представлена шкала ранжирования «силы» корреляции Пирсона по «шкале Чеддока».

Таблица 1 – Шкала интерпретации силы корреляции по Пирсону («шкала Чеддока»)

| Значение коэффициента корреляции | Интерпретация |
|----------------------------------|-------------------------------|
| от 0 до 0,3 | Очень слабая |
| от 0,3 до 0,5 | Слабая |
| от 0,5 до 0,7 | Средняя |
| от 0,7 до 0,9 | Сильная (высокая) |
| от 0,9 до 1 | Очень сильная (очень высокая) |

Для классификации «силы» корреляции между двумя независимыми величинами могут использоваться другие шкалы: «шкала Коэна», «шкала Хинкле», «шкала Эванса» или псевдо-шкала «IBM SPSS Statistica». Данные шкалы используются в основном для проведения психосоциальных исследований⁴.

Для корреляции по Спирмену отмечается применимость «шкалы Эванса», представленная в таблице 2.

Таблица 2 – Шкала интерпретации силы корреляции по Спирмену и Пирсону («шкала Эванса»)

| Значение коэффициента корреляции | Интерпретация |
|----------------------------------|---------------|
| от 0 до 0,19 | Очень слабая |
| от 0,2 до 0,39 | Слабая |
| от 0,4 до 0,59 | Умеренная |
| от 0,6 до 0,79 | Сильная |
| от 0,8 до 1 | Очень сильная |

Авторам не удалось обнаружить работы по выбору шкалы для интерпретации результатов исследований корреляции по Кендаллу, однако, с учетом рангового характера коэффициента корреляции по Кендаллу и его «близости» к коэффициенту корреляции по Спирмену, вполне допустимо применение «шкалы Эванса» для интерпретации силы корреляции по Кендаллу.

Обобщим результаты исследований корреляции по Пирсону, Спирмену и Кендаллу для метрик диагностической точности 100 врачей и времени прохождения теста каждым врачом. Результаты представлены в таблице 3.

³ Баврина А. П., Борисов И. Б. Современные правила применения корреляционного анализа // Медицинский альманах. 2021. №3(68).

⁴ Котеров А. Н., Ушенкова Л. Н., Зубенкова Э. С. [и др.]. Сила связи. Сообщение 2. Градации величины корреляции // Медицинская радиология и радиационная безопасность. 2019. Т. 64. № 6.

Таблица 3 – Результаты корреляционного анализа по Пирсону, Спирмену и Кендаллу метрик диагностической точности и времени прохождения теста

| Тип корреляции | Коэффициент корреляции | p-value | Коэффициент детерминации, % | Интерпретация по «шкале Эванса» |
|----------------|------------------------|---------|-----------------------------|---------------------------------|
| по Пирсону | 0,27 | 0,00618 | 7,4 | Слабая |
| по Спирмену | 0,20 | 0,04948 | 3,9 | Слабая |
| по Кендаллу | 0,14 | 0,03737 | 2 | Очень слабая |

Как было показано ранее, исследуемые данные не подчиняются нормальному закону распределения и применение корреляционного анализа по Пирсону без нормализации данных является некорректным. При проведении вычислений коэффициента корреляции по Спирмену наблюдалось возникновение совпадающих рангов значений, что также некорректно для данного критерия. Корректные вычисления получены при проведении исследований корреляции по Кендаллу, где вывод является математически обоснованным и статистически значимым.

1.1.5. Замечания к результатам корреляционного анализа

При проведении корреляционного анализа больших объемов медицинских данных, содержащих различные факторы, как количественные, так и качественные, могут наблюдаться статистически значимые связи между переменными, ранее не выявленные в клинической практике. При возникновении подобных ситуаций такие закономерности требуют дальнейшей клинической проверки и не могут служить прямой «рекомендацией к действию», даже если сила корреляции и коэффициент детерминации имеют высокие значения.

Пример проведения корреляционного анализа

Проведение корреляционного анализа осуществляется в соответствии с алгоритмом, представленным на рисунке 2 части 1 методических рекомендаций. Первым шагом необходимо проверить количественные данные на принадлежность к нормальному закону распределения. Вторым шагом выбирается метод корреляционного анализа. Третьим шагом подтверждается или опровергается статистическая значимость наблюдаемой корреляции.

Для демонстрации алгоритма проведения анализа используется набор данных *Cushings*, содержащий наблюдаемые показатели экскреции с мочой двух стероидных метаболитов у пациентов с подтвержденным диагнозом «синдром Кушинга» (гипертензивное заболевание, связанное с избыточной секрецией кортизола надпочечниками).

Листинг 4

```
library("MASS") # Подключаем пакет, содержащий набор данных
Cushings
library("gofstest") # Подключаем пакет с тестом Крамера-фон Мизеса
tetraData <- Cushings$Tetrahydrocortisone # Создаем вектор, содержащий
# концентрацию тетрагидрокортизона
pregData <- Cushings$Pregnanetriol # Создаем вектор, содержащий
# концентрацию прегнантриола
```

Продолжение листинга 4

```
# Строим гистограмму распределения тетрагидрокортизона
hist(tetraData, breaks = 12,
     xlab = "Концентрация тетрагидрокортизона, мг/24 ч",
     ylab = "Частота встречаемости значений",
     main = "Гистограмма распределения тетрагидрокортизона", col = "blue")
# На рисунке 7 представлена гистограмма распределения тетрагидрокортизона
у пациентов с подтвержденным диагнозом "синдром Кушинга".
```

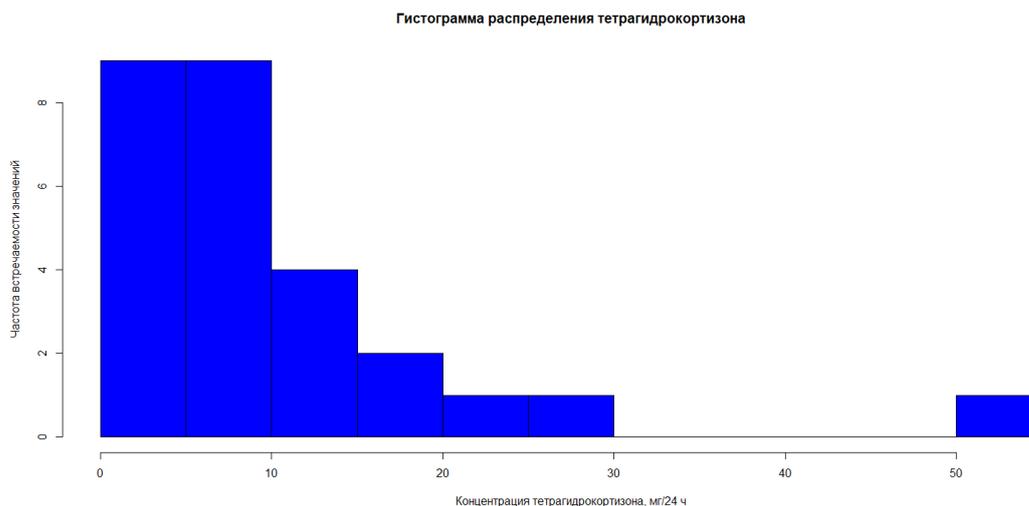


Рисунок 7 – Гистограмма распределения тетрагидрокортизона у пациентов с подтвержденным диагнозом «синдром Кушинга»

```
# Строим гистограмму распределения прегнантриола
hist(pregData, breaks = 12,
     xlab = "Концентрация прегнантриола, мг/24 ч",
     ylab = "Частота встречаемости значений",
     main = "Гистограмма распределения прегнантриола", col = "blue")
# На рисунке 8 представлена гистограмма распределения прегнантриола у
пациентов с подтвержденным диагнозом "синдром Кушинга".
```

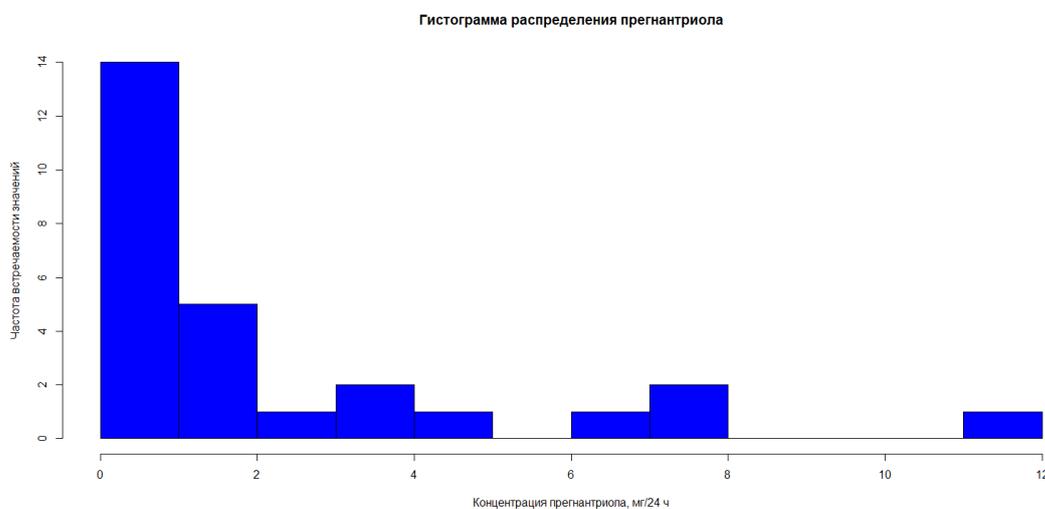


Рисунок 8 – Гистограмма распределения прегнантриола у пациентов с подтвержденным диагнозом «синдром Кушинга»

Продолжение листинга 4

```
#=====
=  
#Проверяем данные на принадлежность к нормальному закону  
распределения  
# с помощью критерия Крамера–фон Мизеса  
#=====
=  
svm.test(tetraData)  
#=====
=  
# Результаты применения теста Крамера–фон Мизеса к данным  
о концентрации тетрагидрокортизона  
#=====
=  
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  
#=====
=  
svm.test(pregData) # Проверяем данные о концентрации  
тетрагидрокортизона на нормальность по критерию  
# Крамера–фон Мизеса  
#=====
=  
# Результаты применения теста Крамера–фон Мизеса  
#=====
=  
Cramer-von Mises test of goodness-of-fit  
Null hypothesis: uniform distribution  
Parameters assumed to be fixed  
data: pregData  
omega2 = 1.7064, p-value = 3.609e-05  
#=====
=====  
# Построение диаграммы рассеяния исследуемых на корреляцию  
# величин  
#=====
=====  
plot(pregData, tetraData,  
xlab= "Скорость выведения прегнантриола, мг/сут",  
ylab= "Скорость выведения тетрагидрокортизона, мг/сут",  
main= "Диаграмма рассеяния исследуемых на корреляцию величин",  
col = "blue", pch = 21, cex = 1.2, bg = "blue") # Строим диаграмму рассеяния  
данных  
# На рисунке 9 представлена диаграмма рассеяния скорости выведения  
# тетрагидрокортизона в зависимости от скорости выведения  
прегнантриола
```

Продолжение листинга 4

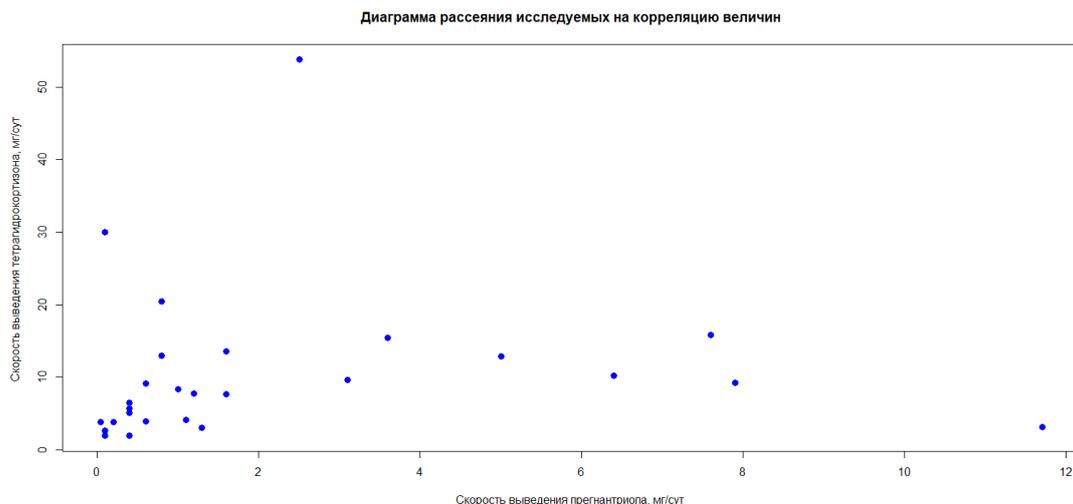


Рисунок 9 – Диаграмма рассеяния зависимости скорости выведения тетрагидрокортизона от скорости выведения pregnanтриола

```
corPirson <- cor.test(pregData, tetraData, method = "pearson") #  
Рассматриваем  
# корреляцию по Пирсону  
print(corPirson) # Выводим результаты в консоль  
#=====
```

```
=====
```

```
# Результаты корреляционного анализа по Пирсону  
#=====
```

```
=====
```

```
Pearson's product-moment correlation  
data: pregData and tetraData  
t = 0.31517, df = 25, p-value = 0.7552  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.3248706 0.4325823  
sample estimates:  
cor  
0.0629099  
#=====
```

```
=====
```

```
corSperman <- cor.test(pregData, tetraData, method = "spearman") #  
Рассматриваем  
# корреляцию по Спирмену  
print(corSperman)  
#=====
```

```
=====
```

```
# Результаты корреляционного анализа по Спирмену  
#=====
```

```
=====
```

```
Spearman's rank correlation rho  
data: pregData and tetraData  
S = 1820.3, p-value = 0.02022
```

Продолжение листинга 4

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4443606
#=====
=====
      corKendall <- cor.test(pregData, tetraData, method = "kendall") #
Рассматриваем
      # корреляцию по Кендаллу
      print(corKendall)
#=====
=====
      # Результаты корреляционного анализа по Кендаллу
#=====
=====
kendall's rank correlation tau

data: pregData and tetraData
z = 2.4064, p-value = 0.01611
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3343376
#=====
=====
```

Результаты корреляционного анализа по Пирсону показывают отсутствие статистически значимой корреляции между исследуемыми данными, тогда как анализ по Спирмену и Кендаллу демонстрируют наличие таковой. Коэффициент корреляции по Кендаллу ниже, чем по Спирмену, что свидетельствует о наличии совпадающих значений в наборе данных, которые завышают силу корреляции.

2. СРАВНЕНИЕ ДАННЫХ В ГРУППАХ

Данные по результатам тестирования 100 врачей содержат не только метрики диагностической точности и время прохождения теста каждым врачом, но и опыт работы каждого врача, и основную специализацию. Необходимо ответить на вопросы: «Влияет ли опыт работы врача на метрики диагностической точности?» и «Влияет ли основная специализация врача на метрики диагностической точности?».

Сделать это можно, проведя сравнение групп между собой и определив, статистически значимо отличаются группы друг от друга или нет. Проведение сравнения групп осуществляется посредством тестов ANOVA (Analysis of Variance – дисперсионный анализ), t-теста Стьюдента, F-критерия Фишера, непараметрических критериев Колмогорова–Смирнова, Манна–Уитни и Краскелла–Уоллиса.

2.1. Дисперсионный анализ ANOVA

Основной смысл дисперсионного анализа заключается в разделении общей дисперсии изучаемого признака на отдельные части. Каждая часть признака выделяется за счет влияния конкретных факторов и проверки гипотез о значимости влияния этих факторов на исследуемый признак [1]. Дисперсионный анализ может быть:

- однофакторным;
- двухфакторным;
- многофакторным.

Основная теорема дисперсионного анализа гласит, что общая дисперсия всей совокупности равна сумме внутригрупповой и межгрупповой дисперсий и может быть представлена в виде уравнения (8):

$$\frac{\sum_{j=1}^p \sum_{i=1}^n (x_{ji} - \bar{x})^2}{N - 1} = \frac{\sum_{j=1}^p \sum_{i=1}^n (x_{ji} - \bar{x}_j)^2}{N - p} + \frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{p - 1}, \quad (8)$$

где N – общее число всех элементов в анализируемой совокупности экспериментальных данных; p – число сравниваемых экспериментальных выборок; \bar{x} – общая средняя всей анализируемой совокупности данных; \bar{x}_j – средняя j -й выборки; x_{ji} – значение всех переменных, полученных в эксперименте, при этом индекс j меняется от 1 до p .

Необходимо отметить, что в основе дисперсионного анализа лежит приближение (допущение) о том, что исследуемые данные подчиняются нормальному (Гауссову) закону распределения. Исходя из этого, прежде чем применить дисперсионный анализ, необходимо провести проверку принадлежности данных к нормальному закону распределения и действовать в соответствии с алгоритмом, представленным на рисунке 2 первой части настоящих методических рекомендаций.

На языке R для проведения одно-, двух- и многофакторного анализа существует команда `aov(formula=x~y,...)`. Нулевой гипотезой в случае дисперсионного анализа является предположение, что все средние значения равны,

а альтернативной гипотезой – что не равны. В исследуемых данных все испытуемые врачи были разделены на 4 группы в зависимости от опыта работы:

- 1) от 0 до 1 года;
- 2) от 1 до 5 лет;
- 3) от 5 до 10 лет;
- 4) более 10 лет.

Результаты проведения дисперсионного анализа показывают значение $F = 3,056^5$ и $p = 0,0321$. Следовательно, у нас есть достаточно доказательств, чтобы отвергнуть нулевую гипотезу о равенстве средних значений метрик диагностической точности в группах, разделенных по признаку опыта работы. Применение теста Тьюки⁶ позволяет сравнить все возможные группы между собой. На языке R данный тест выполняется посредством команды *TukeyHSD(Model)*, где *Model* – результаты, получаемые при выполнении команды *aov()*. Результаты выполнения теста представлены в таблице 4.

Таблица 4 – Результаты сравнения групп из 100 врачей, разделенных по опыту работы

| Сравниваемые пары | Уровень статистической значимости |
|-------------------|-----------------------------------|
| 1-5 – 0-1 | 0,0687048 |
| 10+ – 0-1 | 0,0407361 |
| 5-10 – 0-1 | 0,0281046 |
| 10+ – 1-5 | 0,8500527 |
| 5-10 – 1-5 | 0,5313565 |
| 5-10 – 10+ | 0,9226527 |

Из таблицы 4 следует, что статистически значимые различия наблюдаются между группами с опытом работы от 0 до 1 года и группами с опытом работы свыше 10 лет и от 5 до 10 лет. В остальных случаях полученные различия в группах статистически не значимы.

Проведем такой же анализ для специализации врачей, участвовавших в эксперименте. Используются следующие специализации:

- 1) компьютерная томография;
- 2) маммография;
- 3) магнитно-резонансная томография;
- 4) рентгенография и флюорография.

Проведение дисперсионного анализа показывает значение F-статистики, равное 5,725, а значение $p\text{-value} = 0,0012$, что позволяет отвергнуть нулевую гипотезу о равенстве средних значений в выборках по специализациям.

Так же, как и в случае с опытом работы, проведем попарное сравнение тестом Тьюки по специализациям. Результаты сравнения представлены в таблице 5.

⁵ Значение F-статистики показывает различия в вариациях, т.е. во сколько раз вариация между выборочными средними больше, чем вариации внутри выборок.

⁶ Необходимо отметить, что тест Тьюки не требует поправки на множественное сравнение (более подробно о множественном сравнении рассказано в подразделе 2.7).

Таблица 5 – Результаты сравнения групп из 100 врачей, разделенных по специализации

| Сравниваемые пары | Уровень статистической значимости |
|-------------------|-----------------------------------|
| ММГ – КТ | 0,7985779 |
| МРТ – КТ | 0,9860026 |
| РГ, ФЛГ – КТ | 0,0335126 |
| МРТ – ММГ | 0,6218859 |
| РГ, ФЛГ – ММГ | 0,1225082 |
| РГ, ФЛГ – МРТ | 0,0258812 |

Примечание. ММГ – маммография, КТ – компьютерная томография, МРТ – магнитно-резонансная томография, РГ – рентгенография, ФЛГ – флюорография

По результатам проведенного анализа статистически значимые различия наблюдаются в группах врачей со специализациями «рентгенография», «флюорография» и «магнитно-резонансная томография», аналогичная значимость наблюдается в случае с врачами, специализирующимися на компьютерной томографии.

Дисперсионный анализ в сочетании с тестом Тьюки позволяет установить статистически значимые различия в группе данных, разделенных по факторам. Аналогичного результата можно добиться, применяя t-критерий Стьюдента с учетом поправки на множественное сравнение⁷.

Пример применения дисперсионного анализа

Для примера выявления различий в группах используем набор данных **Melanoma** из пакета **MASS** и проведем исследование разницы в размерах опухоли и в ее наличии в зависимости от гендерной принадлежности пациента и его возраста. Для этих целей используем дисперсионный анализ и тест Тьюки.

Листинг 5

```
library("MASS") # Подключаем пакет, содержащий набор данных
Melanoma
# Рассматриваем различия в диагнозе Меланома в зависимости от
возраста
dispModel_UncAge <- aov(formula = ulcer~as.factor(age), data=Melanoma)
summary(dispModel_UncAge) # Обобщаем результаты дисперсионного
анализа.
#=====
=====
# Обобщенные результаты однофакторного дисперсионного анализа
#=====
=====
              Df Sum Sq Mean Sq F value Pr(>F)
as.factor(age) 69  17.58   0.2547   1.045  0.408
Residuals    135  32.91   0.2438
```

⁷ См. подразделы 2.2 и 2.7.

Продолжение листинга 5

#Рассматриваем различия в размерах патологии в зависимости от возраста

```
dispModel_ThinkAge <- aov(formula = thickness~as.factor(age), data=Melanoma)
```

summary(dispModel_ThinkAge) #Обобщаем результаты дисперсионного анализа.

```
#=====
=====
```

Обобщенные результаты однофакторного дисперсионного анализа

```
#=====
=====
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|-----|--------|---------|---------|--------|
| as.factor(age) | 69 | 608.3 | 8.817 | 1.01 | 0.472 |
| Residuals | 135 | 1178.3 | 8.728 | | |

#Рассматриваем различия в наличии патологии в зависимости от пола

```
dispModel_UncSex <- aov(formula = ulcer~as.factor(sex), data=Melanoma)
```

summary(dispModel_UncSex) # Обобщаем результаты дисперсионного анализа

```
#=====
=====
```

Обобщенные результаты однофакторного дисперсионного анализа

```
#=====
=====
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|-----|--------|---------|---------|----------|
| as.factor(sex) | 1 | 1.42 | 1.4246 | 5.894 | 0.0161 * |
| Residuals | 203 | 49.06 | 0.2417 | | |

#Рассматриваем различия в размерах патологии в зависимости от пола

```
dispModel_ThinkSex <- aov(formula = thickness~as.factor(sex), data=Melanoma)
```

summary(dispModel_ThinkSex) # Обобщаем результаты дисперсионного анализа

```
#=====
=====
```

Обобщенные результаты однофакторного дисперсионного анализа

```
#=====
=====
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|----------------|-----|--------|---------|---------|------------|
| as.factor(sex) | 1 | 61.4 | 61.42 | 7.227 | 0.00778 ** |
| Residuals | 203 | 1725.3 | 8.50 | | |

Интерпретация результатов дисперсионного анализа основана на анализе Pr(>F) (уровня статистической значимости). Если Pr(>F) < 0,05, то у нас нет достаточных оснований для того, чтобы принять нулевую гипотезу (все средние группы равны), в противном случае все группы равны. Результаты дисперсионного анализа показывают, что статистически значимые различия среднего размера опухоли наблюдаются по гендерному признаку и не наблюдаются по возрастному признаку, как в наличии патологии, так и в размерах патологии.

2.1.1. Частный случай дисперсионного анализа. Метод MRMC⁸

Показатели эффективности работы врачей определяются по метрикам диагностической точности, вычисляемым как площадь под ожидаемой кривой полезности ROC⁹, отношения правдоподобия положительных и отрицательных тестов, а также чувствительности и специфичности¹⁰. ROC-кривые могут быть оценены эмпирически, с помощью бинормальных моделей или бинормальных отношений правдоподобия.

Статистическое сравнение диагностических тестов основано на модели ANOVA Обуховского – Рокетта и унифицированной схеме Хиллиса¹¹. Анализ ANOVA может проводиться с данными, полученными в ходе полного факторного¹², гнездового¹³ или частично парного исследования¹⁴, со случайными или фиксированными врачами или исследованиями¹⁵.

На языке программирования R проведение оценок метрик диагностической точности нескольких врачей (экспертов), выполнявших исследование пациентов с имеющимися подтвержденными диагнозами, осуществляется с помощью функции *mrnc()*, входящей в пакет **MRMCaov**.

Пример использования метода MRMC

Для примера использования метода оценки диагностической точности врачей, проводивших анализ нескольких исследований, в пакете **MRMCaov** содержится набор данных *VanDyke* по исследованиям эффективности выявления расслоения грудной аорты врачами при просмотре кинематографического представления МРТ (CINE) и магнитно-резонансной томографии с одним спин-эхо (SE). В таблице 6 представлена структура набора данных *VanDyke*.

⁸ Multi-Reader Multi-Case – множественное прочтение множественных заданий.

⁹ Оценка зрелости технологий искусственного интеллекта для здравоохранения: методические рекомендации / сост. Ю. А. Васильев, А. В. Владимировский, О. В. Омелянская [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». Вып. 123. М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2023. 28 с.

¹⁰ Более подробно о понятиях «чувствительность», «специфичность», «точность» и «ROC-кривые» написано в разделе 5.

¹¹ Hillis S. L., Obuchowski N. A., Schartz K. M., et al. A comparison of the Dorfman – Berbaum – Metz and Obuchowski – Rockette methods for receiver operating characteristic (ROC) data // *Statistics in medicine*. 2005. Vol. 24, №10. P. 1579–1607.

¹² В общем случае исследование, в котором реализуются все возможные сочетания уровней факторов, называется «полным факторным экспериментом».

¹³ Синоним кластерного исследования – изучаемые данные разбиваются на кластеры (гнезда).

¹⁴ Исследуется влияние парного взаимодействия между факторами.

¹⁵ Smith B. J., Hillis S. L. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities // *Medical Imaging. 2020: Image Perception, Observer Performance, and Technology Assessment*. Vol. 11316. SPIE, 2020.

Таблица 6 – Наименование переменных и тип данных, содержащихся в *VanDyke*

| Наименование переменной | Тип данных |
|-------------------------|--|
| reader | Уникальные идентификаторы пяти радиологов |
| treatment | Идентификатор типа снимка МРТ (1 = CINE MRI, 2 = SE MRI) |
| case | Идентификаторы для 114 случаев |
| truth | Истинный показатель расслоения грудной аорты (1 = есть расслоение, 0 = нет расслоения) |
| rating | Пятибалльные оценки, выставленные врачами изображениям кейсов |
| case2 | Примеры идентификаторов, представляющих уровни факторов врача |
| case3 | Примеры идентификаторов, представляющих уровни фактора диагностического теста |

Листинг 6

```

#Подключаем пакет, содержащий набор данных VanDyke и функцию
mrms()
library(MRMCaov)
#Подключаем библиотеку работы с большими объемами данных
library(dplyr)
#Подключаем графическую библиотеку для построения графиков
library(ggplot2)
# Проведение расчета метрик диагностической точности
mrmcResult <- mrmc(empirical_auc(truth, rating), treatment, reader, case,
                    data = VanDyke)
summary(mrmcResult) #Обобщение результатов вычислений
#=====
=====
#Вывод обобщенного результата исследования по методу MRMC
#=====
=====
Multi-Reader Multi-Case Analysis of Variance
Data: VanDyke
Factor types: Random Readers and Random Cases
Covariance method: jackknife

Experimental design: factorial

Obuchowski-Rockette variance component and covariance estimates:

              Estimate Correlation
reader          0.0015349993         NA
treatment:reader 0.0002004025         NA
Error           0.0008022883         NA
Cov1            0.0003466137    0.4320314
Cov2            0.0003440748    0.4288668
Cov3            0.0002390284    0.2979333

ANOVA global test of equal treatment empirical_auc:
      MS(T)      MS(T:R)      Cov2      Cov3 Denominator
F df1  df2  p-value
4.456319  1 0.004796171 0.0005510306 0.0003440748 0.0002390284 0.001076263

```

Продолжение листинга 6

```

95% Cis and tests for treatment empirical_auc pairwise
differences:
  Comparison      Estimate      StdErr      df      CI.Lower
CI.Upper      t      p-value
2 1 - 2 -0.04380032  0.02074862  15.25967  -0.0879594986
0.0003588544 -2.110999 0.05166569

```

```

95% treatment empirical_auc Cis (each analysis based only on data
for the specified treatment):
  Estimate      MS®      Cov2      StdErr      df      CI.Lower
CI.Upper
1 0.8970370  0.003082629  0.0004839618  0.03317360  12.74465
0.8252236 0.9688505
2 0.9408374  0.001304602  0.0002041879  0.02156637  12.71019
0.8941378 0.9875369

```

```

#Графическое отображение результатов вычислений
#Вычисляем эмпирическим методом параметры ROC-кривой
plotData <- with(VanDyke,{
  roc_curves(truth, rating, groups = list(`Модальность` = treatment, `Врач` =
reader),
            method= «empirical»)
})
# Преобразуем результаты вычисления в вид, позволяющий получить
доступ к данным
paramCurv <- parameters(plotData)
# Фильтруем данные для построения графиков
paramCurv %>%
  filter(Group$`Модальность`=="2",Group$`Врач`=="1")->Readers
#Строим график для второй модальности и первого врача
plotRes<-ggplot()
plotRes<-plotRes+geom_smooth(mapping
aes(Readers$FPR[[1]],Readers$TPR[[1]], colour= "blue", size=2, se=FALSE)
plotRes <- plotRes + theme_light()
plotRes <- plotRes + theme(panel.border = element_rect(size = 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(size = 3, colour = "black"),
legend.title = element_text(size = 20, face = "bold", color =
"black"),
legend.text = element_text(size = 20, face = "bold", color =
"black"))
plotRes <- plotRes + labs(x='Ложноположительные значения',
y= 'Истинно положительные значения')
print(plotRes)
#=====
=====
# График метрик диагностической точности первого врача

```

Продолжение листинга 6

#=====

=====
На рисунке 10 изображена ROC-кривая первого врача, который просматривал изображение расслоения грудной аорты по данным МРТ.

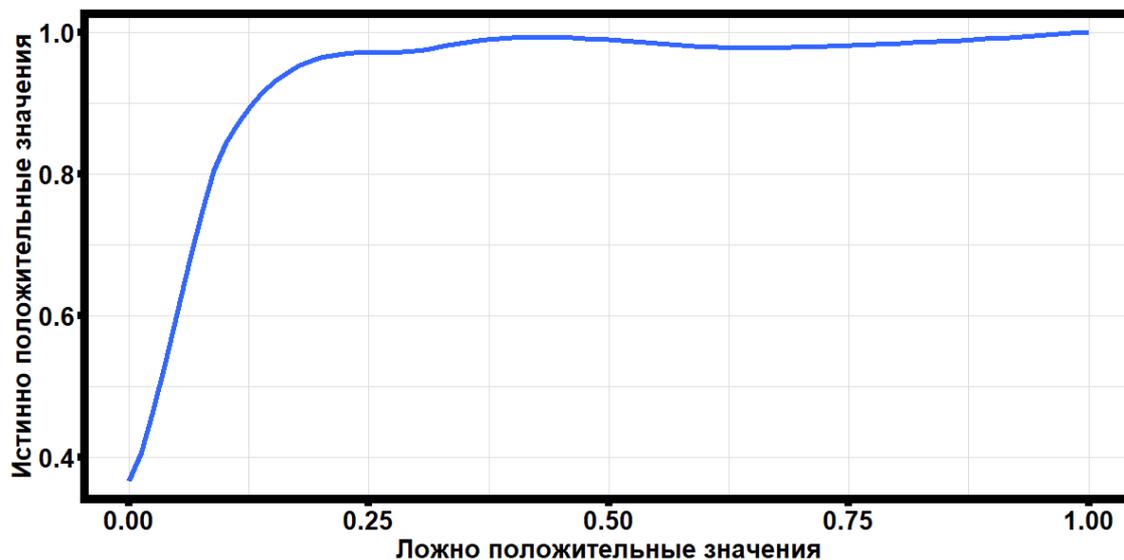


Рисунок 10 – ROC-кривая первого врача, просматривавшего магнитно-резонансную томографию с одним спин-эхо

Обобщенный анализ результатов исследования несколькими врачами некоторого количества исследований, проведенных посредством разных видов аппаратуры, показывает, что изменение метрик диагностической точности по расслоению грудной аорты в зависимости от метода МРТ является статистически не значимым, т. к. $p\text{-value} = 0,05166569 > 0,05$ по методу MRMC¹⁶.

2.2. t-критерий Стьюдента

Критерий Стьюдента направлен на установление различий в средних значениях двух выборок, которые распределены по нормальному закону. Данный критерий имеет широкое применение, а именно, позволяет сравнивать фактические данные относительно ожидаемого среднего значения, проводить анализ статистической значимости различий в средних значениях двух групп, при этом критерий позволяет осуществлять сравнение выборок различного объема. В общем виде критерий Стьюдента рассчитывается по уравнению (9):

$$t = \frac{|\bar{X} - \bar{Y}|}{Sd}, \quad (9)$$

где \bar{X} , \bar{Y} – сравниваемые средние значения; Sd – несмещенная оценка среднего квадратического отклонения выборочных средних, находится по уравнению (10):

¹⁶ С дополнительными возможностями пакета MRMCaov можно познакомиться по ссылке: <https://brian-j-smith.github.io/MRMCaov/using.html>.

$$Sd = \sqrt{S_1^2 + S_2^2}, \quad (10)$$

где S_1 – среднее квадратическое отклонение в выборке 1; S_2 – среднее квадратическое отклонение в выборке 2.

Для случая равночисленных выборок (размер выборки X равен размеру выборки Y) полный вид уравнения (11):

$$\begin{aligned} Sd &= \sqrt{S_1^2 + S_2^2} \\ &= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{(n-1)n}}. \end{aligned} \quad (11)$$

В случае неравночисленных выборок уравнение (12) приобретает вид:

$$Sd = \sqrt{S_1^2 + S_2^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{(n_1 + n_2 - 2)} * \frac{(n_1 + n_2)}{n_1 n_2}}. \quad (12)$$

В случае связанных выборок критерий Стьюдента определяется по уравнению (13):

$$t = \frac{\bar{d}}{Sd}, \quad (13)$$

где

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{\sum_{i=1}^n (x_i - y_i)}{n} \quad (14)$$

$$Sd = \sqrt{\frac{\sum_{i=1}^n d_i^2 - \frac{(\sum_{i=1}^n d_i)^2}{n}}{n(n-1)}}. \quad (15)$$

В основе t -теста Стьюдента лежит предположение о принадлежности данных к нормальному (Гауссову) закону распределения, по этой причине перед применением теста Стьюдента необходимо проверить данные на принадлежность к нормальному закону распределения.

На языке программирования R применение t -теста Стьюдента осуществляется с помощью функции ***t.test()***. Отличительной особенностью данного теста является возможность сравнения только двух уровней одного фактора.

В таблице 7 представлены результаты анализа различий в группах средних значений метрик диагностической точности 100 врачей в зависимости от субспециализации.

Таблица 7 – Сравнение средних значений, по критерию Стьюдента, выборки по субспециализациям метрик диагностической точности 100 врачей

| Подвергаемая сравнению группа по субспециализациям | Уровень статистической значимости по критерию Стьюдента |
|--|---|
| КТ – ММГ | 0,4069 |
| КТ – МРТ | 0,7973 |
| КТ – РГ, ФЛГ | 0,03776 |
| ММГ – МРТ | 0,3792 |
| ММГ – РГ, ФЛГ | 0,02376 |
| МРТ – РГ, ФЛГ | 0,08836 |
| <i>Примечание. КТ – компьютерная томография, ММГ – маммография, МРТ – магнитно-резонансная томография, РГ – рентгенография, ФЛГ – флюорография</i> | |

Из результатов анализа по критерию Стьюдента метрик диагностической точности 100 врачей, разделенных по субспециализациям, статистически значимые различия наблюдаются у субспециализаций ММГ и РГ, ФЛГ, а также у КТ и РГ, ФЛГ, т.к. уровень статистической значимости меньше 0,05. В остальных случаях различия в средних значениях метрик диагностической точности статистически не значимы, т. к. $p\text{-value} > 0,05$.

В таблице 8 представлены результаты анализа различий в группах средних значений метрик диагностической точности 100 врачей в зависимости от стажа работы.

Таблица 8 – Сравнение средних значений, по критерию Стьюдента, независимых выборок по опыту работы метрик диагностической точности 100 врачей

| Сравниваемые пары | Уровень статистической значимости |
|-------------------|-----------------------------------|
| 1-5 – 0-1 | – |
| 10+ – 0-1 | – |
| 5-10 – 0-1 | – |
| 10+ – 1-5 | 0,4109 |
| 5-10 – 1-5 | 0,1755 |
| 5-10 – 10+ | 0,5399 |

Количество врачей с опытом работы от 0 до 1 года в общей выборке оказалось равно одному человеку, и применить критерий Стьюдента для сравнения метрик диагностической точности одного врача с более опытными коллегами не представляется возможным. Сравнение остальных групп между собой показало отсутствие статистической значимости в различиях средних значений, т. к. $p\text{-value} > 0,05$.

Пример применения t-теста Стьюдента

В качестве примера применения t-теста Стьюдента для независимых выборок рассмотрим изменение веса у пациентов с подтвержденным диагнозом «анорексия», находящегося в наборе данных *anorexia* из пакета MASS.

Листинг 7

```
library(MASS) #Подключаем библиотеку, содержащую набор данных
anorexia
#Рассматриваем попарно различия в весе пациентов в трех группах
# Cont – контрольная группа
# СВТ – Когнитивно-поведенческая терапия
# FT - семейное лечение
prewtContData <- anorexia[anorexia$Treat == "Cont", ]$Prewt #Выбираем
значения
#веса контрольной группы до проведения эксперимента
prewtCBTData <- anorexia[anorexia$Treat == "CBT", ]$Prewt #Выбираем
значения
#веса в группе, которой предписана когнитивно-поведенческая терапия
prewtFTData <- anorexia[anorexia$Treat == "FT", ]$Prewt # Выбираем
значения
#веса в группе, которой предписано семейное лечение
#=====
=====
#Проводим сравнительный анализ в группах до применения лечения
#Сравнение среднего веса по t-критерию Стьюдента
#=====
=====
t.test(prewtContData, prewtCBTData) #Сравнение веса в контрольной группе
и
# группе с назначенной когнитивно-поведенческой терапией
#=====
=====
#Результат сравнения
#=====
=====
welch Two Sample t-test

data: prewtContData and prewtCBTData
t = -0.78823, df = 49.352, p-value = 0.4343
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -4.017355  1.753429
sample estimates:
mean of x mean of y
 81.55769  82.68966
#=====
=====
#Сравнение веса пациентов в контрольной группе и с предписанным
семейным
#лечением
#=====
=====
t.test(prewtContData, prewtCBTData)
#=====
=====
#Результат сравнения
```

Продолжение листинга 7

```
#=====
=====
welch Two Sample t-test

data:  prewtContData and prewFTData
t = -1.0112, df = 37.397, p-value = 0.3184
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
  -5.020262  1.676823
sample estimates:
mean of x mean of y
 81.55769  83.22941
#=====

=====
#Сравнение в группах с предписанной семейной и когнитивно-
поведенческой
#терапией
#=====

=====
welch Two Sample t-test

data:  prewCBTData and prewFTData
t = -0.35668, df = 32.697, p-value = 0.7236
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
  -3.619646  2.540132
sample estimates:
mean of x mean of y
 82.68966  83.22941
#=====

=====
#Сравнение в группах после проведения терапии
#=====

=====
postwtContData <- anorexia[anorexia$Treat=="Cont", ]$Postwt #Выбираем
#значения веса в контрольной группе после лечения
postwtCBTData <- anorexia[anorexia$Treat=="CBT", ]$Postwt #Выбираем
#значения веса в группе с когнитивно-поведенческой терапией
postwtFTData <- anorexia[anorexia$Treat=="FT", ]$Postwt #Выбираем
#значения веса в группе, которой предписано семейное лечение
#=====

=====
#Сравниваем средний вес в группах после лечения
#=====

=====
t.test(postwtContData, postwtCBTData)
#=====

=====
#Результат сравнения контрольной группы с группой, прошедшей
# когнитивно-поведенческую терапию
#=====

=====
welch Two Sample t-test

data:  postwtContData and postwtCBTData
```

Продолжение листинга 7

```
t = -2.5372, df = 45.221, p-value = 0.01469
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.2310687 -0.9466501
sample estimates:
mean of x mean of y
 81.10769  85.69655
```

```
#=====
```

```
=====
```

```
t.test(postwtContData,postwtFTData)
```

```
#Результат сравнения контрольной группы с группой, прошедшей  
# семейную терапию
```

```
#=====
```

```
=====
```

welch Two Sample t-test

```
data: postwtContData and postwtFTData
t = -4.1601, df = 22.62, p-value = 0.0003888
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -14.058233 -4.714618
sample estimates:
mean of x mean of y
 81.10769  90.49412
```

```
#=====
```

```
=====
```

```
#Результат сравнения группы с когнитивно-поведенческой и  
# семейной терапией
```

```
t.test(postwtCBTData, postwtFTData)
```

```
#=====
```

```
=====
```

welch Two Sample t-test

```
data: postwtCBTData and postwtFTData
t = -1.8632, df = 33.246, p-value = 0.07129
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.0348982  0.4397663
sample estimates:
mean of x mean of y
 85.69655  90.49412
```

```
#=====
```

```
=====
```

```
#Проверка различий в средних значениях веса до и после лечения в  
#контрольной группе
```

```
t.test(prewtContData,postwtContData)
```

```
#=====
```

```
=====
```

welch Two Sample t-test

```
data: prewtContData and postwtContData
t = 0.30918, df = 48.385, p-value = 0.7585
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.475824 3.375824
sample estimates:
mean of x mean of y
```

Продолжение листинга 7

```
81.55769 81.10769
```

```
#=====
```

```
=====
```

```
#Проверка различий в весе пациентов между группами с  
#когнитивно-поведенческой до и после прохождения терапии  
t.test(rewCBTData,postwtCBTData)
```

```
#=====
```

```
=====
```

```
welch Two Sample t-test
```

```
data:  rewCBTData and postwtCBTData  
t = -1.677, df = 44.931, p-value = 0.1005  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -6.6183919 0.6045988  
sample estimates:  
mean of x mean of y  
 82.68966 85.69655
```

```
#=====
```

```
=====
```

```
#Различия в весе пациентов до и после прохождения семейной терапии  
t.test(rewFTData,postwtFTData)
```

```
#=====
```

```
=====
```

```
welch Two Sample t-test
```

```
data:  rewFTData and postwtFTData  
t = -3.0414, df = 25.986, p-value = 0.005324  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -12.17472 -2.35469  
sample estimates:  
mean of x mean of y  
 83.22941 90.49412
```

```
#=====
```

```
=====
```

Результаты применения критерия Стьюдента к исследованию изменения веса пациентов с подтвержденным диагнозом «анорексия» показывают, что статистически значимые¹⁷ изменения в весе относительно контрольной группы наблюдаются у пациентов, проходивших когнитивно-поведенческую и семейную терапию. При этом статистически значимых различий между группами, прошедшими когнитивно-поведенческую терапию, и группами, завершившими семейную терапию, не выявлено, и установлено, что пациенты, проходившие семейную терапию, статистически значимо набрали вес по сравнению с состоянием до прохождения лечения, тогда как в остальных группах таких изменений не наблюдается. Таким образом, наибольшее влияние на пациентов оказала семейная терапия.

¹⁷ Здесь необходимо уточнить, что при проведении сравнения с количеством групп больше двух требуется введение поправки на множественное сравнение. Об этом более подробно говорится в подразделе 2.7.

2.3. F-критерий Фишера

Критерий Фишера предназначен для сравнения дисперсий двух рядов наблюдений. Для вычисления F -критерия необходимо найти отношение дисперсий двух выборок, причем так, чтобы большая по величине дисперсия находилась бы в числителе, а меньшая – в знаменателе (16):

$$F = \frac{S_1^2}{S_2^2}, \quad (16)$$

где

$$S_1^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}, \quad (17)$$

$$S_2^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}, \quad (18)$$

где S_1 и S_2 – дисперсии распределений величин X и Y ; n – число наблюдений в каждой выборке; \bar{x} и \bar{y} – средние значения выборок X и Y .

На языке R тест Фишера выполняется с помощью функции *fisher.test()* (для выборок, имеющих одинаковое количество данных) и *var.test()* (для выборок, имеющих разное количество данных). В таблице 9 представлены результаты применения теста Фишера к выборкам по субспециализациям метрик диагностической точности 100 врачей.

Таблица 9 – Сопоставление дисперсий выборок по субспециализациям метрик диагностической точности 100 врачей по критерию Фишера

| Сравниваемые пары | Уровень статистической значимости |
|-------------------|-----------------------------------|
| ММГ – КТ | 1 |
| МРТ – КТ | 1 |
| РГ, ФЛГ – КТ | 0,9691 |
| МРТ – ММГ | 1 |
| РГ, ФЛГ – ММГ | 1,869e-07 |
| РГ, ФЛГ – МРТ | 0,8544 |

Примечание. ММГ – маммография, КТ – компьютерная томография, МРТ – магнитно-резонансная томография, РГ – рентгенография, ФЛГ – флюорография

При проведении множественного (когда количество групп больше двух) сравнения дисперсий по критерию Фишера необходимо учитывать поправку на множественное сравнение, более подробно об этом см. в подразделе 2.7.

Из результатов сравнения дисперсий по критерию Фишера следует, что статистически значимая разница присутствует в субспециализациях РГ, ФЛГ и ММГ и отсутствует в остальных случаях.

Проведем такой же анализ, только для выборок метрик диагностической точности в зависимости от опыта работы 100 врачей. Результат представлен в таблице 10.

Таблица 10 – Сопоставление дисперсий выборок по опыту работы метрик диагностической точности 100 врачей по критерию Фишера

| Сравниваемые пары | Уровень статистической значимости |
|-------------------|-----------------------------------|
| 1-5 – 0-1 | 1 |
| 10+ – 0-1 | 0,95 |
| 5-10 – 0-1 | 1 |
| 10+ – 1-5 | 1,658e-09 |
| 5-10 – 1-5 | 1 |
| 5-10 – 10+ | 5,93e-11 |

Результаты анализа по критерию Фишера показывают, что статистически значимые различия присутствуют в дисперсиях¹⁸ метрик диагностической точности врачей с опытом работы от 1 до 5 лет и свыше 10 лет, а также у врачей с опытом работы от 5 до 10 лет и свыше 10 лет.

Стоит отметить, что тест Фишера применим только в случае, если распределение данных обладает дисперсией¹⁹, в остальных случаях данный тест неприменим.

Результаты дисперсионного анализа, анализа по критерию Стьюдента и Фишера, представленные в этой части методических рекомендаций, носят чисто учебный характер и не имеют практической значимости, поскольку проведенный ранее тест Колмогорова – Смирнова показал, что данные не распределены по нормальному закону распределения, что не позволяет применить дисперсионный анализ и критерий Стьюдента. В случае критерия Фишера отсутствуют знания о том, к какому элементарному закону распределения принадлежат данные, и утверждать, что данное распределение обладает дисперсией, нельзя.

Для анализа данных, не подчиняющихся нормальному закону распределения, существуют непараметрические критерии анализа, такие как критерий Манна – Уитни, критерий Колмогорова – Смирнова, критерий Крускала–Уоллиса и критерий Данна.

Пример применения F-теста Фишера

Для примера применения F-теста Фишера используем набор данных *birthwt* из пакета **MASS**, содержащий вес недоношенных детей, сравним дисперсии веса детей, рожденных курящими и некурящими роженицами.

Листинг 8

```
library (MASS) #Подключаем пакет, содержащий набор данных birthwt
smokeWomen <- birthwt[birthwt$smoke==1,]$bwt #Вес детей у курящих
девушек
nosmokeWomen <- birthwt[birthwt$smoke==0,]$bwt #Вес детей у некурящих
# девушек
```

¹⁸ При проведении множественного (когда количество групп больше двух) сравнения дисперсий по критерию Фишера необходимо учитывать поправку на множественное сравнение, более подробно об этом см. в подразделе 2.7.

¹⁹ Обсуждение вопроса вычисления дисперсии для различных законов распределения выходит за рамки настоящих методических рекомендаций. Стоит отметить, что нормальный закон распределения данных (закон Гаусса) обладает дисперсией, тогда как закон распределения Коши дисперсией не обладает.

Продолжение листинга 8

```
#Проверяем длину выборок
length(smokeWomen) # Определяем количество значений в выборке
курящих
[1] 74
length(nosmokeWomen) # Определяем количество значений в выборке
некурящих
[1] 115
#=====
====
# Применяем функцию var.test() для сравнения дисперсий различного
объема
#=====
====
#Проводим тест Фишера на равенство дисперсий
#=====
====
var.test(smokeWomen, nosmokeWomen)
#=====
====
# Результаты применения теста Фишера
#=====
====
                                F test to compare two variances

data:  smokewomen and nosmokewomen
F = 0.76809, num df = 73, denom df = 114, p-value = 0.2254
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5104756 1.1783550
sample estimates:
ratio of variances
 0.7680924
```

Анализ результатов применения теста Фишера показывает, что статистически значимых различий в дисперсиях веса младенцев у курящих и некурящих рожениц нет.

2.4. Критерий Крускала–Уоллиса

Критерий Крускала–Уоллиса применяется для сравнения трех и более независимых групп по количественному показателю, закон распределения случайной величины хотя бы одной из сравниваемых групп должен быть отличен от нормального. Критерий основан на том принципе, что чем меньше взаимное пересечение выборок, тем выше уровень значимости критерия (19):

$$H = \frac{12}{N(N+1)} \sum_{i=1}^p \frac{R_i^2}{n} - 3(N+1), \quad (19)$$

где N – общее число членов в обобщенной выборке; n – число членов в каждой отдельной выборке; R – квадрат суммы рангов выборки; p – число уровней в факторе.

Проведение анализа данных по критерию Крускала–Уоллиса на языке программирования R осуществляется посредством функции *kruskal.test()*. Рассматриваемый критерий был применен для выявления статистически значимых различий в метриках диагностической точности 100 врачей, разделенных по субспециализациям, и в метриках диагностической точности 100 врачей, разделенных по опыту работы. В результате было установлено, что метрики диагностической точности, разделенные по субспециализациям, статистически значимо различаются между собой с $p\text{-value} = 0,002766$, тогда как влияние опыта работы на метрики диагностической точности статистически не значимы и $p\text{-value} = 0,1721$. Данный тест позволяет понять, есть ли между группами статистически значимые различия или нет, но не дает установить, между какими именно это различие наблюдается. Чтобы понять, между какими независимыми группами имеются статистически значимые различия, необходимо провести попарное сравнение тестом Манна–Уитни, тестом Данна (если распределение хотя бы одной из сравниваемых групп отлично от нормального закона распределения) или тестом Стьюдента (если обе группы подчиняются нормальному закону) и учесть поправку на множественное сравнение²⁰.

Применение критерия Крускала–Уоллиса для анализа зависимости длительности прохождения теста 100 врачами от субспециализации и от опыта работы показал, что данные факторы статистически не значимы.

Пример применения теста Крускал – Уоллиса

Для демонстрации применения теста Крускала–Уоллиса используем набор данных *Cushings* из пакета **MASS**, содержащий данные по скорости выведения с мочой (мг/24 ч) тетрагидрокортизона и прегнантриола у пациентов с подтвержденным диагнозом «синдром Кушинга» и выделенными проявлениями, кодируемыми как: а (аденома), b (двусторонняя гиперплазия), с (карцинома) или u (неизвестно).

Листинг 9

```
library(MASS)
#=====
# Проводим тест Крускала – Уоллиса для выявления наличия различий
# скорости выведения тетрагидрокортизона у пациентов с разными
# проявлениями синдрома Кушинга
#=====
=
kruskal.test(Tetrahydrocortisone~Type, data=Cushings)
#=====
=
# Результаты применения теста
#=====
=
kruskal-wallis rank sum test

data: Tetrahydrocortisone by Type
kruskal-wallis chi-squared = 13.509, df = 3, p-value = 0.003655
#=====
==
# Проводим тест Крускала–Уоллиса для выявления наличия различий
```

²⁰ Поправки на множественное сравнение рассматриваются в подразделе 2.7.

Продолжение листинга 9

```
# скорости выведения у пациентов с разными проявлениями  
# синдрома Кушинга  
#=====  
==  
kruskal.test(Pregnanetriol~Type, data=Cushings)  
#=====  
===  
#Результаты применения теста  
#=====  
===  
kruskal-wallis rank sum test  
data: Pregnanetriol by Type  
kruskal-wallis chi-squared = 8.743, df = 3, p-value = 0.03291  
#=====  
===  
# Проводим тест Крускала – Уоллиса для выявления наличия различий  
# скорости выведения прегнантриола и тетрагидрокортизона, у пациентов с  
# разными проявлениями синдрома Кушинга  
#=====  
===  
kruskal.test(Pregnanetriol+Tetrahydrocortisone~Type, data=Cushings)  
#=====  
===  
# Результаты применения теста  
#=====  
===  
kruskal-wallis rank sum test  
data: Pregnanetriol + Tetrahydrocortisone by Type  
kruskal-wallis chi-squared = 11.146, df = 3, p-value = 0.01096  
#=====
```

Результаты анализа, проведенного с использованием теста Крускала–Уоллиса, показывают, что скорость выведения с мочой тетрагидрокортизона и прегнантриола у пациентов, принадлежащих к группам с аденомой, двусторонней гиперплазией, карциномой или неустановленными причинами синдрома Кушинга, статистически значимо различаются между собой, т. к. $p\text{-value} < 0,05$. Для выделения различающихся пар применяется критерий Манна–Уитни с поправкой на множественность²¹.

2.5. Критерий Данна²²

Критерий (тест) Данна является непараметрическим аналогом теста Тьюки, применяемого в дисперсионном анализе (ANOVA). Основным преимуществом данного теста является его независимость от типа распределения данных и возможность применения в тех случаях, когда исследуемые данные распределены

²¹ Подробнее поправки на множественное сравнение рассматриваются в подразделе 2.7.

²² Dunn O. J. Multiple comparisons using rank sums // Technometrics. 1964. Vol. 6, №3. P. 241–252.

отличным от нормального закона распределения. Так же, как и в случае с критерием Крускала–Уоллиса, критерий Данна является ранговым и все вычисления построены на основании рангов, присвоенных значениям в упорядоченном по возрастанию ряду. Основным значением, вычисляемым в данном критерии, является чувствительность или «стандартизованный эффект», рассчитываемый по уравнению (20):

$$z = \frac{y_m}{\sigma_m}, \quad (20)$$

где y_m – вычисляется по уравнению (21):

$$y_m = \frac{\sum_i T_{im}}{\sum_i n_i} - \frac{\sum_{i'} T_{i'm}}{\sum_{i'} n_{i'}}, \quad (21)$$

где i и i' – суммы по различным подмножествам в исследуемых данных, принимающие значения от 1 до k ; m – порядковый номер суммарного количества рангов в подмножествах, принимающих значения от 1 до p ; n – количество данных в исследуемом подмножестве, такое, что (22):

$$N = \sum_{i=1}^k n_i, \quad (22)$$

где N – количество исследований в генеральной совокупности.

Дисперсия, присутствующая в знаменателе уравнения (20), вычисляется по уравнению (23):

$$\sigma_m^2 = \left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^r (T_s^3 - T_s)}{12(N-1)} \right] \cdot \left[\frac{1}{\sum_i n_i + \sum_{i'} n_{i'}} \right], \quad (23)$$

где $N = \sum_i n_i + \sum_{i'} n_{i'}$ – равно количеству исследований, присутствующих в двух группах, а не общему количеству исследований; r – количество связанных рангов; T_s – количество наблюдений, связанных с s -м конкретным связанным значением.

На языке программирования R провести тест Данна можно с помощью функции *dunnTest()*, входящей в пакет **FSA**.

Пример использования теста Данна

Для примера использования теста Данна возьмем набор данных *Cushings* из пакета **MASS**, содержащий скорость выделения с мочой тетрагидрокортизона и прегнантриола у пациентов с подтвержденным диагнозом «синдром Кушинга».

Листинг 10

```
library(MASS) #Пакет, содержащий набор данных Cushings
library(FSA) #Пакет, содержащий функцию dunnTest()
#=====
=====
# Проводим исследование различий в скорости выделения
тетрагидрокортизона
#=====
=====
dunnTest(Tetrahydrocortisone~Type, data=Cushings, method= "bonferroni"23)
#=====
=====
#Результат применения теста
#=====
=====
Comparison      Z          P.unadj24      P.adj25
1      a - b -2.3103355 0.0208695864 0.125217518
2      a - c -3.4167374 0.0006337638 0.003802583
3      b - c -1.5991401 0.1097894782 0.658736869
4      a - u -2.8195085 0.0048097252 0.028858351
5      b - u -0.8419709 0.3998042624 1.000000000
6      c - u  0.7284415 0.4663433840 1.000000000
```

Анализ результатов применения теста Данна показывает, что статистически значимые различия в скорости выведения тетрагидрокортизона наблюдаются у пациентов в группах «а – с» (аденома – карцинома) и «а – u» (аденома – не установлено). В остальных случаях статистически значимых различий не выявлено.

2.6. Критерий Манна–Уитни²⁶

Критерий Манна – Уитни относится к категории непараметрических ранговых критериев, применяющихся для выявления различий по количественному показателю в двух не связанных между собой выборках. Данный критерий может применяться к выборкам, имеющим различную длину, и определяется по уравнению (24):

$$U = \min\{U_1, U_2\}, \quad (24)$$

где

$$U_1 = n_1 * n_2 + \frac{n_1(n_1 + 1)}{2} - R_1, \quad (25)$$

$$U_2 = n_1 * n_2 + \frac{n_2(n_2 + 1)}{2} - R_2, \quad (26)$$

²³ В качестве метода (method) указана поправка Бонферрони на множественное сравнение. Более детально поправки на множественное сравнение рассмотрены в подразделе 2.7.

²⁴ Значение p-value до учета поправки Бонферрони на множественное сравнение.

²⁵ Значение p-value после учета поправки Бонферрони на множественное сравнение.

²⁶ При применении критерия Манна–Уитни должна учитываться поправка на множественное сравнение, более подробно см. в подразделе 2.7.

где R_1, R_2 – сумма рангов, пришедшаяся на долю первой и второй выборок соответственно; n_1, n_2 – количество исследований в первой и второй выборке соответственно.

Вторым вариантом расчета значений критерия Манна–Уитни является уравнение (27):

$$U = n_1 * n_2 + \frac{n_{max}(n_{max} + 1)}{2} - R_{max}, \quad (27)$$

где n_1, n_2 – количество исследований в первой и второй выборках соответственно; n_{max} – число исследований в соответствующей группе с максимальным числом рангов; R_{max} – максимальное число рангов в выборке.

Выявление различий в двух выборках по критерию Манна–Уитни на языке R осуществляется с помощью функции *wilcox.test ()*. Применим критерий Манна–Уитни для выявления различий в выборках по субспециализации в метриках диагностической точности 100 врачей. Результаты применения теста представлены в таблице 11.

Таблица 11 – Результаты сравнения метрик диагностической точности между субспециализациями врачей по критерию Манна–Уитни

| Сравниваемые пары | Уровень статистической значимости |
|--|-----------------------------------|
| ММГ – КТ | 0,3572 |
| МРТ – КТ | 0,7546 |
| РГ, ФЛГ – КТ | 0,01617 |
| МРТ – ММГ | 0,4243 |
| РГ, ФЛГ – ММГ | 0,009784 |
| РГ, ФЛГ – МРТ | 0,03574 |
| <i>Примечание. ММГ – маммография, КТ – компьютерная томография, МРТ – магнитно-резонансная томография, РГ – рентгенография, ФЛГ – флюорография</i> | |

Результаты применения теста Манна–Уитни показывают, что статистически значимые различия в метриках диагностической точности проявляются между врачами с субспециализацией РГ, ФЛГ и КТ, у групп РГ, ФЛГ и ММГ и у групп РГ, ФЛГ–МРТ, в остальных случаях различия статистически не значимы.

Аналогичное сравнение проведем для групп, разделенных по критерию опыта работы. Результаты сравнения представлены в таблице 12.

Таблица 12 – Результаты сравнения метрик диагностической точности между врачами с разным опытом работы по критерию Манна–Уитни

| Сравниваемые пары | Уровень статистической значимости |
|-------------------|-----------------------------------|
| 1-5 – 0-1 | 0,06897 |
| 10+ – 0-1 | 0,1189 |
| 5-10 – 0-1 | 0,06061 |
| 10+ – 1-5 | 0,2633 |
| 5-10 – 1-5 | 0,1581 |
| 5-10 – 10+ | 0,6816 |

Результаты вычислений по критерию Манна–Уитни показывают, что статистически значимых различий в группах по опыту работы нет.

Исходя из проведенных исследований, можно утверждать, что метрики диагностической точности зависят от субспециализации и не зависят от опыта работы врачей, участвовавших в эксперименте.

Пример применения теста Манна–Уитни

Для демонстрации применения теста Манна–Уитни используем набор данных *Cushings* из пакета **MASS**, содержащий значения скорости выведения с мочой (мг/24 ч) тетрагидрокортизона и прегнантриола у пациентов с подтвержденным диагнозом «синдром Кушинга» и выделенными проявлениями, кодируемыми как: a (аденома), b (двусторонняя гиперплазия), c (карцинома) или u – неизвестно.

Листинг 11

```
library(MASS) #Подключаем библиотеку, содержащую набор данных
Cushings
adenCushing <- Cushings[Cushings$Type == "a"],$Pregnanetriol
geperCushing <- Cushings[Cushings$Type == "b"],$Pregnanetriol
#=====
==
# Проводим сравнение в парах с помощью теста Манна–Уитни
#=====
==
wilcox.test(adenCushing, geperCushing)
#=====
===
# Результат сравнения
#=====
===
      wilcoxon rank sum test with continuity correction
data:  adenCushing and geperCushing
w = 25, p-value = 0.6239
alternative hypothesis: true location shift is not equal to 0
```

Сравнительный анализ в двух группах показал, что статистически значимых различий в скорости выведения с мочой прегнантриола у пациентов с аденомой и двусторонней гиперплазией нет. Необходимо отметить, что в данном примере не учтена поправка на множественное сравнение, более детально проведение множественного сравнения рассмотрено в подразделе 2.7.

2.7. Поправки на множественную проверку гипотез

При разделении исследуемой популяции на несколько групп (например, по гендерному признаку и по признаку «курящий/некурящий» и т.д.) можно найти такое разделение, при котором выделенная группа будет статистически значимо отличаться от основной популяции, т.е. возникнет кажущееся отличие. Проблема множественных сравнений в биомедицинских исследованиях заключается в том, что увеличение числа проверяемых в процессе исследования статистических гипотез приводит к увеличению вероятности возникновения ошибок первого рода. Увеличение вероятности возникновения ошибок первого рода приводит к тому, что

исследователи могут констатировать наличие случайных, но статистически значимых различий или связей, что в свою очередь приводит к ошибочным заключениям и выводам. Для решения данной проблемы были разработаны поправки на множественное сравнение.

Поправка на множественную проверку гипотез – это способ устранения ошибки первого рода, накапливающейся при необходимости построения семейства статистических выводов^{27, 28}. Проблема множественного сравнения является одним из краеугольных вопросов статистики и активно обсуждается по настоящее время. Для практического применения чаще всего используются следующие поправки на множественную проверку гипотез:

- 1) поправка Бонферрони²⁹;
- 2) метод Холма³⁰;
- 3) метод Шидака (Сидака)³¹;
- 4) метод Шидака (Сидака)–Холма³²;
- 5) поправка Беньямини–Хохберга³³;
- 6) метод Бенджамина–Иекутиели³⁴.

Каждый из представленных методов может быть использован при интерпретации результатов попарного сравнения, однако высокая консервативность большинства из представленных методов и неоднозначное мнение специалистов в области статистического анализа требуют аккуратности в интерпретации результатов применения различных поправок на множественное сравнение. Наиболее статистически мощной считается поправка Холма, тогда как при проведении сравнения большого количества групп рекомендуется применять поправку Беньямини–Хохберга.

По этой причине при планировании эксперимента и проведении анализа стоит избегать случаев возникновения множественного сравнения (методы дисперсионного анализа ANOVA и Краскелла–Уоллиса не требуют применения поправок на множественное сравнение³⁵).

На языке программирования R учесть поправку на множественное сравнение можно с использованием функции *p.adjust()* из пакета **MASS**.

²⁷ Наркевич А. Н., Виноградов К. А., Гржибовский А. М. Множественные сравнения в биомедицинских исследованиях: проблема и способы решения // Экология человека. 2020. Т. 10. С. 55–64.

²⁸ Майкл Х., Френсис Г., Кларк А. Статистика и планирование эксперимента для непосвященных: как отучить статистику лгать. М.: ДМК Пресс, 2023.

²⁹ Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita // Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936. Vol. 8. P. 3–62.

³⁰ Holm S. A simple sequentially rejective multiple test procedure // Scandinavian journal of statistics. 1979. Vol. 6. P. 65–70.

³¹ Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions // Journal of the American Statistical Association. 1967. Vol. 62 (318). P. 626–633.

³² Holm S. A simple sequentially rejective multiple test procedure // Scandinavian journal of statistics. 1979. Vol. 6. P. 65–70.

³³ Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the Royal statistical society: series B (Methodological). 1995. Vol. 57, №1. P. 289–300.

³⁴ Benjamini Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency // Annals of statistics. 2001. Vol. 29. P. 1165–1188.

³⁵ Майкл Х., Френсис Г., Кларк А. Статистика и планирование эксперимента для непосвященных: как отучить статистику лгать. М.: ДМК Пресс, 2023.

*Пример вычисления уровня статистической значимости
с учетом поправки на множественное сравнение*

Рассмотрим применение поправки на множественное сравнение на примере набора данных *anorexia* из пакета MASS.

Листинг 12

```
library(MASS) #Подключаем пакет, содержащий набор данных anorexia
#Проверяем наличие статистически значимых различий в данных
kruskal.test(Treat~Prewt, data = anorexia) #До проведения лечения
#=====
=====
#Результаты проведения анализа
#=====
=====
kruskal-wallis rank sum test

data: Treat by Prewt
kruskal-wallis chi-squared = 57.067, df = 57, p-value = 0.4726
#=====
=====
kruskal.test(Treat~Postwt, data = anorexia)
#=====
=====
kruskal-wallis rank sum test

data: Treat by Postwt
kruskal-wallis chi-squared = 67.71, df = 66, p-value = 0.4186
#=====
=====
# Различий в группах не выявлено. Проведем парное сравнение тестом
Манна – Уитни
#=====
=====
prewtCont<-anorexia[anorexia$Treat == "Cont",]$Prewt
prewtCBT <- anorexia[anorexia$Treat == "CBT",]$Prewt
prewtFT <- anorexia[anorexia$Treat == "FT",]$Prewt
pContCBT <- wilcox.test(prewtCont, prewtCBT)$`p.value` # Результат
0.4380002
pContFT <- wilcox.test(prewtCont, prewtFT)$`p.value` # Результат 0.3982123
pCbtFT <- wilcox.test(prewtCBT, prewtFT)$`p.value` # Результат 0.7156885
pAll <- c(pContCBT, pContFT, pCbtFT)
p.adjust(pAll, method= "BH") #Рассчитываем уровень статистической
# значимости с учетом поправки Беньямини –
Хохберга
#=====
=====
#Результаты введения поправки
#=====
=====
[1] 0.7156885 0.6570003 0.6570003
```

Результаты парного сравнения с учетом поправки на множественность гипотез Беньямини–Хохберга показывают, что статистически значимых различий в весе пациентов с подтвержденным диагнозом «анорексия» до проведения терапии нет.

*Пример исследования данных без дополнительной информации
по уровням факторов*

Вторая часть методических рекомендаций является обобщающей и демонстрирует ход рассуждений аналитика при работе с данными. Для демонстрации рассуждений используется набор данных по исследованиям метрик диагностической точности 100 врачей, проходивших тестирование в рамках московского эксперимента по компьютерному зрению. Метрики диагностической точности врачей рассчитывались методом SRMC (single reader – multi case³⁶). Используя все вышеописанные статистические методы, предлагается пройти весь путь обработки информации, находящейся в наборе данных, источники которых указаны в настоящих методических рекомендациях.

Проведение анализа данных без наличия предварительно собранных материалов о факторах и уровнях факторов, взаимодействующих между собой и влияющих на конечный результат, является далеко не тривиальной задачей и по большей части опирается на опыт исследователя и на возможность получить дополнительные данные. В этой части представлен подход, основанный на методе последовательных приближений, который позволяет выделить подгруппы данных и провести их исследования. На первом шаге необходимо построить диаграмму рассеяния метрик диагностической точности в зависимости от длительности прохождения теста каждым врачом. Результаты представлены на рисунке 11.

³⁶ См. подраздел 2.1.1.

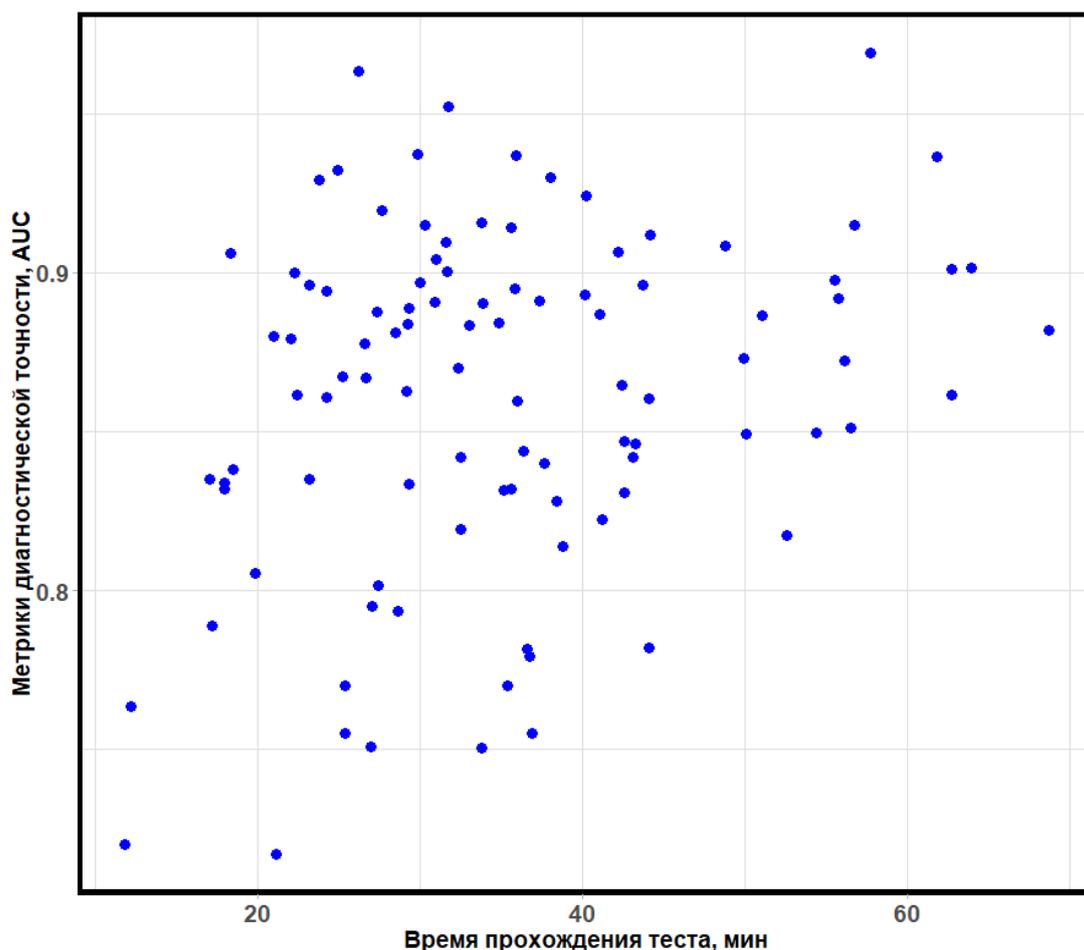


Рисунок 11 – Диаграмма рассеяния метрик диагностической точности в зависимости от времени прохождения теста для 100 врачей

Из диаграммы рассеяния видно, что данные сосредотачиваются вблизи определенной области. Для разделения построенной зависимости и визуального разделения данных на области вычисляем среднее значение времени прохождения теста и среднее значение метрик диагностической точности, затем отображаем их на графике. Такие же вычисления и построения производятся для медианных оценок. На рисунке 12 представлены результаты построения средних и медиан для представленных данных.

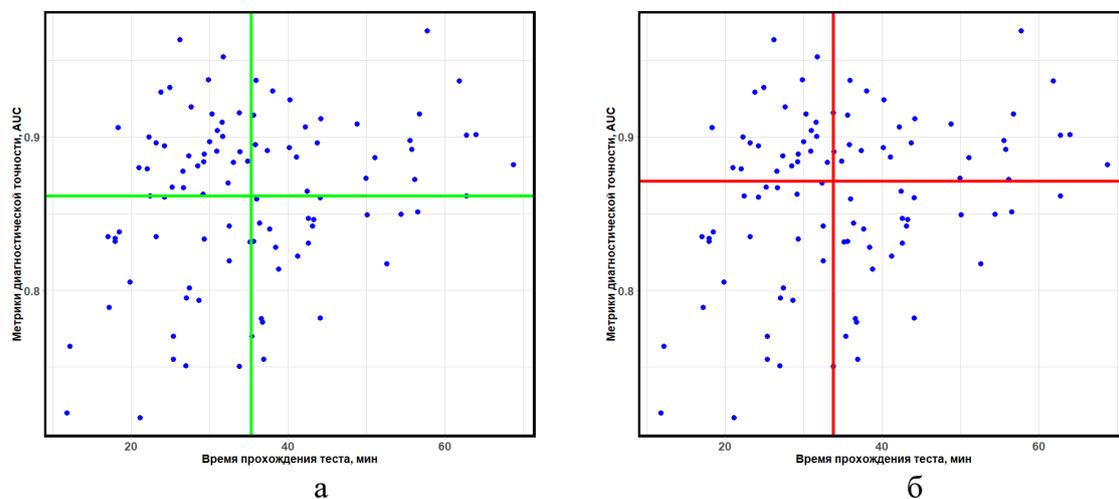


Рисунок 12 – Разделение данных на группы по анализу средних (а) и медиан (б)

Из рисунка 12 видно, что данные разделяются на 4 квадрата, в каждом из которых находится определенное количество исследований. Необходимо установить, принадлежат ли выделенные данные к одному и тому же распределению или нет. Для этого можно воспользоваться критерием Колмогорова–Смирнова. На рисунке 13 представлены результаты выделения групп относительно медиан.

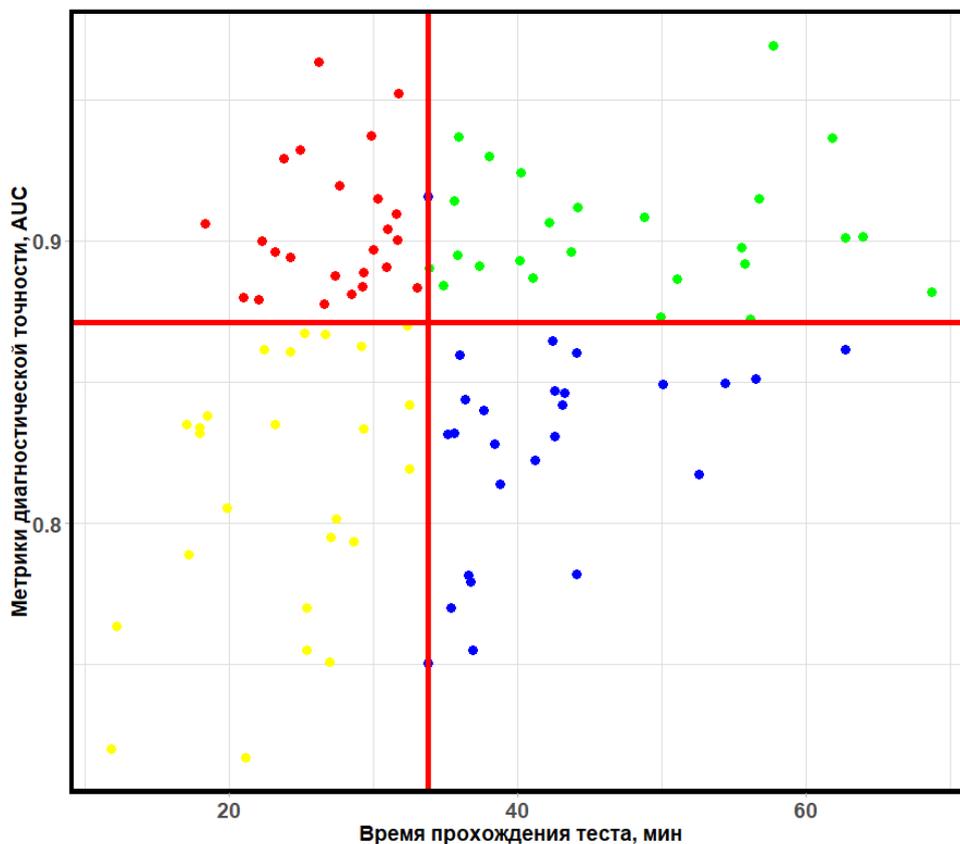


Рисунок 13 – Разделение данных относительно медиан

Для этого выделим медиану времени в зеленых данных и сравним зеленые данные между собой относительно принадлежащей им медианы, а также сравним новые группы с красной группой по критерию Колмогорова–Смирнова. То же самое сделаем для желтой и синей групп данных. Применение теста Колмогорова–Смирнова к данным, выделенным красным, желтым и синим цветом, показывает (красные, желтые и синие точки на рисунке 13), что красные и желтые данные принадлежат к разным типам распределений, красные и синие принадлежат к разным типам распределений. Желтые и зеленые данные также принадлежат к разным типам распределений. В случае желтых и синих, а также красных и зеленых данных критерий Колмогорова–Смирнова не позволяет произвести оценку различий в данных. Проведем оценку принадлежности «приграничных» данных к одному типу распределений между зелеными и красными, а также между желтыми и синими данными.

На рисунке 14 представлено разделение зеленых данных на две подгруппы по медиане.

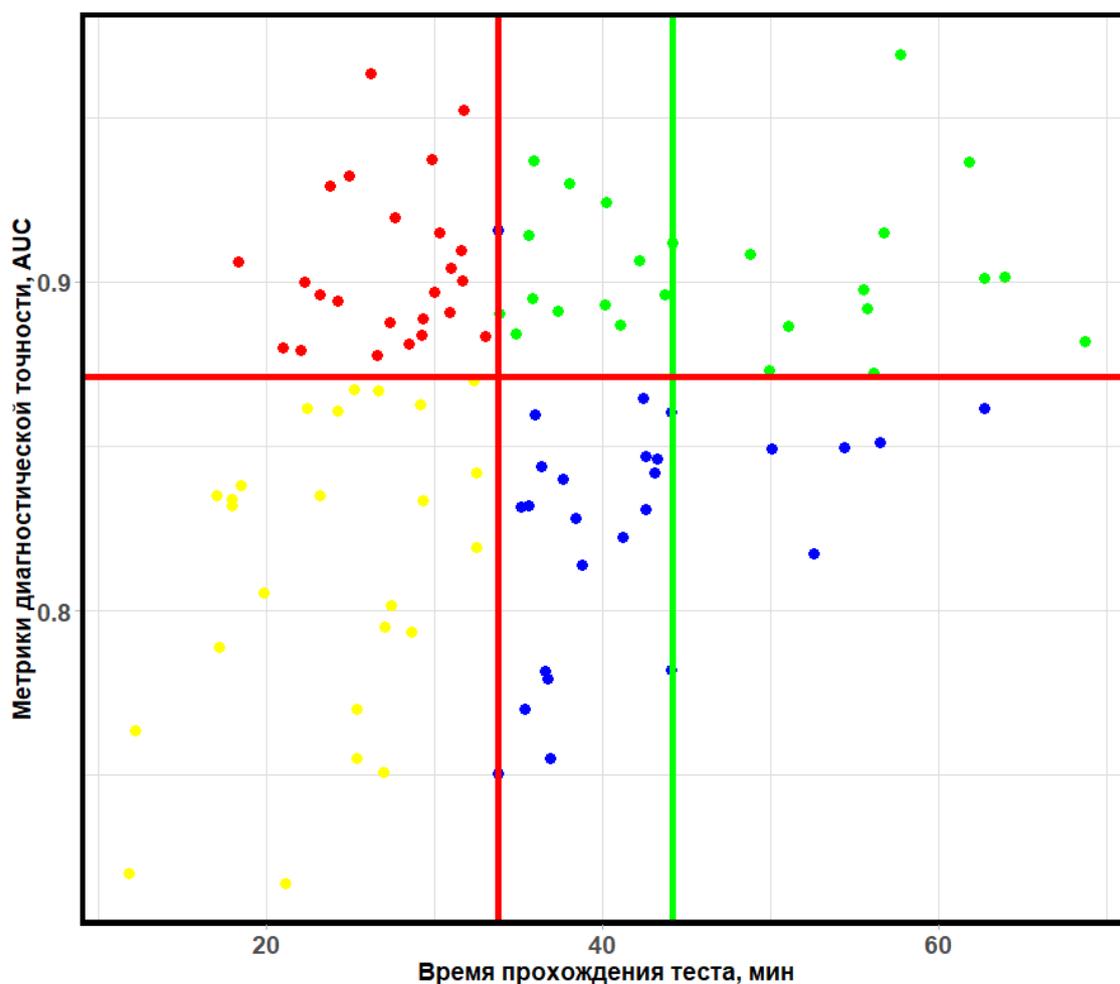


Рисунок 14 – Разделение зеленых данных на две группы по медиане

Применение критерия Колмогорова–Смирнова к двум группам зеленых данных, а также к оценке слева от медианы зеленых данных и красных данных показывает, что зеленые данные слева и справа от медианы относятся к одному и тому же типу распределения, а данные слева от медианы и красные данные – к разным типам распределений. Таким образом, полученный результат свидетельствует о том, что красные и зеленые данные принадлежат к разным типам распределений, т. е. к различным группам. Анализ желтых и синих данных по критерию Колмогорова–Смирнова показал, что данные относятся к одному закону распределения. Таким образом, статистически значимо различаются три группы врачей из 100 прошедших тестирование:

1. Группа с высокими метриками диагностической точности и высокой скоростью просмотра исследований.
2. Группа с высокими метриками диагностической точности и более медленной скоростью просмотра исследований
3. Группа с более низкими метриками диагностической точности и различной скоростью прохождения теста.

Оценка велась относительно медиан распределений диагностической точности и времени прохождения теста. На рисунке 15 представлен результат выделения групп.

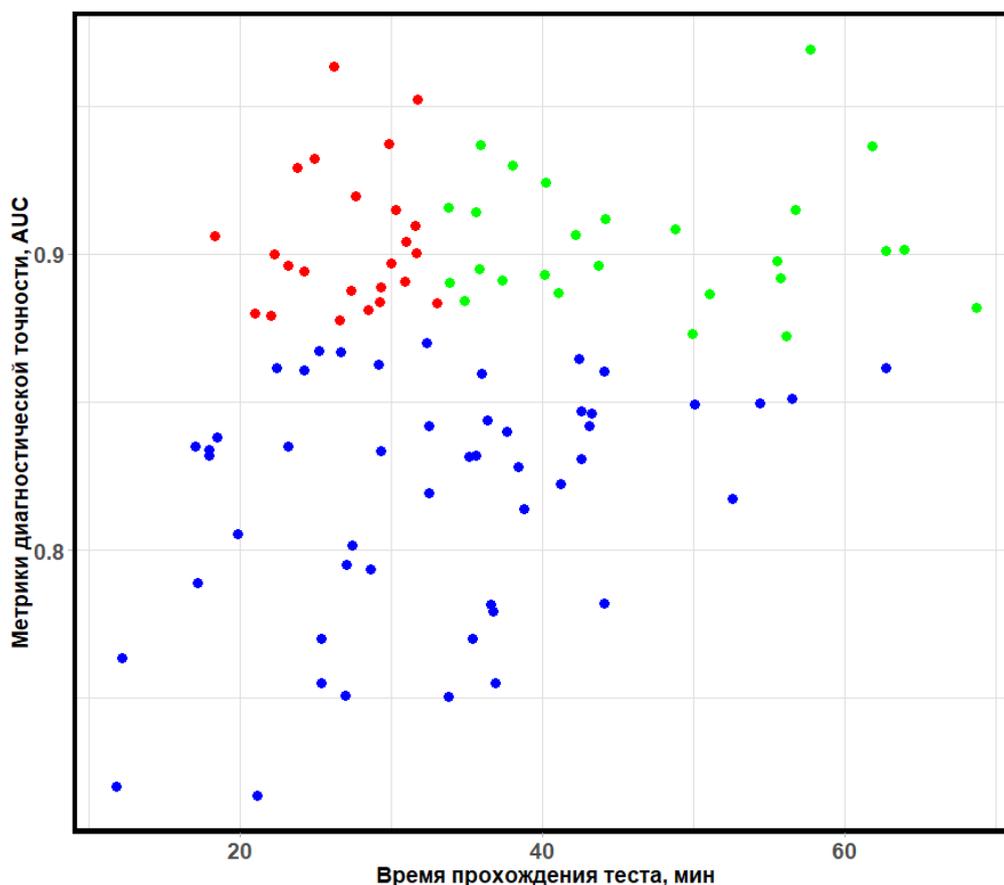


Рисунок 15 – Выделенные группы врачей по метрикам диагностической точности и времени прохождения тестирования

Проведенный анализ позволил по результатам тестирования выделить три группы врачей, статистически значимо различающихся между собой.

На рисунке 16 представлены результаты анализа красной группы по субспециализациям и по опыту работы.

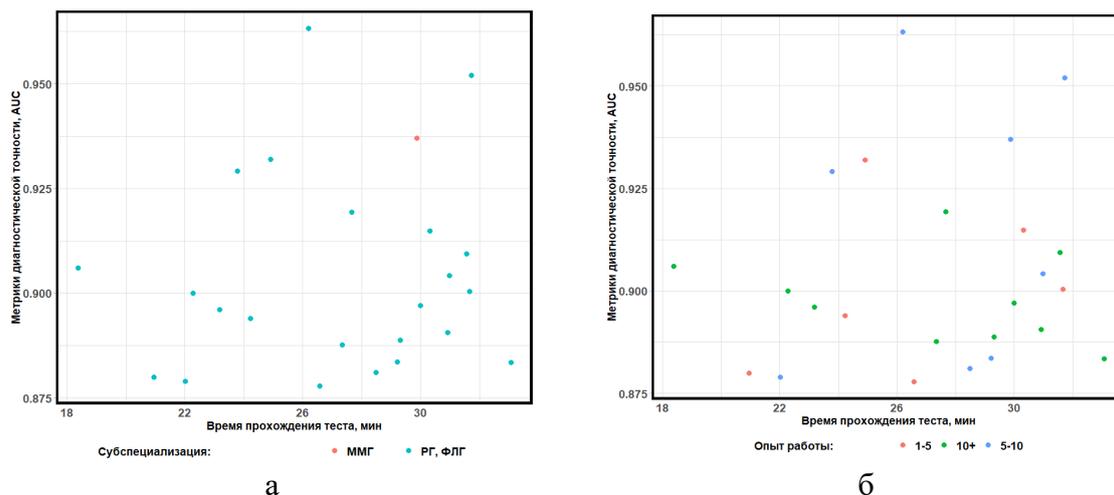


Рисунок 16 – Диаграммы рассеяния метрик диагностической точности в зависимости от времени прохождения теста в красной группе с разделением по субспециализациям (а) и по опыту работы (б)

Анализ диаграмм рассеяния (рисунок 16а) красной группы показывает, что из 24 врачей, обладающих малым временем просмотра исследований и высокой

диагностической точностью, один врач имел субспециализацию ММГ, а 23 врача – субспециализацию РГ, ФЛГ. Диаграмма рассеяния (рисунок 16б) с выделением групп врачей по опыту работы показывает, что в красной группе находятся 6 врачей с опытом работы от 1 до 5 лет, 8 – с опытом работы от 5 до 10 лет и 10 – с опытом работы свыше 10 лет. Таким образом, с увеличением опыта работы в красной группе увеличивается количество врачей на два человека.

На рисунке 17 представлен результат анализа зеленой группы врачей по субспециализации и опыту работы.

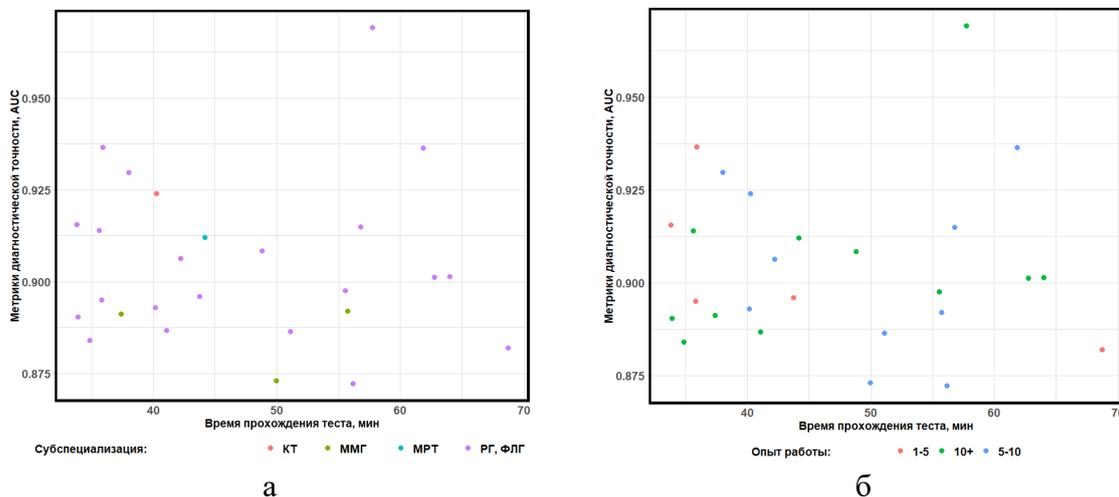


Рисунок 17 – Зависимость метрик диагностической точности врачей зеленой группы от времени прохождения теста с учетом субспециализации (а) и опыта работы (б)

Анализ диаграммы рассеяния зеленой группы показывает, что количество врачей, находящихся в данной группе, составляет 26 человек, из которых 21 имеет субспециализацию РГ, ФЛГ, 3 – ММГ, 1 – КТ и 1 – МРТ (рисунок 17а). В зеленую группу входят 5 врачей с опытом работы от 1 до 5 лет, 10 врачей с опытом работы от 5 до 10 лет и 11 врачей с опытом работы свыше 10 лет (рисунок 17б).

На рисунке 18 представлена зависимость метрик диагностической точности синей группы врачей от времени прохождения теста.

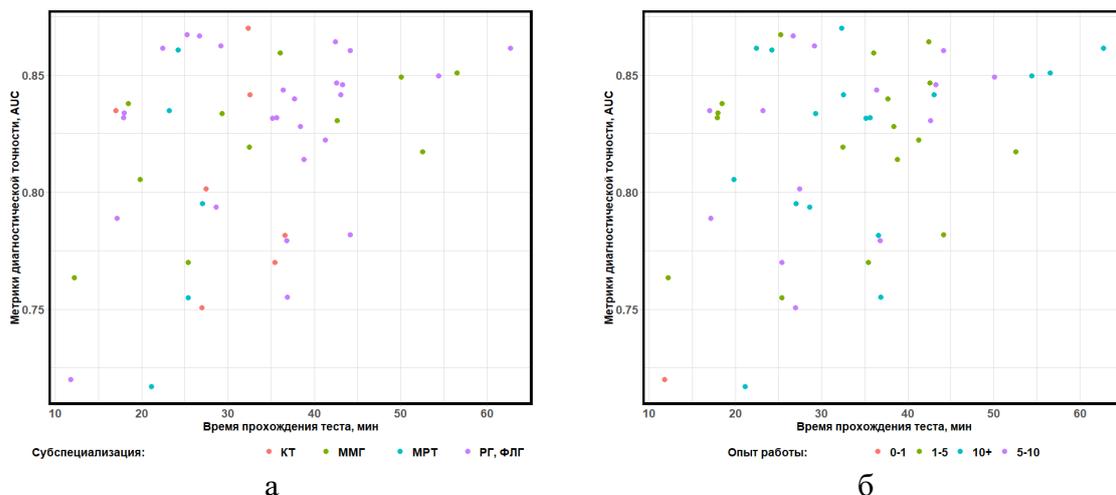


Рисунок 18 – Зависимость метрик диагностической точности врачей синей группы от времени прохождения теста с учетом субспециализации (а) и опыта работы (б)

Анализ диаграммы рассеяния синей группы показывает, что количество врачей в синей группе составляет 50 человек, из них 26 имеют субспециализацию РГ, ФЛГ, 5 – МРТ, 11 – ММГ и 8 – КТ. Один врач имеет опыт работы менее года, 17 – от 1 года до 5 лет, 14 – от 5 до 10 лет и 18 – свыше 10 лет.

В качестве курсового проекта аспирантам предлагается самостоятельно выбрать набор данных из приведенных ниже источников и провести их статистический анализ.

2.8. Источники наборов данных для проведения самостоятельного статистического анализа

При работе над курсовым проектом аспиранты могут воспользоваться следующими источниками данных:

- 1) <https://www.mdpi.com/journal/data>;
- 2) <https://www.nature.com/sdata>;
- 3) <https://www.ncei.noaa.gov/products>.

Представленный перечень источников не является исчерпывающим, и для курсовой работы могут быть использованы любые источники данных, представленные в академической литературе, в том числе собранные аспирантами при проведении самостоятельных исследований. Если аспиранты осуществляют самостоятельный сбор и подготовку данных, то необходимо руководствоваться методическими рекомендациями по подготовке наборов данных³⁷.

³⁷ Васильев Ю. А., Арзамасов К. М., Владимирский 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.

3. СТАТИСТИЧЕСКИЕ КРИТЕРИИ СОГЛАСИЯ

3.1. Критерий согласия X^2 Пирсона

При обработке результатов исследований и построении различных статистических моделей может возникнуть задача определить, принадлежит ли опытное распределение данных к простому теоретическому типу и имеется ли статистически значимая зависимость между анализируемыми данными. Для определения принадлежности распределения данных к теоретическому типу существуют критерии согласия; наиболее часто применяемым и рекомендуемым³⁸ является критерий X^2 Пирсона.

В основе применения критерия Хи-квадрат лежит группировка наблюдений по возрастанию исследуемой случайной величины с последующим разбиением области определения случайной величины на k непересекающихся интервалов, подсчитывается количество интервалов n_i , и вероятность попадания случайной величины в i -й интервал определяется по уравнению (28):

$$P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x, \theta) dx, \quad (28)$$

где $f(x, \theta)$ – функция плотности теоретического распределения.

Критерий Хи-квадрат определяет отклонение $\frac{n_i}{N}$ от $P_i(\theta)$ по уравнению (29):

$$S_{X^2} = N \sum_{i=1}^k \frac{(\frac{n_i}{N} - P_i(\theta))^2}{P_i(\theta)}, \quad (29)$$

где $N = \sum_{i=1}^k n_i$ – общее количество интервалов.

Необходимо отметить, что критерий X^2 Пирсона является достаточно старым и классическим критерием, но не единственным. Существует ряд информационных критериев, например критерии Такиучи, Акаике и Байеса, с помощью которых возможно осуществлять выбор наиболее близких теоретических распределений, а также проводить оценку регрессионных моделей. Рассмотрение информационных критериев выходит за рамки настоящих методических рекомендаций; при наличии заинтересованности аспиранты могут самостоятельно с ними ознакомиться.

Для проведения тестирования данных по методу Хи-квадрат на языке R существует команда `chisq.test()`, `pchisq()` и `qchisq()`, входящая в библиотеку `stats` языка R.

Примеры применения теста X^2 Пирсона

Используем набор данных, содержащий фармакологическую кинетику индометацина (*Indometh*). В данном примере будут рассмотрены вычисления критерия согласия со статистикой X^2 Пирсона и уровня статистической значимости.

³⁸ Р 50.1.033-2001. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерий Хи-квадрат.

Листинг 13

```

library(MASS) # Подключаем пакет, содержащий набор данных Indometh
library(fitdistrplus) # Подключаем пакет, содержащий функции
вычисления
# параметров распределения по методу
максимального
# правдоподобия
#=====
=
# Создаем вектор, содержащий данные по фармакологической кинетике
# индометацина
#=====
=
indometData <- Indometh$conc
#=====
=
# Вычисляем параметры нормального распределения по методу
# максимального правдоподобия
#=====
=
normMLE <- fitdist(indometData, "norm")
meanIndometh <- normMLE$estimate[1] #Среднее значение
sdIndometh <- normMLE$estimate[2] # Среднее квадратическое
отклонение
#=====
==
# Создаем массив нормально распределенных данных
#=====
==
nData <- length(Indometh$conc)
teorIndometh <- rnorm(nData, mean = meanIndometh, sd = sdIndometh)
#=====
==
# Рассчитываем значение критерия  $\chi^2$  Пирсона
chisq <- sum((Indometh$conc-teorIndometh)^2/teorIndometh)
#Вычисляем значение уровня значимости
pchisq(q=chisq, df=length(teorIndometh)-1)
#=====
==
# Результат вычисления
#=====
==
[1] 7.001251e-24

```

Результаты вычислений уровня статистической значимости меньше 0,05, т. е. есть достаточно оснований сказать, что полученные в ходе эксперимента данные отличаются от нормального закона распределения.

3.2. Графический метод определения согласия Бланда–Альтмана³⁹

Графический метод анализа согласованности результатов измерений Бланда–Альтмана применяется в аналитической химии и биомедицине для проведения соответствия между различными методами аналитического контроля химического состава вещества, биохимических или биомедицинских анализов, проводимых в клинической практике. Для примера построения графика Бланда–Альтмана используем данные, приведенные в оригинальной публикации⁴⁰ и представленные в таблице 13.

Таблица 13 – Пиковая скорость выдыхаемого потока воздуха, измеренная двумя методами: с помощью расходометра Райта и с помощью мини-измерителя Райта (измерения проводились в случайном порядке)

| Субъект | Измерения, проведенные с помощью расходометра Райта | | Измерения, проведенные с помощью мини-измерителя Райта | |
|---------|---|------------------------|--|------------------------|
| | Первый метод измерения | Второй метод измерения | Первый метод измерения | Второй метод измерения |
| 1 | 494 | 490 | 512 | 525 |
| 2 | 395 | 397 | 430 | 415 |
| 3 | 516 | 512 | 520 | 508 |
| 4 | 434 | 401 | 428 | 444 |
| 5 | 476 | 470 | 500 | 500 |
| 6 | 557 | 611 | 600 | 625 |
| 7 | 413 | 415 | 364 | 460 |
| 8 | 442 | 431 | 380 | 390 |
| 9 | 650 | 638 | 658 | 642 |
| 10 | 433 | 429 | 445 | 432 |
| 11 | 417 | 420 | 432 | 420 |
| 12 | 656 | 633 | 626 | 605 |
| 13 | 267 | 275 | 260 | 227 |
| 14 | 478 | 492 | 477 | 467 |
| 15 | 178 | 165 | 259 | 268 |
| 16 | 423 | 372 | 350 | 370 |
| 17 | 427 | 421 | 451 | 443 |

Анализ согласованности состоит из нескольких шагов:

1. Анализируем корреляции между результатами измерений, проведенных различными методами (в оригинальной публикации коэффициенты корреляции оцениваются по Пирсону, принадлежность данных к нормальному закону распределения не проверяется).

2. Вычисляем разницу между значениями, полученными с помощью основного прибора и с помощью его миниатюрной версии.

³⁹ Altman D. G., Bland, J. M. Measurement in medicine: the analysis of method comparison studies // Journal of the Royal Statistical Society Series D: The Statistician. 1983. Vol. 32, №3. P. 307–317.

⁴⁰ Bland J. M., Altman D. G. Statistical methods for assessing agreement between two methods of clinical measurement // The Lancet. 1986. Vol. 327, №8476. P. 307–310.

3. Вычисляем среднее значение и среднеквадратическое отклонение разниц.
4. Вычисляем среднее значение по двум измерениям из старого метода (критерий вычисления среднего значения – известная систематическая ошибка метода).
5. Строим график по вычисленным величинам.
6. Различия, выходящие за доверительный интервал, равный двум сигмам, являются не допустимыми.

В листинге 12 представлен код на языке R, позволяющий получить график согласованности результатов измерений, проведенных посредством двух различных устройств.

Листинг 14

```

library(dplyr) # Подключаем библиотеку анализа больших объемов данных
library(ggplot2) # Подключаем библиотеку для построения графиков
#Построение графика согласия двух методов измерений по Бланду
# Загружаем данные, содержащиеся в Таблице 13, из заранее
# подготовленного CSV файла.
BlantDataSet <- read.table(file = file.choose(), sep=";",dec = ".", header =
TRUE)
#Вычисляем матрицу корреляции по Пирсону
cor(BlantDataSet)
cor.test(BlantDataSet$WrightFPEFR,BlantDataSet$MiniWrightFPEFR)
#Используем функции пакета dplyr для добавления новых переменных в
# загруженный набор данных.
BlantDataSet %>%
# Добавляем переменную, содержащую разницу в измерениях двух
приборов
mutate(DiferFerst=WrightFPEFR-MiniWrightFPEFR) %>%
# Добавляем переменную, содержащую разницу во второй серии
измерений
mutate(DiferentSecond= WrightSPEFR-MiniWrightSPEFR)%>%
# Добавляем переменную, содержащую выборочное среднее между двумя
# сериями измерений и сохраняем результат в новый набор данных
mutate(meanLarge =(WrightFPEFR+WrightSPEFR)/2)-> NewBlantDataSet
# Вычисляем параметры линейной регрессионной модели
LineModel <-
lm(BlantDataSet$MiniWrightFPEFR~BlantDataSet$WrightFPEFR)
summary(LineModel) # Обобщаем результаты построения линейной
# регрессионной модели
# Строим график зависимости результатов измерений второго прибора
# от результатов измерений первого прибора и модель линейной
регрессии
corDataPlot <- ggplot()
corDataPlot <- corDataPlot+geom_point(mapping =
aes(BlantDataSet$WrightFPEFR,BlantDataSet$MiniWrightFPEFR,
colour="Исходные данные"), size=3)

```

Продолжение листинга 14

```

corDataPlot <- corDataPlot+geom_line(mapping =
aes(BlantDataSet$WrightFPEFR,
predict(LineModel),
colour="Линейная модель"),
size=3)
corDataPlot <- corDataPlot+theme_light()
corDataPlot <- corDataPlot+theme(panel.border = element_rect(linewidth = 6,
colour = "black"),
axis.title = element_text(size = 20, face = "bold",colour =
"black"),
axis.text.x = element_text(size = 20,face = "bold",colour =
"black"),
axis.text.y = element_text(size = 20, face = "bold",
colour = "black"),
axis.ticks = element_line(linewidth = 3, colour = "black"),
legend.title = element_text(size = 20, face = "bold",
colour = "black"),
legend.text = element_text(size = 20, face = "bold",
color = "black"))
corDataPlot <- corDataPlot+labs(x="Скорость потока по прибору 1",
y="Скорость потока по прибору 2", colour="Тун
данных: ")
print(corDataPlot)
# На основе подготовленных вычислений разниц и средних значений по
двум
# сериям измерений строим график Бланда – Альтмана
diffPlot <- ggplot()
diffPlot <- diffPlot + geom_point(mapping = aes(NewBlantDataSet$meanLarge,
NewBlantDataSet$DiferFerst),
colour="blue", size=3)
diffPlot <- diffPlot + geom_hline(yintercept =
mean(NewBlantDataSet$DiferFerst),
colour="green", linewidth=2)
diffPlot <- diffPlot +
geom_hline(yintercept = mean(NewBlantDataSet$DiferFerst)+
2*sd(NewBlantDataSet$DiferFerst),
colour="red", linewidth=2)
diffPlot <- diffPlot + geom_hline(yintercept =
mean(NewBlantDataSet$DiferFerst)-
2*sd(NewBlantDataSet$DiferFerst),
colour="red", linewidth=2)
diffPlot <- diffPlot+theme_light()
diffPlot <- diffPlot+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",

```

Продолжение листинга 14

```
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"))  
diffPlot <- diffPlot+labs(x="Среднее значение по прибору 1",  
y="Разница между приборами")  
print(diffPlot)
```

Результаты выполнения кода листинга 14 представлены на рисунках 19 и 20.

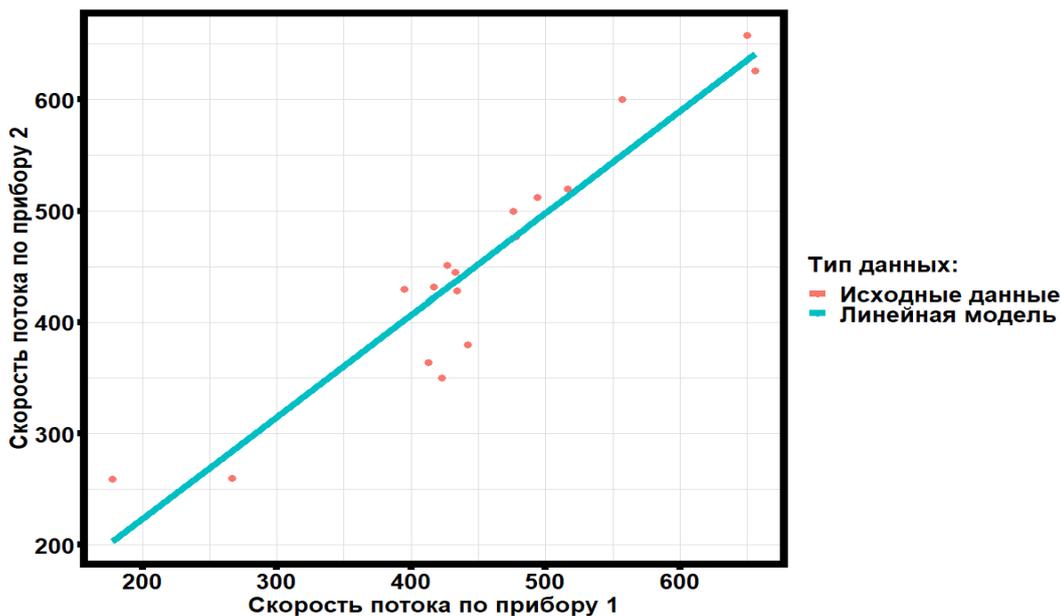


Рисунок 19 – Зависимость скорости выдыхаемого потока воздуха, замеренного посредством прибора 1, от скорости потока, замеренного посредством прибора 2 (коэффициент корреляции по Пирсону между измерениями составляет 0,94 с уровнем статистической значимости $p\text{-value} = 1,4 \times 10^{-08}$)

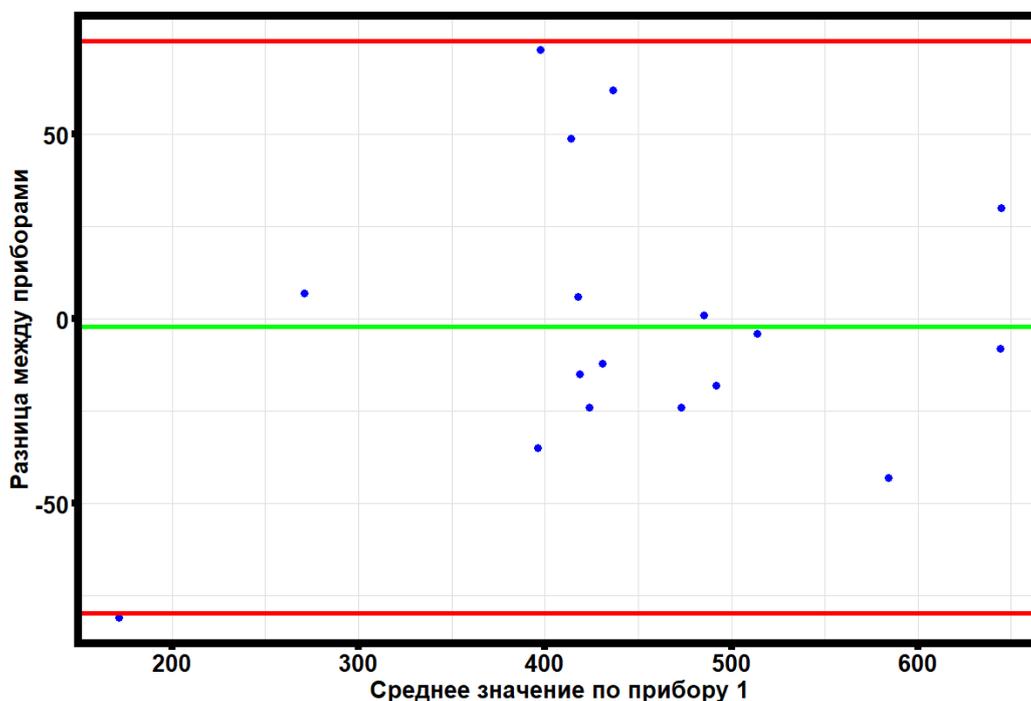


Рисунок 20 – Зависимость разницы измерений между двумя приборами по первой серии измерений от среднего значения скорости выдыхаемого потока воздуха, вычисленного по двум сериям измерений, проведенных с помощью первого прибора. Зеленая горизонтальная линия – среднее значение разницы измерений между двумя приборами. Верхняя и нижняя горизонтальные линии (обозначены красным цветом) – верхняя и нижняя границы доверительного интервала, соответственно

Соответственно, все значения разниц между измерениями, выходящие за пределы доверительного интервала, являются выбросами измерений. Допустимое количество выбросов составляет не более 5 %.

4. РЕГРЕССИОННЫЙ АНАЛИЗ. ЛИНЕЙНАЯ РЕГРЕССИЯ

Регрессионный анализ является одним из самых распространенных типов статистического анализа, позволяющих строить фенологические модели, описывающие изменение одних численных параметров в зависимости от других. В простом линейном однопараметрическом представлении моделью регрессионного анализа является уравнение, которое записывается, как (30)⁴¹:

$$a = \bar{y} - b\bar{x}, \quad (30)$$

где \bar{y} , \bar{x} – средние значения зависимой и независимой переменных; коэффициент b определяется по уравнению (31):

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (31)$$

Для многопараметрических моделей уравнение регрессии приобретает более сложную форму, но основной смысл остается тем же. В случае необходимости расширения объема знаний в области регрессионного анализа читатель может обратиться к работе [2].

На языке программирования R для построения линейных однопараметрических и многопараметрических моделей, а также нелинейных одно- и многопараметрических моделей используется команда **lm()** из пакета **stats**.

Пример проведения регрессионного анализа

Рассмотрим пример построения регрессионной модели на наборе данных **muscle**, содержащем длину сокращения сердечной мышцы мышцей и концентрацию хлорида кальция в растворе.

Листинг 15

```
library(MASS) # Подключаем пакет, содержащий набор данных muscle
library(ggplot2) # Подключаем пакет построения графиков
#Формируем линейную модель
modelLengthHeath <- lm(Length~Conc, data=muscle)
summary(modelLengthHead) # Выводим результаты вычисления модели
#=====
# Результаты построения регрессионной модели
#=====
Call:
lm(formula = Length ~ Conc, data = muscle)

Residuals:
    Min       1Q   Median       3Q      Max
-12.884   -4.097    1.060    4.487   10.064

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.5330     1.4229   9.511 1.93e-13 ***
Conc         5.4030     0.7653   7.060 2.32e-09 ***
```

⁴¹ Altman D. G. Practical statistics for medical research. CRC press, 1990.

Продолжение листинга 15

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.411 on 58 degrees of freedom  
Multiple R-squared:  0.4622,    Adjusted R-squared:  0.4529  
F-statistic: 49.85 on 1 and 58 DF, p-value: 2.322e-09
```

Коэффициенты a и b регрессионного уравнения (32) представлены в колонке Estimate соответственно, уравнение регрессии будет иметь вид:

$$y = 13.533 + 5.403 \cdot x \quad (32)$$

Коэффициент $\text{Pr}(> |t|)$ меньше заданного уровня статистической значимости, следовательно, все коэффициенты регрессионного уравнения статистически значимы. Физический смысл коэффициента a – средняя длина сердечной мышцы при отсутствии раствора хлорида кальция. А коэффициент b показывает, насколько увеличится длина сердечной мышцы при увеличении концентрации хлорида кальция на 1 единицу.

Residual standard error – среднее значение отклонений данных от линии регрессии.

Multiple R-squared – данная величина говорит о том, какой процент изменения длины сердечной мышцы связан с изменением концентрации хлорида кальция в растворе.

F-statistic и p-value – значения данных параметров говорят об общей значимости построенной модели. В рассмотренном случае построенная модель статистически значима.

Для графического представления результатов построения линейной регрессионной модели строят три графика:

1. Диаграмму рассеяния анализируемых данных с нанесенной линией регрессии.
2. Диаграмму остатков с нанесенным средним значением.
3. Квантиль-квантильный график для определения наиболее близкого типа распределения.

Представленный ниже код на языке R является продолжением листинга 15 и демонстрирует построение указанных графиков.

Продолжение листинга 15

```
modelPlot <- ggplot()  
modelPlot <- modelPlot + geom_point(mapping = aes(muscle$Conc, muscle$Length),  
  colour = "blue", size=3)  
modelPlot <- modelPlot + geom_line(mapping =  
  aes(muscle$Conc, predict(modelLengthHeath)),  
  colour = "red", linewidth=3)  
modelPlot <- modelPlot + labs(x = "Концентрация хлорида кальция",  
  y = "Длина сердечной мышцы").  
print(modelPlot)
```

Результат применения данной части кода листинга 15 представлен на рисунке 21.

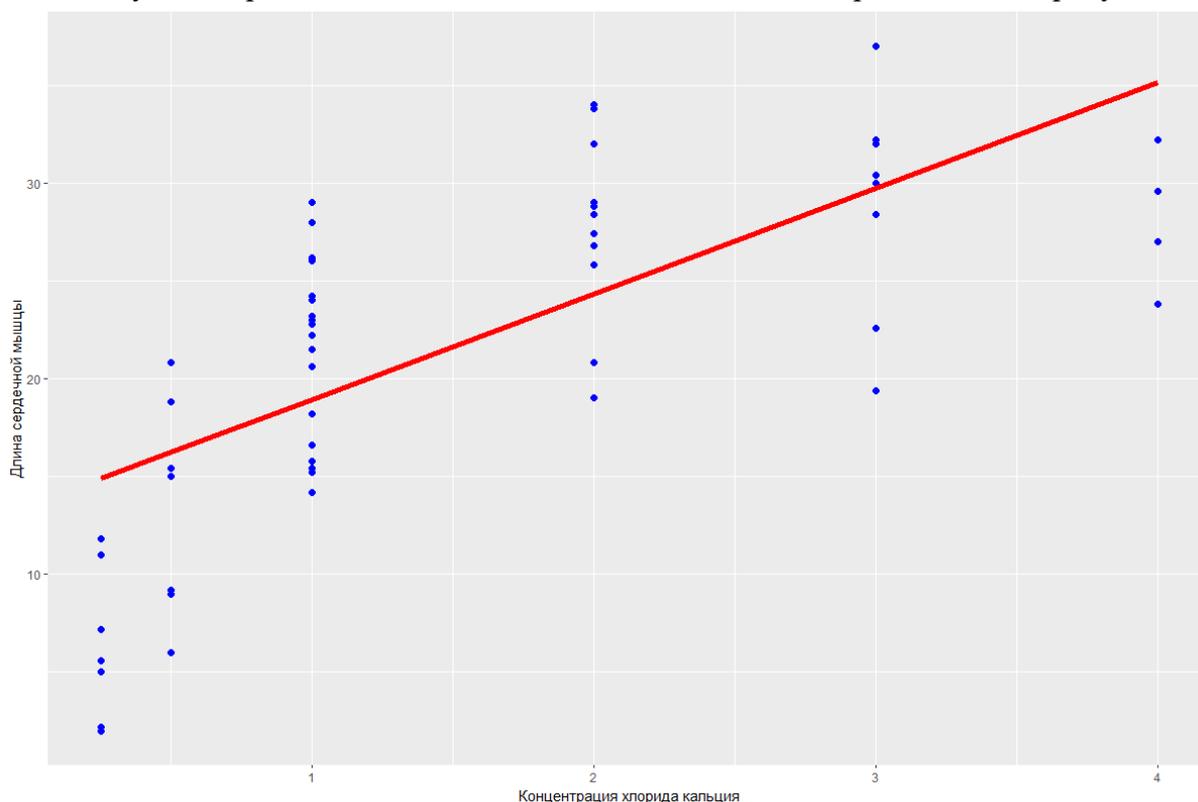


Рисунок 21 – Диаграмма рассеяния значений длины сердечной мышцы в зависимости от концентрации хлорида кальция. Отмечены синими точками на графике. Линейная регрессионная модель – красная линия на графике

Построение графика остатков с помощью библиотеки **ggplot2**.

Продолжение листинга 15

```
resPlot <- ggplot(mdoelLengthHeath, aes(x=.fitted, y=.resid))
resPlot <- resPlot+geom_point(colour = "blue", size=2)
resPlot <- resPlot+geom_hline(yintercept = 0, colour="red", size=2)
resPlot <- resPlot+labs(x = "Подознанные значения", y="Разницы значений")
print(resPlot)
```

Результат применения данной части кода листинга 15 представлен на рисунке 22.

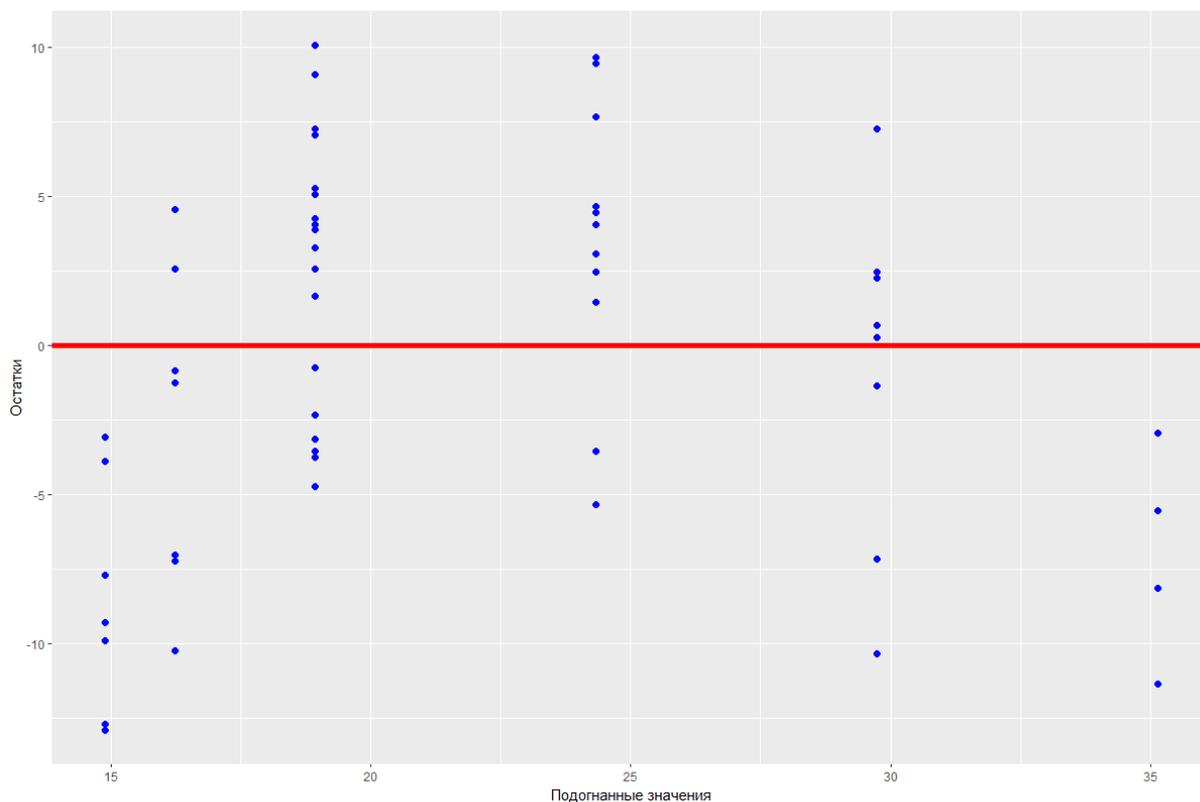


Рисунок 22 – Оценка остатков подобранной регрессионной модели

Оценка типа распределения данных может быть проведена по квантиль-квантильным графикам.

Неравномерность распределения остатков в зависимости от подобранных значений свидетельствует о не очень хорошем выборе модели. Окончательное решение о доработке линейной модели принимается на основании квантиль-квантильного (Q-Q) графика нормального распределения.

Продолжение листинга 15

```

qqPlot <- ggplot()
qqPlot <- qqPlot+geom_qq(aes(rstandard(mdoelLengthHeath)),
                           colour="blue", size=3))
qqPlot <- qqPlot + geom_abline(colour="red", size=2)
qqPlot <- qqPlot + labs(x = "Квантили нормального распределения",
                        y = "Квантили распределения остатков")
print(qqPlot)

```

Результат построения Q-Q графика представлен на рисунке 23.

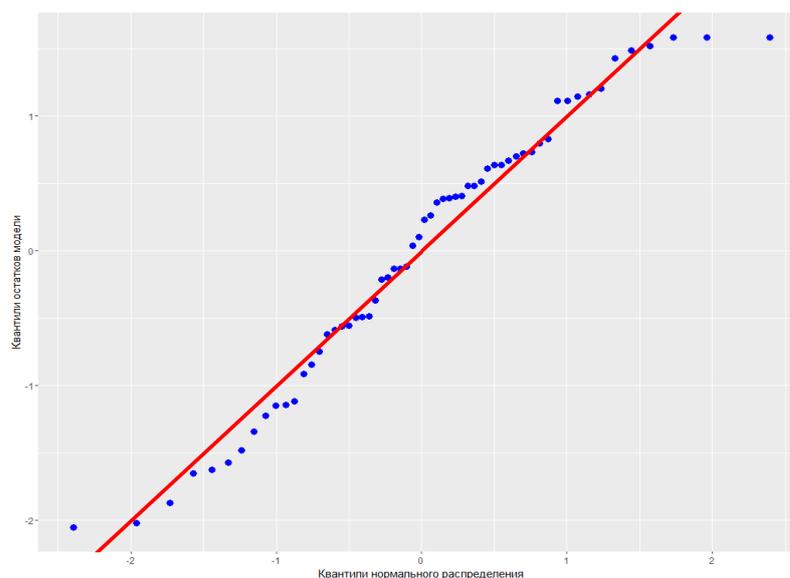


Рисунок 23 – Квантиль-квантильный график зависимости распределения квантилей остатков регрессионной модели от теоретических квантилей нормального закона распределения

Анализ квантиль-квантильного графика показывает, что наибольшее отклонение от нормального распределения остатков наблюдается на «хвостах» нормального распределения. Данное наблюдение является поводом для поиска альтернативной модели описания данных либо добора данных и повторного построения модели.

5. РЕГРЕССИОННЫЙ АНАЛИЗ. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Регрессионный анализ не ограничивается построением линейных моделей, связывающих между собой различные количественные переменные. В случае, если исследователю необходимо построить вероятностный прогноз на основе имеющихся у него данных, то возможно применение логистической регрессии, которая описывается уравнением (33):

$$f(z) = \frac{1}{1 + \exp^{42}(-z)}, \quad (33)$$

где $f(z)$ принимает значение 0 до 1, а z определяется из уравнения (34):

$$z = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_n \cdot x_n. \quad (34)$$

Функция независимых переменных $x_1 \dots x_n$, каждая из которых имеет свой вес $\theta_1 \dots \theta_n$.

Стоит отметить, что логистическая регрессия лежит в основе многих систем искусственного интеллекта (другое название функции 27 – перцептрон). На языке программирования R логистическая регрессия строится с помощью функции *glm()*, входящей в пакет **stats**.

Пример построения модели на основе логистической регрессии

Для примера построения и проверки модели логистической регрессии используем *Pima.tr*, *Pima.tr2* и *Pima.te*, которые входят в состав пакета **MASS** и содержат данные о популяции женщин в возрасте 21 года и старше, принадлежащих к индейскому племени Пима и проживающих в окрестностях Феникса (штат Аризона). Пациенты были обследованы на наличие сахарного диабета в соответствии с критериями Всемирной организации здравоохранения. Сбор данных осуществлялся Национальным институтом диабета и болезней органов пищеварения и почек США. Набор данных содержит 632 полные записи по величине инсулина в сыворотке крови.

При построении модели логистической регрессии необходимо провести оценку данных на предмет наличия корреляционных связей между независимыми переменными и различиями в данных, относящихся к одному из классов зависимой переменной – 0 или 1.

Листинг 16

```
library(MASS) #Подключаем пакет, содержащий набор данных Pima.tr2  
#Исследуем данные на наличие статистически значимых корреляционных  
# связей  
#=====  
# Исследуем данные на предмет структуры и наполнения  
#=====  
str(Pima.tr2)  
#=====  
# Вывод результата
```

⁴² \exp – в уравнениях означает экспоненту, математическую константу, приблизительно равную 2,718. Альтернативное название – число Эйлера.

Продолжение листинга 16

```
#=====
'data.frame':   200 obs. of  8 variables:
 $ npreg: int   5  7  5  0  0  5  3  1  3  2  ...
 $ glu   : int  86 195 77 165 107 97 83 193 142 128  ...
 $ bp    : int  68 70 82 76 60 76 58 50 80 78  ...
 $ skin  : int  28 33 41 43 25 27 31 16 15 37  ...
 $ bmi   : num  30.2 25.1 35.8 47.9 26.4 35.6 34.3 25.9 32.4 43.3  ...
 $ ped   : num   0.364 0.163 0.156 0.259 0.133  ...
 $ age   : int   24 55 35 26 23 52 25 24 63 31  ...
 $ type  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 1 1 2  ...
#=====
# Из имеющихся данных исключаем колонку, содержащую факторные данные
# проверяем данные и вычисляем матрицу корреляций по Пирсону43
#=====
cor(Pima.tr2)
#=====
# Результаты вычисления матрицы корреляции
#=====
      npreg      glu      bp skin bmi      ped      age
npreg 1.0000000 0.14758579 NA  NA NA  NA -0.09618250  0.55743415
glu   0.1475858 1.00000000 NA  NA NA  NA  0.09186663  0.28401844
bp    NA        NA        NA  1  NA NA  NA        NA        NA
skin  NA        NA        NA  NA  1  NA NA        NA        NA
bmi   NA        NA        NA  NA  NA  1  NA        NA        NA
ped   -0.0961825 0.09186663 NA  NA NA  NA  1.00000000 -0.05226369
age   0.5574341 0.28401844 NA  NA NA  NA -0.05226369  1.00000000
#=====
# Результаты вычислений показывают, что корреляция присутствует между
# количеством родов и возрастом исследуемого
#=====
# Проверяем уровень статистической значимости корреляции по Пирсону
# между исследуемыми параметрами
#=====
cor.test(Pima.tr2$npreg,Pima.tr2$age)
#=====
# Результаты теста
#=====
Pearson's product-moment correlation

data: Pima.tr2$npreg and Pima.tr2$age
t = 11.591, df = 298, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4741219 0.6308528
sample estimates:
 cor
0.5574341
#=====
# Результат теста показывает статистическую значимость корреляции
# между возрастом и количеством родов. Исключаем столбец с количеством
# родов из анализа
#=====
# Используя t-тест Стьюдента44 для независимых выборок, проверяем наличие
```

⁴³ В данном случае пример является демонстрационным, в качестве дополнительной задачи читатель может самостоятельно воспользоваться алгоритмом, приведенным на рисунке 2 первой части методических рекомендаций.

Продолжение листинга 16

```
# статистически значимых различий между группами с диабетом и без
#=====
t.test(bmi~type,data=Pima.tr2)
#=====
# Результат применения теста Стьюдента для независимых выборок
#=====
                                welch Two Sample t-test

data:  bmi by type
t = -5.4141, df = 243.68, p-value = 1.474e-07
alternative hypothesis: true difference in means between group No and
group Yes is not equal to 0
95 percent confidence interval:
 -5.331380 -2.486953
sample estimates:
 mean in group No mean in group Yes
      30.67083      34.58000
#=====
# p-value < 0,05 различия в группах статистически значимы
#=====
# Вычисляем параметры логистической регрессии
#=====
glmPima2 <- glm(type~bmi, data = Pima.tr2, family = "binomial")
summary(glmPima2)
#=====
#Результат построения модели логистической регрессии
#=====
Call:
glm(formula = type ~ bmi, family = "binomial", data = Pima.tr2)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.82024    0.69527  -5.495 3.92e-08 ***
bmi           0.09871    0.02070   4.769 1.85e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 385.87  on 296  degrees of freedom
Residual deviance: 360.55  on 295  degrees of freedom
(3 пропущенных наблюдений удалены)
AIC: 364.55

Number of Fisher Scoring iterations: 4
#=====
# Строим график биномиальной классификации
#=====
bmiGrid <- 18:55 #Новый интервал изменения индекса массы тела
newBMI <- data.frame(bmi = bmiGrid) # Формируем срез данных
PimaModel <- predict(glmPima2, newdata=newBMI, type="response")
plot(bmiGrid, PimaModel, ylim = c(-0,05, 1,05), t="l",
```

⁴⁴ Применение теста Стьюдента приведено в учебных целях. На практике, прежде чем проводить какой-либо тест, необходимо выполнить проверку на соответствие нормальному закону распределения и на основании результатов этой проверки следовать алгоритму, изображенному на рисунке 2 первой части методических рекомендаций.

Продолжение листинга 16

```
main = "Логистическая модель риска диабета",
xlab = "Индекс массы тела",
ylab = "Вероятность возникновения диабета")
stripchart(Pima.tr2$bmi[Pima.tr2$type == "No"], at = -0.05, add = T,
  method = "jitter", jitter = 0.02, col = "blue")
stripchart(Pima.tr2$bmi[Pima.tr2$type == "Yes"], at = 1.05, add = T,
  method = "jitter", jitter = 0.02, col = "red")
```

Результат построения логистической регрессионной модели риска возникновения диабета у женщин племени Пима в зависимости от индекса массы тела представлен на рисунке 24.

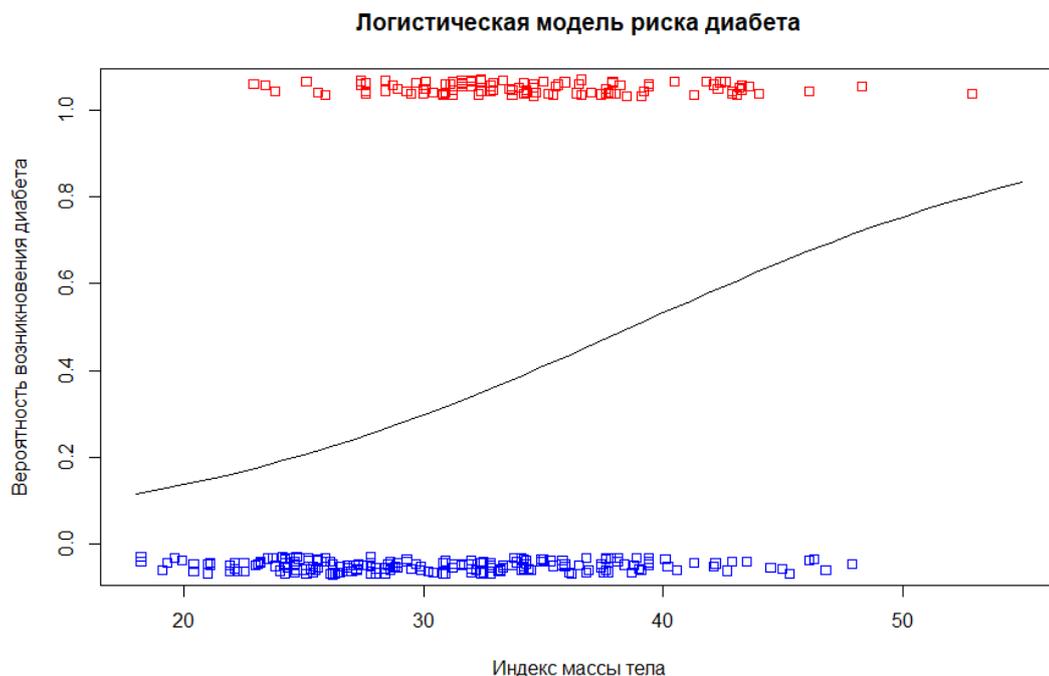


Рисунок 24 – Логистическая модель, описывающая риск возникновения диабета в зависимости от индекса массы тела. Синими квадратами обозначены пациенты, не имеющие сахарного диабета. Красными квадратами обозначены пациенты с подтвержденным диагнозом «сахарный диабет». Черной кривой обозначена вероятность возникновения сахарного диабета в зависимости от индекса массы тела

Анализ результатов моделирования показывает, что максимальный риск развития диабета возникает у пациентов с индексом массы тела от 28 до 45, т.к. в данной области наблюдается наибольшее количество пациентов с подтвержденным диагнозом «сахарный диабет» и наименьшее – с отсутствием данного диагноза.

5.1. Оценка моделей, построенных на основе логистической регрессии

Процесс оценки моделей, построенных на основе логистической регрессии, можно разделить на несколько подзадач:

1. Оценка корректности классификации (ошибка классификации).
2. Оценка диагностической точности классификатора.

3. Оценка построенной модели на альтернативном (тестовом) наборе данных, содержащим ту же патологию.
 4. Оценка моделей логистической регрессии при введении дополнительных признаков.
- Для проведения всех вышеперечисленных оценок применяются различные методы сравнительного анализа и статистические критерии оценки сервисов. Рассмотрим каждую из подзадач на примерах⁴⁵.

5.2. Оценка корректности классификации

Оценка корректности классификации производится с использованием так называемой «матрицы ошибок»⁴⁶ и формируется на основе сравнительного анализа между результатами классификации, построенной с помощью логистической регрессии, и результатами классификации, представленной в наборе данных⁴⁷. Наглядное представление данного сравнения приведено в таблице 14.

Таблица 14 – Наглядное представление сравнительного анализа и способа классификации ошибок и истинных значений при проведении оценки модели, построенной на основе логистической регрессии

| Тип диагноза/№ пациента | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--|----|----|----|----|----|----|----|----|----|
| Подтвержденный диагноз | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| Диагноз по модели | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| Условное обозначение | FN | TN | TN | TP | TN | TP | FP | FN | TP |
| <p><i>Примечание. FN (False negative) – случай ложноотрицательной постановки диагноза; TN (True negative) – случай истинно отрицательной постановки диагноза; TP (True positive) – случай истинно положительной постановки диагноза; FP (False positive) – случай ложноположительной постановки диагноза</i></p> | | | | | | | | | |

Используя данные из таблицы 14, можно преобразовать результаты классификации в вид, представленный в таблице 15.

Таблица 15 – Матрица ошибок, составленная на основании сравнительного анализа модели и подтвержденного диагноза пациента

| | | Диагноз, поставленный врачом | | Итого: |
|--|-----|------------------------------|-----|--------|
| | | Нет | Да | |
| Результаты моделирования | Нет | A | B | A+B |
| | Да | C | D | C+D |
| Итого: | | A+C | B+D | N |
| <p><i>Примечание. A – общее количество TN; B – общее количество FN; C – общее количество FP; D – общее количество TP</i></p> | | | | |

⁴⁵ Методы оценки и статистического анализа моделей логистической регрессии не ограничиваются приведенными в настоящих методических рекомендациях, но являются наиболее часто применяемыми на практике.

⁴⁶ Данное название является переводом с английского языка термина «confusion matrix». Дословный перевод – «матрица путаницы» или «матрица соответствия».

⁴⁷ Ground truth – диагноз, подтвержденный любым методом исследований с высокой степенью доверия.

На основе матрицы ошибок вычисляются три основные операционные характеристики модели:

1. Чувствительность модели вычисляется по уравнению (35):

$$Sensitivity = \frac{A}{A + B}. \quad (35)$$

2. Специфичность модели вычисляется по уравнению (36):

$$Specificity = \frac{D}{D + C}. \quad (36)$$

3. Точность модели вычисляется по уравнению (37):

$$Accuracy = \frac{A + C}{N}. \quad (37)$$

Алгоритм матрицы ошибок может быть применен для анализа нулевых гипотез⁴⁸.

Сложность применения данного подхода заключается в том, что исследователю данных неизвестно, какое распределение данных является истинным, и возможна только оценка наиболее близкого теоретического распределения. А в случае, когда гипотеза является сложной, данный метод оценки становится практически неприменим на практике.

Таким образом, единственное, что может быть применено – это проверка данных на принадлежность тому или иному теоретическому типу распределения в соответствии с различными критериями и их мощностью.

Рассмотрим построение матрицы ошибок на примере модели логистической регрессии пациентов с диагнозом «диабет».

Пример построения матрицы ошибок бинарной классификации

Данный пример является продолжением листинга 16, представленного в примере построения модели на основе логистической регрессии. На основе набора данных *Pima.tr2*, содержащегося в пакете **MASS**, построена модель логистической регрессии возникновения диабета у женщин племени Пима.

Продолжение листинга 16

```
#=====
# Строим модель на основании логистической регрессии
#=====
predGlm <- predict(glmPima2, type = "response")
#=====
# Разделяем модель на две части, если вероятность по модели меньше 0,5
# то норма – 0, если выше 0,5 то патология – 1
#=====
typePred <- ifelse(predGlm < 0,5, 0, 1)
#=====
#Формируем матрицу ошибок
```

⁴⁸ Майкл, Х. [и др.]. Статистика и планирование эксперимента для непосвященных: как отучить статистику лгать. М.: ДМК Пресс, 2023.

Продолжение листинга 16

```
#=====
table(typePred[1:length(Pima.tr2$type)], Pima.tr2$type)
#=====
# Матрица ошибок построенной модели
#=====
      No Yes
0 166  85
1   26  20
#=====
# Вычисление чувствительности, специфичности и точности модели
#=====
sensData <- confMatrix[1]/(confMatrix[1]+confMatrix[3]) #Чувствительность
print(sensData)
#=====
# Результат вычисления специфичности
#=====
[1] 0.6613546
#=====
specData <- confMatrix[4]/(confMatrix[4]+confMatrix[2]) #Специфичность
print(specData)
#=====
# Результат вычисления чувствительности
#=====
[1] 0.4347826
#=====
      accurData <- (confMatrix[1]+confMatrix[4])/(confMatrix[4]+confMatrix[3]+
confMatrix[2]+confMatrix[1]) #Точность
print(accurData)
#=====
# Результат вычисления точности
#=====
[1] 0.6262626
```

Анализ результатов построения матрицы ошибок модели бинарной классификации на основе логистической регрессии показывает, что доля правильно классифицированных патологий (единиц) составляет 66 % (чувствительность), тогда как доля правильно классифицированных норм (нулей) – 43 % (специфичность)⁴⁹. В теории построения классификатор считается идеальным тогда и только тогда, когда чувствительность и специфичность равна 1. Однако на практике при построении моделей и дальнейшем прогнозировании модель обладает или высокой чувствительностью, или высокой специфичностью. Поведение чувствительности и специфичности построенной модели зависит от порога принятия решения, выбор порога является достаточно творческим процессом, но может быть статистически обоснован за счет применения различных тестов.

⁴⁹ Не существует единого способа интерпретации полученных оценок показателей как хороших или плохих, но в целом можно считать, что показатели, превышающие 95 %, являются очень хорошими, от 85 до 95 % – хорошими, более низкие величины – плохими.

Оценка классификации и выбор порога принятия могут быть проведены с помощью ряда статистических тестов, таких как тест Макнемара, индекс Юдена, или на основе оценки метрик диагностической точности (ROC-AUC⁵⁰).

5.3. Тест Макнемара

При исследовании данных могут возникать задачи, требующие установления статистической значимости в различиях между двумя коррелирующими пропорциями или проведения статистического анализа матрицы ошибок классификации логистической регрессии. Для подтверждения или опровержения статистической значимости между коррелирующими пропорциями служит тест Макнемара⁵¹. Значение теста Макнемара вычисляется по уравнению (38):

$$\chi^2 = \frac{(D - A)^2}{(D + A)}, \quad (38)$$

где D и A – количество исследований, содержащееся в матрице ошибок следующего вида (таблица 16).

Таблица 16 – Матрица ошибок классификации модели бинарной классификации на основе логистической регрессии

| | | Результаты теста 2 | | Итого в строках |
|--------------------|---------------|--------------------|---------------|-----------------|
| | | Отрицательный | Положительный | |
| Результаты теста 1 | Положительный | A | B | A+B |
| | Отрицательный | C | D | C+D |
| Итого в столбцах: | | A+C | B+D | A+B+C+D |

Тест Макнемара имеет ограничения по применению и неприменим в случае, если $A+D < 10^{25}$.

Применение теста Макнемара не ограничивается оценкой результатов бинарной классификации: он может быть использован для оценки статистической значимости в результатах исследований, проведенных посредством двух различных тестов.

На языке программирования R тест Макнемара можно провести посредством функции `mcnemar.test()` библиотеки `stats`.

Пример применения теста Макнемара

Рассмотрим применение теста Макнемара для оценки статистически значимых различий в бинарной классификации при постановке диагноза врачом и моделью, построенной на основе логистической регрессии. Для построения модели используем информацию, содержащуюся в наборе данных *biopsy* пакета **MASS**.

⁵⁰ ROC-AUC – площадь под кривой «рабочих характеристик приемника». Более подробно она рассмотрена в подразделе 5.5.

⁵¹ McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages // Psychometrika. 1947. Vol. 12, №2. P. 153–157.

Листинг 17

```

library(MASS) #Подключаем пакет, содержащий набор данных biopsy
# Исследуем структуру данных, содержащихся в biopsy
str(biopsy)
#=====
# Результаты применения функции
#=====
'data.frame':   699 obs. of  11 variables:
 $ ID   : chr  "1000025" "1002945" "1015425" "1016277" ...
 $ V1   : int   5 5 3 6 4 8 1 2 2 4 ...
 $ V2   : int   1 4 1 8 1 10 1 1 1 2 ...
 $ V3   : int   1 4 1 8 1 10 1 2 1 1 ...
 $ V4   : int   1 5 1 1 3 8 1 1 1 1 ...
 $ V5   : int   2 7 2 3 2 7 2 2 2 2 ...
 $ V6   : int   1 10 2 4 1 10 10 1 1 1 ...
 $ V7   : int   3 3 3 3 3 9 3 3 1 2 ...
 $ V8   : int   1 2 1 7 1 7 1 1 1 1 ...
 $ V9   : int   1 1 1 1 1 1 1 1 5 1 ...
 $ class: Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1
#=====
#Фактор, содержащий бинарные данные class, ID – идентификатор
пациента
# Для количественных данных проводим анализ корреляции
#=====
cor(biopsy[,c(1,11)]) # Строим корреляционную матрицу
#=====
#Результат построения корреляционной матрицы52
#=====

```

| | v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 |
|-----------|-----------|-----------|-----------|-----------|-----------|----|-----------|-----------|
| v9 | 1.0000000 | 0.6449125 | 0.6545891 | 0.4863562 | 0.5218162 | NA | 0.5584282 | 0.5358345 |
| 0.3500339 | | 1.0000000 | 0.9068819 | 0.7055818 | 0.7517991 | NA | 0.7557210 | 0.7228648 |
| 0.4586931 | | | 1.0000000 | 0.6830792 | 0.7196684 | NA | 0.7359485 | 0.7194463 |
| 0.4389109 | | | | 1.0000000 | 0.5995991 | NA | 0.6667153 | 0.6033524 |
| 0.4176328 | | | | | 1.0000000 | NA | 0.6161018 | 0.6288807 |
| 0.4791015 | | | | | | 1 | NA | NA |
| NA | | | | | | | 1.0000000 | 0.6658778 |
| 0.3441695 | | | | | | | | 1.0000000 |
| 0.4283357 | | | | | | | | |
| 1.0000000 | | | | | | | | |

```

#=====
# Анализ корреляционной матрицы показывает, что для построения модели
# бинарной классификации на основе логистической регрессии могут быть
# использованы единственные независимые переменные V9 и V6, т.к. между
# ними корреляция отсутствует или крайне низкая
#=====

```

⁵² При построении регрессионных моделей существуют две основные стратегии. Первая заключается в учете только сильно коррелирующих переменных (данная стратегия используется при построении линейных и нелинейных регрессионных моделей). Вторая стратегия – это учет только слабо коррелирующих параметров (стратегия, применяющаяся при построении моделей классифицирующих моделей). Каждая из стратегий основана на математических особенностях вычислений коэффициентов регрессии, рассмотрение которых выходит за рамки настоящих рекомендаций.

Продолжение листинга 17

```
#=====
# Строим модель логистической регрессии
#=====
glmBiopsy <- glm(class~V6+V9, data = biopsy, family= "binomial")
#=====
# Строим модель
#=====
predGlm <- predict(glmBiopsy, type="response")
#=====
typePred <- ifelse(predGlm < 0.5, 0, 1) # Отсечение нормы и патологии
# Построение матрицы ошибок
confMatrix <- table(typePred[1:length(biopsy$class)],biopsy$class)
# Выводим матрицу ошибок
print(confMatrix)
#=====
#Результаты построения матрицы ошибок
#=====
      benign malignant
0      313         146
1      133          91
#=====
# Рассчитываем чувствительность, специфичность и точность модели
#=====
sensData <- confMatrix[1]/(confMatrix[1]+confMatrix[3]) #Чувствительность
print(sensData)
#=====
#Результат выведения чувствительности
#=====
[1] 0.6819172
#=====
specData <- confMatrix[4]/(confMatrix[4]+confMatrix[2]) #Специфичность
print(specData)
#=====
# Результат вычисления специфичности
#=====
[1] 0.40625
#=====
accurData <- (confMatrix[1]+confMatrix[4])/(confMatrix[4]+confMatrix[3]+
confMatrix[2]+confMatrix[1]) #Точность
print(accurData)
#=====
# Результат вычисления точности
#=====
[1] 0.5915081
#=====
mcnemar.test(confMatrix)
#=====
#Результаты применения теста Макнемара
#=====
McNemar's Chi-squared test with continuity correction

data:  confMatrix
McNemar's chi-squared = 0.51613, df = 1, p-value = 0.4725
```

Применение теста Макнемара показывает, что доля верно классифицированных результатов биопсии с помощью построенной модели логистической регрессии статистически незначимо ($p\text{-value} = 0,4725 > 0,05$) отличается от результатов постановки диагноза врачом. Однако чувствительность составляет 68 %, специфичность – 40 %, а точность модели – 59 %. Соответственно, полученная модель требует дополнительной настройки с целью снижения статистически значимых различий (увеличение $p\text{-value}$) между решением, принятым врачом, и результатами прогнозирования модели.

5.4. Индекс Юдена⁵³

Индекс Юдена является альтернативой теста Макнемара. Основная задача, решаемая с помощью индекса Юдена, – выбор самого чувствительного диагностического теста, который наиболее точно сортирует пациентов по признаку «норма – патология». Вычисление индекса Юдена основано на матрице ошибок, приведенной в таблице 17.

Таблица 17 – Матрица ошибок для вычисления индекса Юдена

| | | Результаты теста 2 | | Итого в строках |
|--------------------|---------------|--------------------|---------------|-----------------|
| | | Отрицательный | Положительный | |
| Результаты теста 1 | Положительный | A | B | A+B |
| | Отрицательный | C | D | C+D |

Используя данные таблицы 17, индекс Юдена рассчитывается по уравнению (39):

$$J = \frac{A}{A+B} + \frac{D}{C+D} - 1, \quad (39)$$

где A – общее количество TN; B – общее количество FN; C – общее количество FP; D – общее количество TP.

Стандартное отклонение для индекса Юдена вычисляется по уравнению (40):

$$S_{eJ} = \sqrt{\frac{A \cdot B}{(A+B)^3} + \frac{C \cdot D}{(C+D)^3}}. \quad (40)$$

Оценка эффекта от изменения параметров диагностических тестов в оригинальной работе Юдена проводится с помощью t-теста Стьюдента.

В представленном ниже примере разбирается применение индекса Юдена для выбора порога принятия решения в модели бинарной классификации, построенной на основе логистической регрессии.

Пример применения индекса Юдена для выбора оптимального порога принятия решения

Используя набор данных, содержащих исследование женщин племени Пима на наличие диабета, проведем оптимизацию порога принятия решения модели логистической регрессии на основании поведения индекса Юдена в зависимости от порога принятия решения и выполним сравнительный анализ статистической значимости выявленных различий.

⁵³ Youden W. J. Index for rating diagnostic tests // Cancer. 1950. Vol. 3, №1. P. 32–35.

Листинг 18

```

library(MASS) #Подключаем пакет, содержащий набор данных biopsy
library(ggplot2) # Подключаем пакет построения графиков
library(dplyr) # Подключаем библиотеку работы с большими массивами
данных
#Строим модель логистической регрессии
glmBiopsy <- glm(class~V6+V9, data = biopsy, family = "binomial")
predGlm <- predict (glmBiopsy, type = "response")
Index <- c() #Вектор для хранения индекса принятия решения
Juden <- c() #Вектор для хранения значений индекса Юдена
seJuden <- c() # Вектор для хранения значений ошибки индекса Юдена
Specificity <- c() # Вектор для хранения значений Специфичности модели
Sensivity <- c() # Вектор для хранения значений Чувствительности модели
Accuracy <- c() # Вектор для хранения значений Точности модели
Macnemar <- c() # Вектор для хранения значений уровня статистической
значимости
# теста Макнемара
#=====
# Цикл перебора значений порога принятия решения
#=====
for(stepIndex in seq(0,05;0,99;0,01)){
#=====
typePred <- ifelse(predGlm < stepIndex, 0, 1) # Классифицируем результаты
#расчета вероятности заболевания в зависимости от значений порога
#=====
confMatrix <- table(typePred[1:length(biopsy$class)],biopsy$class) #Вычисляем
# матрицу ошибок
#=====
# Вычисляем индекс Юдена
J <- (confMatrix[1]/(confMatrix[1]+confMatrix[3]))+
(confMatrix[4]/(confMatrix[4]+confMatrix[2]))-1
# Вычисляем среднюю квадратическую ошибку индекса Юдена
De <- ((confMatrix[1]*confMatrix[3])/((confMatrix[1]+confMatrix[3])^3))+
((confMatrix[4]*confMatrix[2])/((confMatrix[4]+confMatrix[2])^3))
Se <- sqrt(De)
# Проводим тест Макнемара для каждой вычисленной матрицы ошибок
macTest<-mcnemar.test(confMatrix)
# Вычисляем чувствительность модели логистической регрессии
sensData<- confMatrix[1]/(confMatrix[1]+confMatrix[3])
# Вычисляем специфичность модели логистической регрессии
specData <- (confMatrix[4]/(confMatrix[4]+confMatrix[2]))
# Вычисляем Точность построенной модели
accurData <- (confMatrix[1]+confMatrix[4]/(confMatrix[4]+confMatrix[3]+
confMatrix[2]+confMatrix[1])
# Сохраняем результаты вычислений в вектора
Index <- c(Index, stepIndex)
Juden <- c(Juden, J)
seJuden <- c(seJuden, Se)
Macnemar <- c(Macnemar,macTest$p.value)
Specificity <- c(Specificity, specData)

```

Продолжение листинга 18

```
Sensivity <- c(Sensivity, sensData)
Accurancy <- c(Accurancy, accurData)
}
# Формируем Фрейм данных, содержащий результаты всех вычислений.
resultData <- data.frame(Index, Juden, seJuden, Specifity, Sensivity, Accurancy,
Macnemar)
#=====
#Создаем дополнительные колонки с верхним и нижним значением
# доверительного интервала индекса Юдена
resultData %>%
  mutate(JudenUp=Juden+seJuden) %>%
  mutate(JudenDown=Juden-seJuden) -> judenData
#=====
# Строим графическое представление результатов вычисления зависимости
# индекса Юдена от порога принятия решения
#=====
# Модели, аппроксимирующие результаты вычисления индекса Юдена,
# Верхнюю и нижнюю границу доверительного интервала
#=====
modelJuden <- lm(judenData$Juden~splines::bs(judenData$Index,5))
modelJudenUp <- lm(judenData$JudenUp~splines::bs(judenData$Index,5))
modelJudenDown <- lm(judenData$JudenDown~splines::bs(judenData$Index,5))
#=====
plotJuden <- ggplot()
plotJuden <- plotJuden +
  geom_point(mapping = aes(judenData$Index,judenData$Juden),
  colour="red", size=2)
plotJuden <- plotJuden +
  geom_line(mapping=aes(judenData$Index,predict(modelJuden)),
  colour="blue", linewidth=2)
plotJuden <- plotJuden +
  geom_line(mapping=aes(judenData$Index,predict(modelJudenUp)),
  colour="grey", linewidth=2)
plotJuden <- plotJuden +
  geom_line(mapping=aes(judenData$Index,predict(modelJudenDown)),
  colour="grey", linewidth=2)
plotJuden <- plotJuden+labs(x="Порог принятия решения", y="Индекс Юдена")
print(plotJuden)
```

Результаты построения зависимости индекса Юдена от порога принятия решения представлены на рисунке 25.

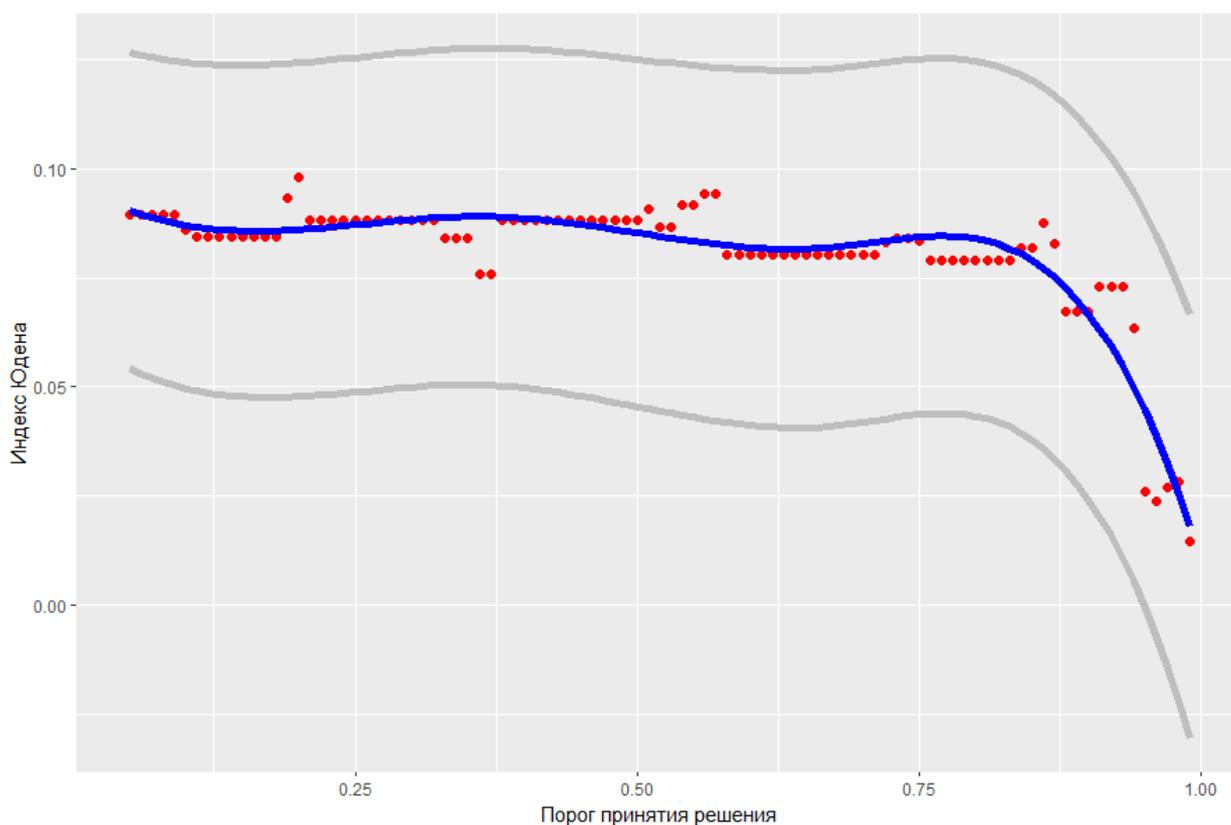


Рисунок 25 – Зависимость индекса Юдена от порога принятия решения в бинарной классификации, построенной на основе логистической регрессии

Из определения индекса Юдена следует, что чем больше значение индекса Юдена, тем более точно сортирует исследуемая модель пациентов по признаку «норма – патология». Анализ зависимости индекса Юдена от порога принятия решения (рисунок 22) показывает, что максимальное значение индекса достигается при значении порога принятия решения, равного 0,57. В данной точке тест Макнемара показывает статистическую значимость, равную $p\text{-value} = 0,306$, что свидетельствует об отсутствии статистически значимых различий между диагнозом, поставленным врачом, и диагнозом, полученным на основании модели. При этом чувствительность модели равна 0,68, специфичность – 0,41 и точность – 0,60.

5.5. Метрика диагностической точности ROC-AUC⁵⁴

Определение метрики диагностической точности начинается с построения так называемой «рабочей характеристики приемника» (ROC⁵⁵). Фактически это зависимость всех правильно идентифицированных положительных результатов от всех правильно идентифицированных отрицательных результатов для всего диапазона значений точки принятия решения бинарной классификации. Рассмотрим построение ROC-кривой на примере модели бинарной классификации пациентов с подозрением на сахарный диабет.

⁵⁴ ROC-анализ – частный анализ связей признаков (количественного и бинарного). Receiver Operating Characteristic curve или, сокращенно, ROC (receiver operating characteristics) – рабочая характеристика приемника, AUC (area under the curve) – площадь под кривой ROC.

⁵⁵ Оценка зрелости технологий искусственного интеллекта для здравоохранения: методические рекомендации / сост. Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». Вып. 123. М: ГБУЗ «НПКЦ ДиТ ДЗМ», 2023. 28 с.

Пример построения ROC-кривой для модели логистической регрессии

Используя набор данных **Pima.tr2** из пакета **MASS**, содержащих результаты обследования женщин племени Пима на наличие диабета, построим модель логистической регрессии, вычисляющей вероятность наличия патологии и проводящей бинарную классификацию пациентов на основе вычисленной вероятности. Далее построим рабочую характеристику модели и проведем анализ построенной модели на тестовом наборе данных.

Листинг 19

```
library(MASS) #Подключаем пакет, содержащий набор данных Pima.tr2
# графиков
library(ROCR) #Подключаем пакет, содержащий функции построения ROC-
# кривой.
# Строим модель логистической регрессии для одной переменной
glmPima2_1 <- glm(type~bmi, data = Pima.tr2, family = "binomial")
predGlm_1 <- predict (glmPima2_1, type = "response")
# Строим модель логистической регрессии для двух переменных
glmPima2_2 <- glm(type~bmi+glu, data=Pima.tr2, family = "binomial")
predGlm_2 <- predict (glmPima2_2, type = "response")
#=====
# Готовим модель для построения ROC-кривой
#=====
#Однопараметрическая модель
rocGlm_1 <- prediction(predGlm_1, labels = Pima.tr2$type [1:length(predGlm_1)])
rocPerf_1 <- performance(rocGlm_1, measure = "tpr", x.measure = "fpr")
plot(rocPerf_1, col="skyblue")
#Двухпараметрическая модель
rocGlm_2 <- prediction(predGlm_2, labels = Pima.tr2$type [1:length(predGlm_2)])
rocPerf_2 <- performance(rocGlm_2, measure = "tpr", x.measure = "fpr")
plot(rocPerf_2, , col = "darkblue", add = TRUE)
#=====
#Построение графика ROC-кривой
#=====
abline(a = 0, b = 1, lty = 2)
legend("bottomright", leg = c("bmi", "bmi+glu"),
col = c("skyblue", "darkblue"), lwd = 1,5)
#=====
# Вычисляем AUC однопараметрической модели
#=====
aucGlm_1 <- performance(rocGlm_1, measure = "auc")
print(aucGlm_1@y.values[[1]])
#=====
# Вычисляем AUC двухпараметрической модели
#=====
aucGlm_2 <- performance(rocGlm_2, measure = "auc")
print(aucGlm_2@y.values[[1]])
```

Результаты построения ROC-кривой представлены на рисунке 26.

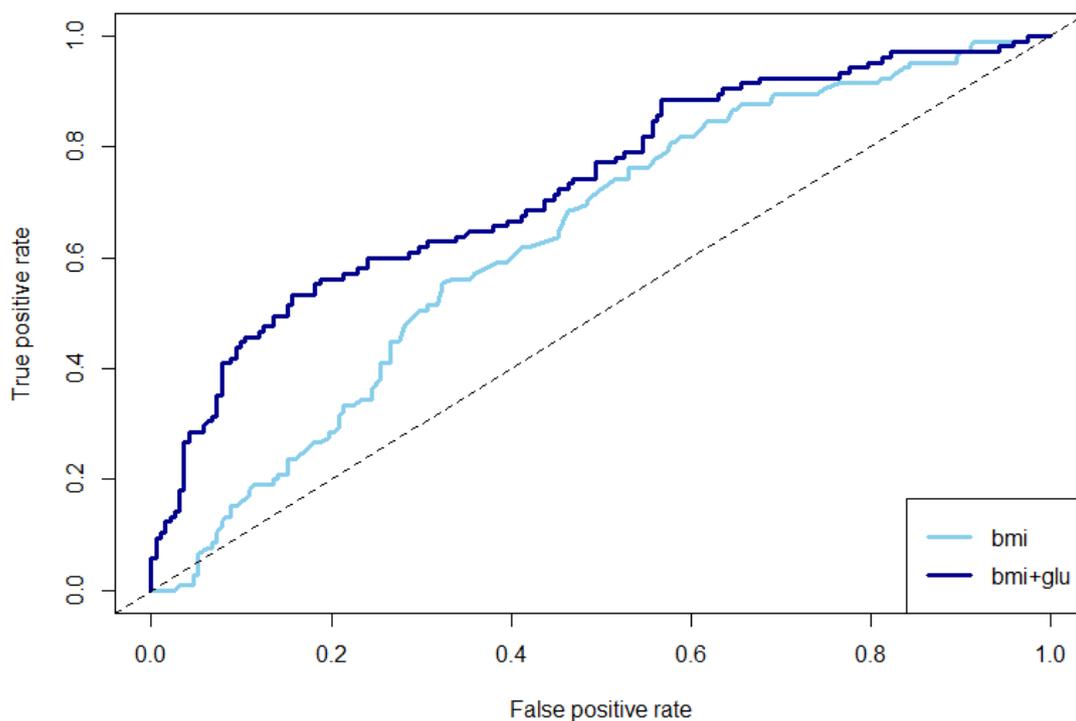


Рисунок 26 – ROC-кривые, полученные для двух моделей бинарной классификации, построенной на основе логистической регрессии

Анализ построенного графика показывает, что ROC-кривая, построенная на основании модели с дополненными данными, лежит выше, чем ROC-кривая, построенная на основе только индекса массы тела. Визуальный анализ должен быть дополнен вычислениями AUC. По результатам вычислений для однопараметрической модели было получено значение AUC, равное 0,64, и для двухпараметрической модели AUC равен 0,73.

ЗАКЛЮЧЕНИЕ

В двух частях представленных методических рекомендаций изложены основные методы статистического анализа данных. В качестве инструмента для проведения статистического анализа выбран свободно распространяемый язык программирования R.

Особенностью данного языка программирования является наличие в его составе большого количества библиотек и функций, предназначенных для проведения статистического анализа данных. Каждая из включенных в состав языка библиотек была ранее опубликована в академических рецензируемых журналах, что придает достаточную степень уверенности в достоверности результатов, получаемых при корректном ее применении.

В методических рекомендациях была предпринята попытка охватить достаточно большой объем основных методов статистического анализа, применяемых на практике в работе медицинских статистиков и биостатистиков, однако ряд тем остались неохваченными. Первая тема, которая не была обозначена в первой части методических рекомендаций, – факторный анализ, или анализ качественных данных. Это обширная тема, с которой сталкивается исследователь при необходимости анализа социологических, маркетинговых, психологических или медико-биологических данных.

Вторая тема, которая не была рассмотрена в методических рекомендациях – это построение прогностических моделей, таких как модели выживаемости или авторегрессионные модели.

И третья, нерассмотренная, обширная тема – это построение моделей принятия решений, в том числе с применением нейронных сетей. Каждая из этих тем требует отдельного пособия и отдельного рассмотрения. Язык программирования R позволяет достаточно быстро проводить построение обозначенных моделей, сосредотачиваясь на сути исследования, и не расходовать большой объем времени на написание программного кода.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Джейм Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. М.: ДМК Пресс, 2016. 450 с.
2. Гельман Э., Хилл Д., Вехтари А. Регрессия: теория и практика. С примерами на R и Stan. М.: ДМК Пресс, 2022. 748 с.

Серия «Лучшие практики лучевой и инструментальной диагностики»

Выпуск 140

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 2

Методические рекомендации

Отдел координации научной деятельности ГБУЗ «НПКЦ ДиТ ДЗМ»
Технические редакторы В. П. Гамарина, А. И. Овчарова
Компьютерная верстка Е. Д. Бугаенко

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