МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«Новосибирский национальный исследовательский государственный университет»

(Новосибирский государственный университет, НГУ)

Структурное подразделение Новосибирского государственного университета –

Высший колледж информатики Университета (ВКИ НГУ)

КАФЕДРА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

1. Направление подготовки: 09.03.01 Информатика и вычислительная техника
2. Образовательная программа: 09.03.01 Информатика и вычислительная техника

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА**

СОЗДАНИЕ БИБЛИОТЕКИ АЛГОРИТМОВ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ КЛИНИЧЕСКИХ ИССЛЕДОВАНИЙ В ПАРАЛЛЕЛЬНЫХ ГРУППАХ

утверждена приказом по ВКИ НГУ № 02/3-204 от «27» апреля 2017 г.

|  |  |  |  |
| --- | --- | --- | --- |
| Кошкаревой Софьи Владимировны, | группа 14214 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| (фамилия, имя, отчество студента) |  | (подпись студента) |  |

|  |  |  |
| --- | --- | --- |
| **«К защите допущена»** |  | **Руководитель ВКР** |
| Заведующий кафедрой, к.ф-м.н |  | к.ф-м.н, |
| (ученая степень, звание) |  | (ученая степень, звание) |
| Попов Л.К., /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  | Лукинов В.Л., /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| (ФИО) / (подпись) |  | (ФИО) / (подпись) |
| «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_г. |  | «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_г. |

Дата защиты: «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_ 20\_\_г.

Новосибирск

2017

СОДЕРЖАНИЕ

[ПЕРЕЧЕНЬ ТЕРМИНОВ И СОКРАЩЕНИЙ 4](#_Toc494307788)

[ВВЕДЕНИЕ 5](#_Toc494307789)

[1. ПОСТАНОВКА ЗАДАЧИ 6](#_Toc494307790)

[1.1. Описание предметной области 6](#_Toc494307791)

[1.2. Формулировка задачи 7](#_Toc494307792)

[1.3. Функциональные требования 8](#_Toc494307793)

[1.4. Нефункциональные требования 8](#_Toc494307794)

[1.5. Характеристики выбранных программных средств 9](#_Toc494307795)

[2. РЕАЛИЗАЦИЯ 10](#_Toc494307796)

[2.1. Алгоритмы решения задач 10](#_Toc494307797)

[2.1.1. Опечатки 10](#_Toc494307798)

[2.1.2. Выбросы 11](#_Toc494307799)

[2.1.3. Генерация отчетов 12](#_Toc494307800)

[2.2. Объектно-ориентированная модель 12](#_Toc494307801)

[2.2.1. Файлы 13](#_Toc494307802)

[2.2.2. Типы значений столбцов таблицы 15](#_Toc494307803)

[2.2.3. Типы ошибок 17](#_Toc494307804)

[2.2.4. Описание реализации поиска опечаток 19](#_Toc494307805)

[2.2.4.1. Метод поиска опечаток для дискретных значений 19](#_Toc494307806)

[2.2.4.2. Метода поиска опечаток для непрерывных значений 21](#_Toc494307807)

[2.2.4.3. Метод поиска опечаток для дат 23](#_Toc494307808)

[2.2.5. Реализация поиска выбросов 25](#_Toc494307809)

[2.2.6. Реализация поиска неупорядоченных дат 28](#_Toc494307810)

[2.2.7. Создание сводной таблицы значений 28](#_Toc494307811)

[3. РАЗРАБОТКА БИБЛИОТЕКИ 32](#_Toc494307812)

[4. ОТЛАДКА И ТЕСТИРОВАНИЕ 34](#_Toc494307813)

[РЕЗУЛЬТАТЫ 35](#_Toc494307814)

[ЗАКЛЮЧЕНИЕ 38](#_Toc494307815)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 39](#_Toc494307816)

[ПРИЛОЖЕНИЕ А 40](#_Toc494307817)

ПЕРЕЧЕНЬ ТЕРМИНОВ И СОКРАЩЕНИЙ

|  |  |
| --- | --- |
| **Выброс** | Значение, которое слишком велико или слишком мало по сравнению с большинством других имеющихся значений. |
| **Параллельные группы** | Это две или более группы испытуемых, одна или более из которых получают исследуемый препарат, а другая группа является контрольной. |
| **Медиана** | Такое число, что вероятность получить значение случайной величины справа от него равна вероятности получить значение слева от него. |
| **Математическое ожидание** | Это среднее значение случайной величины. |
| **Дисперсия** | Среднее значение квадрата отклонения случайной величины от её математического ожидания. |
| **Среднеквадратическое отклонение** | Значение, равное квадратному корню из дисперсии случайной величины. |
| **Нормальное распределение** | Это распределение вероятностей, совпадающее с функцией Гаусса. |
| **Квантиль** | Это значение, которое заданная случайная величина не превышает с фиксированной вероятностью. |
| **ИКР** | Интерквартильный размах. |
| **Интерквартильный размах** | Разность между третьим и первым квартилями. |

ВВЕДЕНИЕ

Целью данной работы является разработка программного обеспечения, предназначенного для проведения статистического анализа данных, полученных в результате клинических исследований.

На протяжении своего развития медицинское сообщество всегда старалось найти более эффективные способы лечения и диагностики болезней. Первоначальные методы были неэффективны из-за применения метода проб и ошибок и интуитивных обобщений. Для решения этой проблемы в медицине сформировалась новая область – доказательная медицина.

Доказательная медицина подразумевает такой подход к медицинской практике, при котором каждое решение, относящееся к выбору метода лечения, обязано иметь научное обоснование. Оно основывается на данных, полученных в ходе четко спланированного и документированного исследования, использующего методы статистического анализа.

Опечатки и пропуски отдельных значений в данных исследований являются постоянной проблемой. Поэтому, прежде чем начать применять статистические методы, обрабатываемые данные следует привести к приемлемому для обработки виду. Для этого необходимо идентифицировать возможные проблемы, а в дальнейшем, либо удалить некорректные данные, либо заменить имеющиеся пропуски и опечатки разумными значениями, что и было реализовано в данной работе. Еще одной важной проблемой является наличие выбросов. Под «выбросом» понимается значение, которое слишком велико или слишком мало по сравнению с большинством других имеющихся значений.

Разработанная библиотека решает проблему подготовки входных данных для статистического анализа в автоматическом режиме. Библиотека сводит к минимуму затраченное на это время, а также позволяет биостатистикам анализировать данные исследования в удобной форме.

1. ПОСТАНОВКА ЗАДАЧИ
   1. Оп**исание предметной области**

В большинстве случаев клинические исследования проводятся в параллельных группах. Параллельные группы – это две или более группы испытуемых, одна или более из которых получают исследуемый препарат, а другая группа является контрольной. Для достижения статистической достоверности испытуемые распределяются по группам методом случайной выборки.

При подготовке данных к статистическому анализу клинических исследований необходимо выполнять следующие рутинные процедуры:

* Поиск пропущенных значений (незаполненных полей).
* Исправление опечаток.
* Обнаружение выбросов.
* Проверка данных на условие нормальности распределения.
* Поиск неупорядоченных дат.

На данный момент выполнение этих процедур происходит в ручном режиме. Такой подход является неэффективным из-за человеческой невнимательности. При таком подходе невозможно обработать большой объем данных за разумное время.

Поэтому программная проверка существенно ускорит дальнейший анализ, а так же окажется более эффективной по сравнению с проверкой, выполняемой в ручном режиме.

Методы статистического анализа делятся на два вида:

* Параметрические.
* Непараметрические.

Параметрические методы имеют более высокую точность и эффективность по сравнению с непараметрическими, но имеют ограничения на входные данные – данные должны быть нормально распределены.

Нормальное распределение (или распределение Гаусса) – это распределение вероятностей, совпадающее с функцией Гаусса, вычисляемой по формуле, представленной на рисунке 1.

,

1. – функция Гаусса

где

μ – математическое ожидание (среднее значение),

σ – среднеквадратическое отклонение.

Математическое ожидание – это среднее значение случайной величины Дисперсией случайной величины называется среднее значение квадрата отклонения случайной величины от её математического ожидания. Среднеквадратическим отклонением случайной величины называется квадратный корень из дисперсии этой величины.

Стандартным нормальным распределением называется нормальное распределение с математическим ожиданием μ = 0 и стандартным отклонением σ = 1. Стандартное нормальное распределение отображено на рисунке 2 зеленой линией.



1. – нормальное распределение

Перед применением статистических моделей нужно удостовериться, что данные имеют нормальное распределение.

Создание библиотеки алгоритмов статистического анализа данных для клинических исследований в параллельных группах, позволит существенно сократить время проведения статистического анализа и поможет проводить более качественные исследования в короткие сроки.

* 1. Формулировка задачи

Целью работы являлась разработка библиотеки языка R с инструкцией по эксплуатации, предназначенной для проведения статистического анализа данных клинических исследований в параллельных группах.

Программирование осуществлялось на языке R, с использованием следующих библиотек:

* base.
* methods.
* utils.
* grDevices.
* plyr.
* xlsx.
* sm.
* nortest.
* devtools.
* roxygen2.

Для обеспечения возможности возврата к определённым старым версиям разработки, использовалась система контроля версий git и репозиторий на сервере GitHub.

* 1. Функциональные требования

В рамках дипломной работы были поставлены следующие задачи:

* Поиск пропущенных значений (незаполненных полей).
* Поиск опечаток.
* Поиск выбросов.
* Проверка на упорядочение дат.
* Исследование нормальности распределения данных различными статистическими критериями.
  1. Нефункциональные требования

В рамках дипломной работы были определены следующие нефункциональные требования:

* Ввод данных должен осуществляться в виде Excel-файла или внутренней структуры языка R – таблиц данных (data.frame).
* Результаты работы должны записываться в Excel-файл.
* Реализация методами ООП, используя объектную модель S4 в языке R.
* Использование системы контроля версий Git в связке с сервером GitHub.
* Тестирование созданной библиотеки на реальных данных.
  1. Характеристики выбранных программных средств

Для решения задачи был выбран язык R и среда разработки RStudio.

R – язык программирования высокого уровня, предназначенный для статистической обработки данных с упором на визуализацию и воспроизводимость. Он так же является свободной кроссплатформенной программной средой вычислений с открытым исходным кодом в рамках проекта GNU. R – интерпретируемый язык с интерфейсом командной строки. Он сочетает в себе процедурное, функциональное и объектно-ориентированное программирование.

На сегодняшний день R является широко известным среди свободно распространяемых систем статистического анализа. Он обладает хорошей расширяемостью с помощью библиотек, содержащих набор специфических функций. В среде R реализованы многие статистические методы: линейные и нелинейные модели, проверка статистических гипотез, анализ временных рядов, классификация, кластеризация, графическая визуализация. Еще одной из особенностей языка является поддержка графических возможностей, позволяющая визуализировать данные в виде различных графиков и диаграмм.

1. РЕАЛИЗАЦИЯ
   1. Алгоритмы решения задач

Различные статистические методы по-разному относятся к наличию выбросов во входных данных. Наличие выбросов может сделать использование определенных статистических моделей *невозможным*, и в то же время, никак не сказаться на результатах других.

Именно поэтому перед проведением статистического анализа необходимо выполнять проверку начальных данных на валидность.

Валидизация данных – это процесс обнаружения и исправления ошибок.

Для каждого рода ошибок был разработан свой алгоритм обнаружения, о которых будет рассказано ниже.

* + 1. Опечатки

Опечатки – это некорректно введенные пользователем данные. Примером опечаток может служить:

* Данные, не входящие в набор определенных допустимых значений.
* Наличие буквенных символов в числовых данных.
* Неправильные разделители между числами в записи дат.
* Лишние пробелы в записи десятичных дробей.

На рисунке 3 изображена блок-схема алгоритма поиска опечаток для данных, имеющих набор определенных допустимых значений. Например, записи о тяжести состояния пациента (только следующие значения: «удовлетворительное», «средней тяжести», «тяжёлое», «крайне тяжёлое»).

Остальные виды опечаток определяются с помощью шаблонов, заданных регулярными выражениями.



1. – блок-схема алгоритма поиска опечаток для данных с набором допустимых значений
   * 1. Выбросы

Под «выбросом» понимается значение, которое слишком велико или слишком мало по сравнению с большинством других имеющихся значений. К примеру, средний вес большинства участников исследования – 75 кг. Но в исследовании также приняли участие два человека весом 55 кг и 110 кг, именно эти значения будут выбросами.

Причины появления выбросов могут быть различными:

* Из-за ошибки измерения.
* Из-за необычной природы входных данных. Например, если наугад измерять температуру предметов в комнате, то большинство полученных значений будет лежать в диапазоне от 18 до 22 °C, но радиатор отопления будет иметь температуру в 70°.
* Выбросы могут быть и частью распределения.

Для обнаружения выбросов была использована функция boxplot.stats(), производная графической функции высокого уровня boxplot(), которая служит для построения диаграмм размахов. А boxplot.stats(), в свою очередь, используется для сбора статистики, необходимой для создания диаграмм размахов. Подробный механизм реализации описан в пункте 2.2.5.

К выбросам всегда следует относиться внимательно. Они вполне могут оказаться «нормальными» для исследуемой совокупности, и поэтому требуют дополнительного изучения причин их появления. Поэтому производится только идентификация таких значений, а не поспешное удаление и исключение их из статистического анализа.

* + 1. Генерация отчетов

Библиотека позволяет создавать пользовательские отчеты об ошибках нескольких типов:

* Текстовый файл, в который построчно записываются сообщения о найденных некорректных данных. Текст сообщения содержит тип ошибки, позицию некорректного значения в таблице данных, само значение.
* Excel-файл с несколькими стилями, которые применяются для наглядного отображения найденных ошибок.
* Сводная таблица данных в виде Excel-файла, которая содержит различные значения каждого столбца исходных данных и частоту встречаемости этих значений. Она дает верное представление о содержимом входных данных перед началом их обработки, а также помогает в определении допустимого набора значений для исправления опечаток.
  1. Объектно-ориентированная модель

Поставленные задачи были реализованы с использованием объектно-ориентированной (ООП) модели S4 языка R. Для решения поставленных задач были разработаны три независимых структуры классов:

* Файлы.
* Столбцы таблицы.
* Типы ошибок.

Подробнее каждая структура описана в следующих пунктах.

* + 1. Файлы

В процессе выявления различных ошибок требуется обеспечить ввод и вывод различных данных. Для этого была разработана структура классов *Файлы*, которая представлена на рисунке 4.



1. – структура классов Файлы

Класс-родитель File имеет одно поле path и два метода для взаимодействия с ним. Метод SetPath() устанавливает полный путь к файлу, а метод GetPath() служит для получения текущего. У класса File существует три дочерних класса: In, Out и Report. Каждый из них наследует родительские методы и поле path, значение которого для каждого класса будет разным.

Класс In имеет метод Read(), реализующий считывание начальных данных из Excel-файла, и поле table, куда помещается результат считывания в виде внутренней структуры языка R, таблицы данных (data.frame).

Результат исправления опечаток в исходных данных присваивается полю table класса Out.

Также, в классе Out реализованы методы для записи результатов в Excel и созданы необходимые для этих методов поля. Поля wb и sheet являются объектами класса jobjRef из библиотеки rJava, которая, в свою очередь, используется библиотекой xlsx для связи Java и R. Поле sheet\_name содержит название листа для новой рабочей книги Excel. Назначить новое название можно используя SetExcelSheetName(), а чтобы узнать текущее – метод GetExcelSheetName().

Метод CreateExcelWB() нужен для создания новой рабочей книги с именованным листом, и добавления на него итоговой таблицы данных, к которой будут применены стили. Также в этом методе создается новая пустая строка в шапке таблицы для того, чтобы разместить там обозначения различных типов ошибок.

Метод SaveExcelWB() сохраняет рабочую книгу, используя в качестве полного пути содержимое поля path объекта Out. Перед сохранением файла устанавливается автоподбор ширины столбцов и закрепляется первая строка таблицы (легенда). Методы CreateExcelWB() и SaveExcelWB() являются «оберткой» для работы с библиотекой xlsx.

Для работы с текстовым файлом и записью в него сообщений для пользователя был создан класс Report. У него есть поле file, которое используется для создания файла-отчета. Полный путь к файлу генерируется методом setDirectory(), который добавляет к указанной при вызове метода директории строку «Report\_» и текущую дату и время. Метод Create() создает по указанному полному пути файл и открывает соединение для записи в текстовом режиме, а метод Close() закрывает это соединение.

Сводная таблица должна содержать названия столбцов исходной таблицы, их значения и частоту встречаемости каждого из значений. Для создания сводной таблицы был создан класс SummaryTable, являющийся потомком класса Out, и метод ColumnsValues(), механизм работы которого описан в пункте 2.2.7.

* + 1. Типы значений столбцов таблицы

Структура данных таблицы, с которой работает библиотека, может состоять из четырех типов переменных:

* Дата.
* Непрерывные переменные.
* Дискретные переменные.
* Категориальные переменные.

В виде даты в результатах исследования может указываться время измерения различных показателей пациента, время его поступления в клинику и время выписки. Важно следить за тем, чтобы даты повторных измерений сохраняли упорядоченность, т.е. чтобы дата повторного измерения была больше даты первичного.

Непрерывные переменные могут принимать любые численные значения, которые естественным образом упорядочены на числовой оси (например, рост, вес).

Дискретные переменные могут принимать счётное множество упорядоченных значений, которые могут просто обозначать целочисленные данные или ранжировать данные по степени проявления на упорядоченной ранговой шкале (клиническая стадия опухоли, тяжесть состояния пациента).

Категориальные переменные являются неупорядоченными и используются для качественной классификации (пол, цвет глаз, место жительства); в частности, они могут быть бинарными (дихотомическими) и иметь категорические значения: 1/0, да/нет, имеется/отсутствует. Поэтому каждый из четырех типов значений описывает свой класс, и для каждого из них реализован свой метод поиска ошибок. Структура классов *Столбцы таблицы* представлена на рисунке 5.



1. – структура классов Столбцы таблицы

Родительский класс Column имеет поле column\_index, в котором хранится номер столбца в исследуемой таблице, и два метода GetColumnIndex(), SetColumnIndex() для задания нового значения этого поля и получения текущего. Данные методы доступны во всех четырех дочерних классах.

Объекты класса Continuous описывают столбцы таблицы, содержащие непрерывные переменные. Для объектов класса Continuous метод FindErrors() состоит из поиска опечаток и поиска выбросов.

Объекты класса Discrete описывают дискретные переменные. Для поиска и исправления опечаток была разработана структура словаря.

Словарь состоит из ключей и значений. Одному значению может соответствовать множество ключей.

Поля объекта Discrete, key (ключ) и value (значение), являются переменными R-типа list, т.е. они имеют вложенную структуру, представляющую из себя массив массивов, и могут содержать в себе сочетания любых типов данных. Это позволяет эффективно, т.е. в одном объекте, хранить разнородную информацию. Более подробное описание работы со структурой словаря описано ниже, в пункте 2.2.4.1.

Задать новые значения для полей класса Discrete, key и value, можно при помощи созданных методов SetValue() и SetKey(), а получить их текущие значения при помощи GetValue() и GetKey() соответственно. Эти методы также доступны в дочернем классе.

Дочерний класс Binary имеет более узкую специализацию и описывает только бинарные категориальные переменные (например, пол). Во время инициализации класса Binary значения поля value устанавливаются по умолчанию: 0 и 1. При необходимости, они могут быть изменены при помощи родительского метода SetValue().

Объекты класса Dates необходимы для описания столбцов таблицы, содержащих даты.

Метод FindErrors() для классов Discrete и Dates выполняет поиск опечаток, но для каждого из этих двух классов существует своя реализация данного метода, более подробно это описано в пункте 2.2.4.3.

* + 1. Типы ошибок

Процесс валидации «сырых» данных должен выявлять следующие типы ошибок:

* Опечатки.
* Неупорядоченные даты.
* Выбросы.
* Пропущенные значения (незаполненные поля).

Для каждого типа ошибок был создан свой класс, который наследуется от класса Error. Это сделано для объединения их по общим признакам, таким как:

* Индексы ошибок в таблице.
* Стиль типа ошибки для раскраски итоговой таблицы.
* Название типа ошибки в легенде таблицы.
* Позиция условного обозначения типа ошибки внутри легенды таблицы.

На рисунке 6. изображена структура классов *Типы ошибок*.



1. – структура классов Типы ошибок

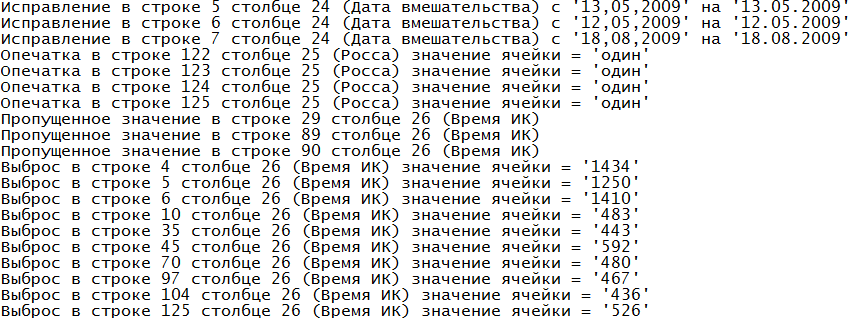
Родительский класс Error имеет следующие поля: indices, style, title и col\_index\_legend. Поле indices содержит индексы ячеек с ошибками в исходной таблице, поле style для хранения названия стиля, применяемого для раскраски неправильно заполненных ячеек, поле title, содержащее название типа ошибки (это название используется для легенды итоговой таблицы), и col\_index\_legend, в котором хранится позиция ячейки внутри легенды таблицы.

Значения для полей style, title и col\_index\_legend устанавливаются во время инициализации каждого из дочерних классов Error. Для этого используется созданный метод Initialize(), а не стандартный конструктор.

Каждый вызов метода FindErrors() изменяет поле indices, добавляя новые индексы ячеек таблицы, содержащих ошибки.

После того, как был выполнен анализ всех столбцов таблицы, итоговая таблица может быть раскрашена. Для этого был создан метод SetColor(), который взаимодействует с объектом новой рабочей книги Excel, созданной при вызове метода CreateExcelWB(). В методе SetColor() осуществляется доступ ко всем ячейкам рабой книги и устанавливаются новые стили для тех ячеек, индексы которых содержит поле indices. Выбор стиля для типа ошибки осуществляется исходя из значения, содержащегося в поле style. Внутри метода SetColor() также осуществляется вызов метода AddTableLegend(), при помощи которого происходит добавление условного обозначения в легенду таблицы.

Метод PrintReport() отвечает за запись сообщений о найденных ошибках в пользовательский текстовый файл. Он принимает в качестве одного из аргументов объект класса Report и использует его поле file с открытым соединением. Тип ошибки определяется автоматически, путем получения значения поля title переданного методу объекта. Фрагмент отчета об ошибках представлен на рисунке 7.



1. – фрагмент отчета об ошибках
   * 1. Описание реализации поиска опечаток

Поиск опечаток и их исправление, осуществляется при помощи метода FindMisprints(). У него существует три реализации для каждого из четырех типов значений исследуемой таблицы, описанных выше.

* + - 1. Метод поиска опечаток для дискретных значений

Доступ к элементам определенного столбца в исходной таблице данных осуществляется по его индексу, который был передан при вызове метода. Полученные значения элементов присваиваются новой переменной, представляющей собой одномерный массив. В цикле, последовательно проверяется каждый элемент этого массива на наличие различных ошибок.

В первую очередь осуществляется поиск пропущенных значений. Если проверка выявила существование пропущенного значения, то его индексы передается полю объекта класса missingValue для дальнейшей раскраски, и вызывается метод PrintReport(), который производит запись сообщения о найденной ошибке в пользовательский текстовый файл-отчет.

Если результат поверки на заполнение оказался успешным и значение ячейки отлично от пустой строки, то начинается проверка на совпадение с одним из значений словаря. Для корректности сравнения, значение словаря и значение элемента столбца приводятся к верхнему регистру, во избежание ошибок.

Если значение столбца не совпало ни с одним значение словаря, проверяется совпадение элемента с одним из ключей. Оба значения так же приводятся к верхнему регистру.

Если совпадение найдено, то по нужному индексу в итоговой таблице производится замена ключа на соответствующее ему значение в словаре, а индексы элемента передаются полю объекта класса Misprint для раскраски итоговой таблицы. Далее происходит вызов метода PrintReport(). Ячейка, содержащая данную ошибку, в итоговой таблице будет выделена как исправленная опечатка.

В случае, когда элемент столбца не совпал ни с одним из ключей, индексы этого элемента в таблице (номер строки и номер столбца) передаются полю объекта класса UnsolvedMisprint, и производится запись сообщения об ошибке в файл при помощи метода PrintReport().

Наглядное представление описанного алгоритма приведено на рисунке 8 в виде блок-схемы.



1. – блок-схема алгоритма поиска опечаток для дискретных значений
   * + 1. Метода поиска опечаток для непрерывных значений

Так же, как и в методе, реализованном для категориальных переменных, в первую очередь производится доступ к значениям определенного столбца и проверка на заполнение всех его элементов.

Далее значение элемента сравнивается с шаблоном, заданным регулярным выражением, показанным на рисунке 9.



1. – шаблон регулярного выражения для непрерывных значений

где

^ – символ начала строки,

(\\d)+ – одно число и более,

[,.]– один из двух указанных символов (точка или запятая),

| – логическое ИЛИ,

[[:space:]])?– пустые символы ноль или один раз,

$ – символ конца строки.

Данный шаблон был разработан для работы с целыми и дробными числами.

Если значение элемента не соответствует шаблону, это может означать, что оно содержит буквы, или любые другие символы. Следовательно, индексы элемента передаются полю объекта класса UnsolvedMisprint, который, как было сказано ранее, хранит индексы элементов таблицы, которые будут раскрашены как неисправленные опечатки. Далее производится запись сообщения о найденной ошибке в файл при помощи метода PrintReport().

Если значение элемента соответствует шаблону, то производится дополнительная проверка разделителя в записи десятичной дроби и поиск лишних пробелов. Если разделитель неверный, то производится его замена, а если найден лишний пробел – он удаляется, а индексы элемента передаются полю объекта класса Misprint для раскраски итоговой таблицы. Метод PrintReport() сообщает пользователю о найденной и исправленной опечатке, если таковая выявлена. Наглядное представление описанного алгоритма приведено на рисунке 10 в виде блок-схемы.



1. – блок-схема алгоритма поиска опечаток для непрерывных значений
   * + 1. Метод поиска опечаток для дат

Данная реализация схожа с предыдущей, но отличается применяемым шаблоном регулярного выражения, который используется для сравнения со значением элемента. Шаблон регулярного выражения приведен на рисунке 11.



1. – шаблон регулярного выражения для непрерывных значений

где

^ – символ начала строки,

(\\d){2}) –два числа,

([,.]|[-/]) – один из четырех указанных символов («,», или «.» , или «\», или «/» или «-»),

((\\d){2}|(\\d){4}) – два или четыре числа,

$ – символ конца строки.

Если значение элемента не соответствует шаблону, то производится запись сообщения о найденной опечатке в пользовательский файл, а индексы этого элемента передаются полю объекта класса UnsolvedMisprint.

Все значения элементов колонки, которые подошли под указанный выше шаблон, являются датами, но существует возможность того, что они записаны неверно. Поэтому производится дополнительная проверка разделителей между числами в записи даты и замена их на точку, если разделитель неправильный. Далее, вызванный метод PrintReport() сообщает о найденной и исправленной опечатке, если таковая выявлена, и добавляет индексы такого элемента полю объекта класса Misprint. Наглядное представление описанного алгоритма приведено на рисунке 12 в виде блок-схемы.



1. – блок-схема алгоритма поиска опечаток для дат
   * 1. Реализация поиска выбросов

Поиск выбросов осуществляется только среди значений элементов столбцов, которые описывает класс Continuous, т.е. непрерывных переменных. Анализ значений возможен лишь в том случае, когда среди них нет неисправленных опечаток. В противном случае результаты анализа будут ошибочными и, следовательно, бесполезными.

Выбросы в столбце определяются при помощи функции boxplot.stats(), производной графической функции высокого уровня boxplot(), которая служит для построения диаграмм размахов. boxplot.stats(), в свою очередь, используется для сбора статистики при создании этих диаграмм.

Диаграммы размахов, или «ящики с усами» (англ. box-whisker plots), получили свое название за характерный вид. Точку или линию, соответствующую медиане, окружает прямоугольник («ящик»), длина которого соответствует одному из показателей разброса или точности оценки генерального параметра. Дополнительно от этого прямоугольника отходят «усы», также соответствующие по длине одному из показателей разброса или точности. Строение получаемых при помощи этой функции «ящиков с усами» представлено ниже на рисунке 13.

[](http://3.bp.blogspot.com/-sqSGopnp0lo/Uvu_wl_dPQI/AAAAAAAAAgs/F2DBOSdfiU4/s1600/boxplot.PNG)

1. – диаграмма размахов

В R при построении диаграмм размахов используются устойчивые оценки центральной тенденции (медиана) и разброса (интерквартильный размах, далее ИКР).

Медианой случайной величины является такое число, что вероятность получить значение случайной величины справа от него равна вероятности получить значение слева от него (обе они равны 1/2).

Квантиль в математической статистике – это значение, которое заданная случайная величина не превышает с фиксированной вероятностью. α-квантиль – числовая характеристика закона распределения случайной величины; такое число, что данная случайная величина попадает левее его с вероятностью, не превосходящей α.

Термин «квартиль» (от «кварта» – одна четвертая часть) используют для обозначения квантиля порядка α, когда α кратно 1/4. Выделяют нижний квартиль при α =1/4 и верхний квартиль при α = 3/4. Квартиль, соответствующий α = 1/2, имеет собственное название – медиана.

Интерквартильным размахом называется разность между третьим и первым квартилями. Простейшие способы определения выбросов основаны на интерквартильном расстоянии. Например, всё, что не попадает в диапазон, представленный формулой, представленной ниже, считается выбросами.

где

– первый квартиль,

– третий квартиль,

– неотрицательная постоянная Тьюки, равная 1,5.

Верхний «ус» простирается от верхней границы «ящика» до наибольшего выборочного значения, находящегося в пределах заданного расстояния произведения коэффициента 1,5 на ИКР от этой границы. Нижний «ус» строится аналогично верхнему от нижней границы «ящика» до наименьшего выборочного значения. Наблюдения, находящиеся за пределами «усов», потенциально могут быть выбросами.

Далее описана программная реализация поиска выбросов.

В первую очередь выполняется проверка значения счетчика неисправленных опечаток. Если он равен нулю, то это означает, что можно производить дальнейший анализ значений элементов столбца. В противном случае, происходит вызов метода печати в файл PrintReport() и запись предупреждения о невозможности проведения дальнейшего анализа, т.к. среди значений столбца присутствуют неисправленные опечатки.

Далее осуществляется доступ к элементам столбца и поиск среди них пропущенных значений.

Функция boxplot.stats() работает только с числовыми значениями. Опечатки в столбце, содержащем непрерывные переменные, препятствуют правильному распознаванию его значений средой R как числовых, вместо этого они представляются как значения строкового типа. Поэтому, для того, чтобы включить эти значения в анализ выбросов, используется шаблон, изображенный на рисунке 14, заданный при помощи регулярного выражения.



1. – шаблон регулярного выражения для непрерывных значений

где

^ – символ начала строки,

(\\d)+ – одно число и более,

[,.] – один из двух указанных символов (точка или запятая),

([,.](\\d)+)? – разделитель, одно число и более ноль или один раз,

$ – символ конца строки.

Все значения элементов столбца сравниваются с данным шаблоном и при совпадении сохраняются в новую буферную переменную, являющуюся одномерным массивом. Далее, среди значений этой переменной производится замена разделителей в десятичных дробях на точку. После замены используется приведение типа каждого элемента переменной к числовому, и выполняется поиск выбросов с помощью функции boxplot.stats(), которая анализирует сразу весь массив элементов. Результат работы функции хранится в виде вложенной структуры R-типа list. Значения (не их индексы) из переданного массива, потенциально являющимися выбросами, содержатся в элементе out данной структуры. Если он содержит какие-либо значения, то они сопоставляются с исходными значениями столбца, для того, чтобы определить позицию уже найденных выбросов в таблице. Для сравнения разделители в исходных элементах, все еще являющихся строковыми, тоже преобразуются. Во время вызова метода PrintReport() происходит печать сообщения о найденных выбросах.

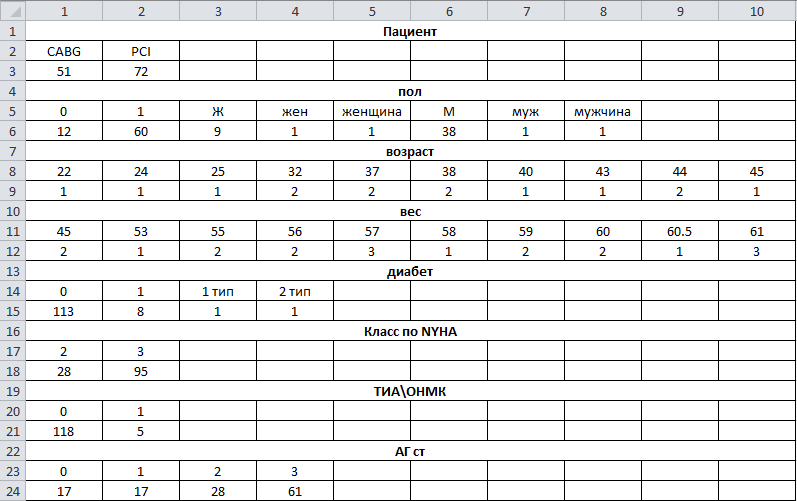
* + 1. Реализация поиска неупорядоченных дат

Метод Find() для класса DateMisprint реализует поиск непоследовательных дат внутри таблицы. Двумя из его аргументов являются два объекта класса Dates, которые описывают столбцы с разными датами, которые предстоит проверить. Например, это могут быть даты первичных и повторных замеров или даты поступления и выписки пациента. После сортировки столбцов по индексам производится приведение каждого из их элементов к типу данных R, date, с форматом представления месяц.день.год. После приведения типов производится присваивание полученных значений двум новым переменным. Далее эти переменные поэлементно сравниваются между собой. Если какое-либо из значений элементов первого столбца дат оказалось больше значения второго столбца дат в этой же строке, то индексы такого элемента добавляются в поле indices объекта DateMisprint. Далее происходит вызов метода PrintReport(), с помощью которого добавляется новое сообщение об ошибке в текстовый файл. В результирующей таблице будут выделены ячейки обоих столбцов в этой строке.

* + 1. Создание сводной таблицы значений

Данный функционал необходим для того, чтобы перед началом проверки «сырых» данных таблицы на валидность, пользователь получил представление о значениях, которые она содержит. Исходя из этого, он сможет верно сопоставить столбцы таблицы (вернее, их индексы) и классы, которые должны их описывать, а так же задать ключи и значения словарей, которые будут использоваться для поиска и исправления опечаток.

Из приведенного на рисунке 15 фрагмента сводной таблицы видно, что в столбце «пол» исходной таблицы содержатся значения: «0», «1», «Ж», «жен», «женщина», «М», «муж», «мужчина». Частота встречаемости каждого из этих значений указана на следующей строке сводной таблицы. Это можно рассматривать как гистограмму, изображенную на рисунке 16. Столбцы гистограммы – это различные значения, а их высота – частота встречаемости. Исходя из этих данных, можно определить словарь для столбца «пол». Ключи: «М», «муж», «мужчина» и «Ж», «жен», «женщина». И соответствующие этим ключам допустимые значения: «1» и «0». Данный словарь будет использоваться для поиска и замены опечаток. Результат обработки столбца «пол» будет содержать только значения «1» и «0».



1. – фрагмент сводной таблицы
2. – гистограмма для столбца «пол» исходной таблицы

Реализацией такого функционала является метод ColumnsValues(), который работает с объектом класса SummaryTable. В нем названия столбцов исходной таблицы, полученные при помощи функции colnames(), присваиваются новой переменной. Далее, в цикле к элементам каждого столбца исходной таблицы применяется функция table(). Ее результат содержит все различные значения столбца и частоту встречаемости каждого из этих значений. Производится преобразование типа результата функция table() в R-тип таблицу данных (data.frame). Для того чтобы получить нужный формат используется функция транспонирования таблицы данных как матрицы, а после используется обратное приведение типа. Полученная таблица добавляется в созданную ранее пустую сводную таблицу данных в виде новых строк. Т.к. количество различных значений в разных столбцах таблицы неодинаковое, это создает проблемы для создания сводной таблицы обычным путем – ее строки должны быть одной длины. Поэтому для решения этой проблемы используется функция rbind.fill() из библиотеки plyr, которая заменяет значения недостающих столбцов на NA. Если количество различных значений в одном столбце больше указанного пользователем, например, десять, то для отображения будут использованы только первые десять значений столбца.

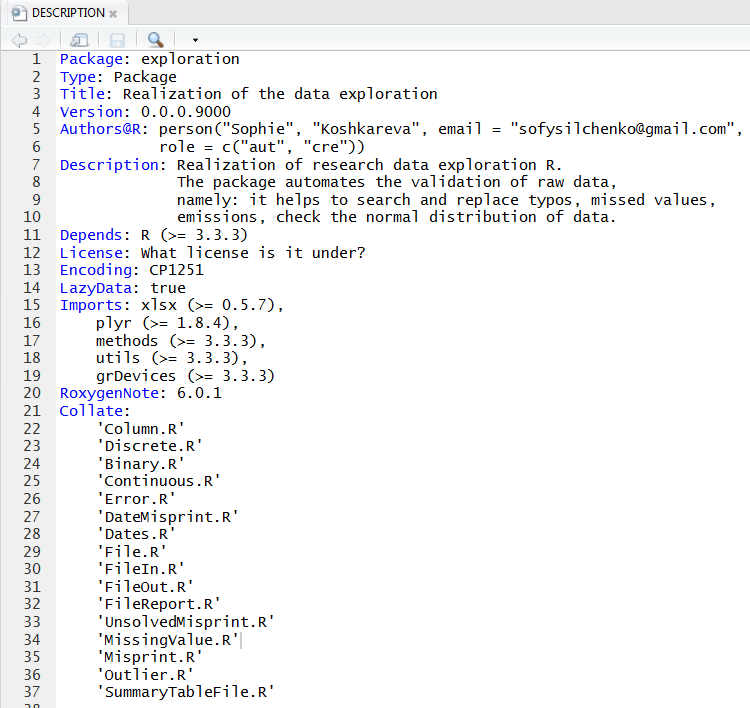
Названия столбцов исходной таблицы вставляются перед каждой новой строкой, описывающей различные значения и их частоту. Далее происходит запись полученной сводной таблицы в новую рабочую книгу Excel, путем вызова родительского метода CreateExcelWB(). Устанавливаются необходимые стили оформления и при помощи метода SaveExcelWB() производится сохранение нового Excel-файла.

1. РАЗРАБОТКА БИБЛИОТЕКИ

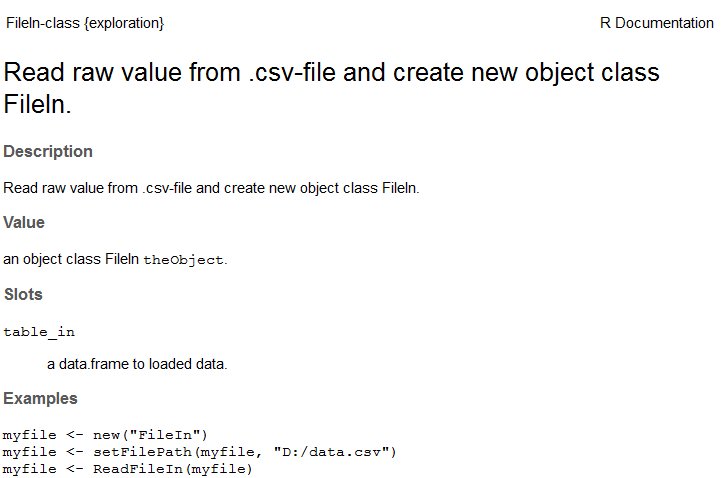
После написания классов и методов, все содержащие их файлы были объединены в новую библиотеку. Для сборки использовались вспомогательные библиотеки devtools и roxygen2.

Порядок действий:

* Установка библиотек devtools и roxygen2.
* Создание каркаса новой библиотеки.
* Перенос файлов, в которых определены классы и методы в нужную директорию внутри библиотеки.
* Изменение файла DESCRIPTION (на рисунке 17).
* Изменение файла NAMESPACE.
* Добавление документации (на рисунке 18).



1. – настройки файла DESCRIPTION



1. – страница из документации библиотеки
2. ОТЛАДКА И ТЕСТИРОВАНИЕ

На этапе отладки и тестирования на реальных данных было выявлено и исправлено несколько ошибок.

Текстовый файл-отчет, сводная таблица и файл с итоговой раскрашенной таблицей каждый раз перезаписывались, было невозможно посмотреть их ранние версии. Эта проблема была решена путем добавления текущего времени и даты в названия файлов.

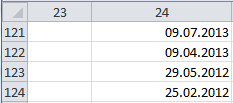
Среда R не всегда верно интерпретирует значения в столбцах таблицы. Из-за опечаток, например, лишнего пробела после запятой в десятичной дроби, все числовые значения столбца становятся строковыми и не включаются в анализ наряду с числовыми. Данная проблема была исправлена при помощи шаблонов регулярных выражений и приведения типов.

Во время тестирования метода для создания сводной таблицы, которая содержит значения каждого столбца, и частоту встречаемости этих значений, была выявлена ошибка. Длины столбцов получаются разными, а для компоновки таблицы данных или матрицы они должны быть одной длины. Ошибка была устранена с помощью функции rbind.fill из библиотеки plyr, которая заменяет значения недостающих столбцов на NA.

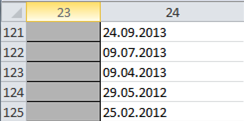
РЕЗУЛЬТАТЫ

Разработанные алгоритмы были проверены на реальных входных данных. Далее на рисунках будут представлены входные данные, содержащие ошибки и результат их обработки.

На рисунке 19 показан фрагмент входных данных, содержащих пропущенные значения. На рисунке 20 показан фрагмент итоговой таблицы с найденными и выделенными незаполненными ячейками.



1. – фрагмент данных с опечатками и пропущенными значениями

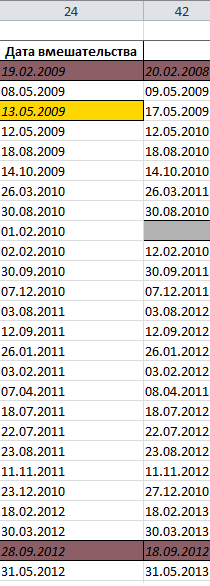


1. – результат обработки входных данных

На рисунке 21 изображен фрагмент входных данных, который содержит нарушение упорядоченности дат, а на рисунке 22 представлена итоговая раскрашенная таблица с выделением этой ошибки.

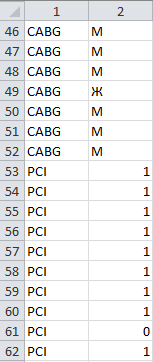


1. – фрагмент данных с неупорядоченными датами

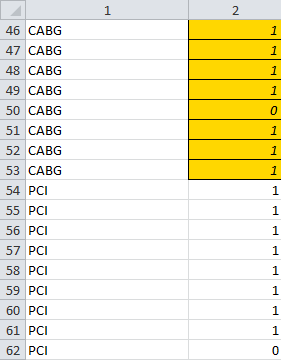


1. – фрагмент исправленной таблицы

Опечатки в виде данных, не входящих в набор определенных допустимых значений отображены на рисунке 23. Исправление этих опечаток показано на рисунке 24.

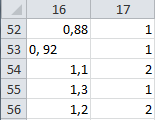


1. – данные, не входящие в набор допустимых значений

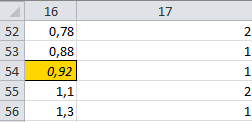


1. –исправленные опечатки

Наличие такой опечатки, как лишний пробел в записи десятичных дробей, хорошо видно на рисунке 25. На рисунке 26 представлен фрагмент таблицы, в которой данная ошибка исправлена.

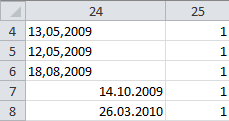


1. – лишний пробел в исходных данных

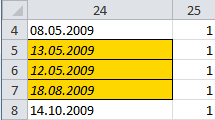


1. – фрагмент таблицы с удаленным лишним пробелом

Неправильные разделители между числами в записи дат показаны на рисунке 27. В результате обработки они были найдены и заменены на точки, это показано на рисунке 28.

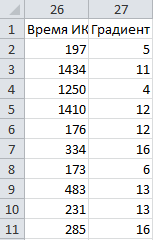


1. – неправильные разделители в записи дат

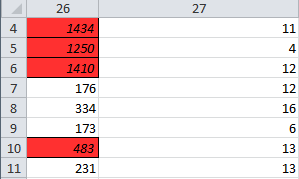


1. – исправленные разделители

На рисунке 29 отображен фрагмент входных данных, а на рисунке 30 выделены значения из этого же фрагмента, потенциально являющиеся выбросами.

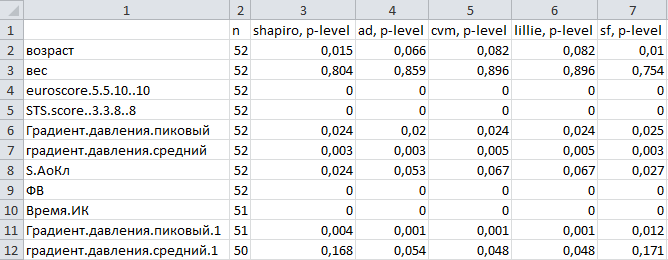


1. – фрагмент входных данных

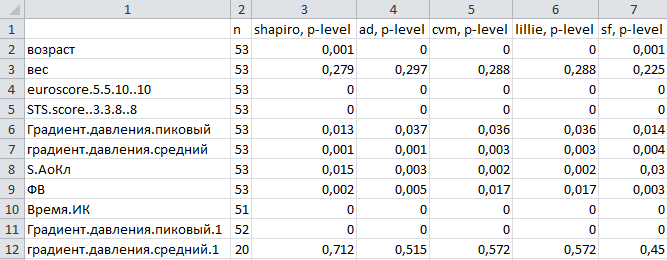


1. – значения, потенциально являющиеся выбросами

Проверка нормальности распределения проводится после того, как все найденные во входных данных ошибки исправлены, в том числе и выбросы. На рисунке 31 приведены результаты проверки нормальности распределения данных пациентов из группы «CABG», а на рисунке 32 – для данных пациентов из группы «PCI».

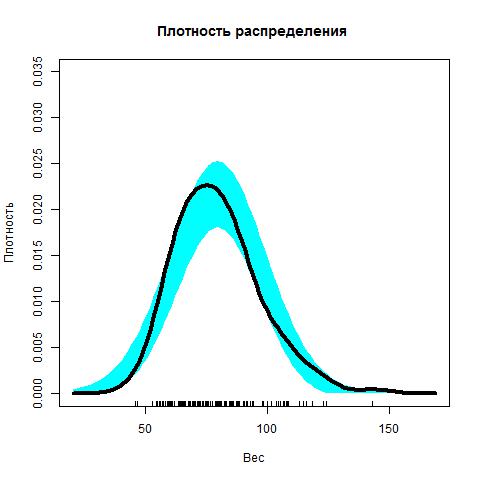


1. – результаты проверки нормальности распределения данных пациентов из группы «CABG»



1. – результаты проверки нормальности распределения данных пациентов из группы «PCI»

На рисунке 33 приведен график плотности распределения для столбца «Вес» исправленных данных.



1. – график плотности распределения для столбца «Вес»

Фрагмент сводной таблицы показан на рисунке 34. Она содержит названия столбцов исходной таблицы, различные значения каждого столбца и частоту встречаемости этих значений.



1. – фрагмент сводной таблицы

ЗАКЛЮЧЕНИЕ

Разработанная в рамках дипломной работы библиотека, решает поставленные задачи проверки входных данных, а также позволяет биостатистикам анализировать данные исследования в удобной форме за короткий промежуток времени.

Созданная библиотека алгоритмов для статистического анализа данных клинических исследований удовлетворяет всем поставленным требованиям:

* Выявление пропущенных значений (незаполненных полей).
* Поиск опечаток.
* Поиск выбросов.
* Исследование нормальности распределения различными статистическими критериями.
* Проверка на упорядочение дат.

В ходе выполнения дипломной работы была изучена предметная область, разработана и описана архитектура библиотеки, идентифицирующей потенциальные проблемы исследования данных, используя ООП модель S4 языка R. Был рассмотрен пример применения библиотеки на реальных задачах.

Таким образом, удалось создать библиотеку проверки входных данных, которая в дальнейшем позволит переложить рутинные действия на компьютер и даст возможность биостатистикам анализировать данные исследования удобной форме.

Выражаю искреннюю благодарность моему научному руководителю, Лукинову Виталию Леонидовичу, за помощь в подготовке данной работы и поддержку.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

ПРИЛОЖЕНИЕ А