Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное
учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт естественных и точных наук
Факультет математики, механики и компьютерных технологий
Кафедра прикладной математики и программирования
Направление подготовки: 01.03.02 Прикладная математика и информатика

РАБОТА ПРОВЕРЕНА
Рецензент, зам. директора
ВШЭКН по научной работе, д.к.т.н.
_____/Н.В. Плотникова
«____»_____20___г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.ф.-м.н.,
профессор
_____/А.А.Замышляева
«____»_____20___ г.

Разработка модулю статистического анализа потока пациентов в информационной системе МАУЗ ГКБ №2

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–01.03.02.2020.096.ПЗ ВКР

Руководитель работы,
ст.преп. кафедры ПМиП
_____/М.Ю. Сартасова
«____»_____2020 г.

Автор работы
Студент группы ЕТ-413
_____/Е.О. Фомина
«____»_____2020 г.
Нормоконтролер,
ст. преподаватель
_____/Н.С. Мидоночева
«____»_____2020 г.

Челябинск
2020

АННОТАЦИЯ

Фомина Е.О. Разработка модуля статистического анализа потока пациентов в информационной системе МАУЗ ГКБ №2. – Челябинск: ЮУрГУ, ЕТ-413, 58 с., 27 ил., 12 табл., библиогр. список – 30 наим.

Целью данной работы является разработка модуля статистического анализа потока пациентов в информационной системе МАУЗ ГКБ № 2.

В первом разделе был проведен анализ предметной области, проанализированы методы, применяемые для статистического анализа. Приводится обоснование выбранного метода.

Во втором разделе была составлена математическая модель статистического анализа потока пациентов.

В третьем разделе описана проведён сравнительный анализ различных конфигураций генетического алгоритма. Выполнено тестирование работы алгоритма на экспериментальных данных.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

МАУЗ ГКБ № 2 – это медицинское учреждение, которое представляет собой сложную систему с взаимодействующими сущностями и множеством как формальных, так и не формальных связей. Наладить работу медицинского учреждения – стремление каждого главного врача, потому что по итогу работы оптимизируется нагрузка на сотрудников, улучшается психологический климат в медицинском коллективе, что делает специалистов более динамичными и работоспособными. В результате увеличивается эффективность работы МАУЗ ГКБ № 2 и качество медицинской помощи пациентам.

Подходом к решению этой задачи является анализ потока пациентов в МАУЗ ГКБ № 2. В рамках решения данной задачи должны анализироваться и представляться в удобном для специалиста виде данные о траекториях потоков, интенсивности потоков, длине очередей, количестве и частоте отказов в обслуживании, а также суточных, недельных, месячных и сезонных изменениях данных характеристик.

Современная медицина это не только высококвалифицированные специалисты, это так же высокий уровень технологий. Выдвинутые экспертами гипотезы необходимо проверить, и математическая статистика является одним из инструментов анализа экспериментальных данных. Использование статистических программ предполагает знание основных методов и этапов статистического анализа: их последовательности, необходимости и достаточности.

Разработка модуля статистического анализа потока пациентов в информационной системе представляет определенную сложность, обусловленную объективными причинами:

1) сильная изменчивость исследуемых признаков ввиду влияния очень большого количества неуправляемых и неконтролируемых факторов;

2) проблемы в формировании выборок (планов экспериментов) требуемого объема и структуры;

3) трудности в освоении методов статистического анализа сотрудниками, не имеющими специального математического образования.

Статистическая обработка медицинских исследований основывается на принципе того, что истинное для случайной выборки истинно и для генеральной совокупности (популяции), из которой эта выборка получена. Выбрать или набрать истинно случайную выборку из генеральной совокупности очень сложно, следовательно, необходимо стремиться к тому, чтобы выборка была репрезентативной по отношению к изучаемой популяции.

В реабилитационном центре МАУЗ ГКБ № 2 текущие данные записывались в файл формата Excel, что затрудняло доступ к информации одновременно нескольким специалистам. И представляло определенные трудности при выполнении массовых операций с большими объемами данных.

Разработка данного модуля статистического анализа потока пациентов в информационной системе позволит улучшить качество и доступность медицинского обслуживания за счет увеличения объема и качества доступной информации, упростит информационный обмен между специалистами и пациентами, а также обеспечит своевременный сбор и анализ медицинской статистики.

1 МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА КОЛИЧЕСТВЕННЫХ И КАЧЕСТВЕННЫХ МЕДИЦИНСКИХ ДАННЫХ ПАЦИЕНТОВ В МАУЗ ГКБ № 2

1.1 Понятие и требования статистического наблюдения медицинских осмотров в МАУЗ ГКБ № 2

Статистическое наблюдение – это массовое, планомерное, научно организованное наблюдение за явлениями социальной и экономической жизни, которое заключается в регистрации отобранных признаков у каждой единицы совокупности [18].

Процесс проведения статистического наблюдения включает следующие этапы:

1) организационная подготовка к статистическому наблюдению;

2) сбор необходимых данных;

3) подготовка данных к автоматизированной обработке;

4) разработка вариантов по повышению качества статистических наблюдений.

Различные виды работы включает в себя процесс подготовки статистического наблюдения. Сперва решаются методологические вопросы. Важными вопросами являются: определение цели и объекта наблюдения, состава признаков, подлежащих регистрации; разработка документов для сбора данных; выбор отчетной единицы и единицы, относительно которой будет проводиться наблюдение, а также методов и средств получения данных [18].

Помимо методологических проблем существуют так же проблемы организационного характера. Главной задачей является составление календарного плана работ по подготовке медицинского персонала, поскольку проанализировать необходимо весь спектр медицинских услуг, представленный в МАУЗ ГКБ № 2.

Сбор данных включает работы, связанные непосредственно с заполнением медицинских форм отчета в специально разработанной информационной системе.

Все данные подвергаются логическому и арифметическому контролю данные на этапе их подготовки к автоматизированной обработке. Оба эти контроля основываются на знании взаимосвязей между показателями и качественными признаками.

На заключительном этапе проведения наблюдения анализируются причины, которые привели к неправильному заполнению медицинских форм отчета, и разрабатываются варианты по повышению качества статистических наблюдений.

1.2 Виды статистических данных в медицине

Поскольку размеры групп (выборок) и объемы данных могут варьироваться, а данные могут быть разнообразными, возникает необходимость использования методов статистического анализа.

Статистические данные можно разделить на количественные (числовые непрерывные или дискретные) и качественные (категориальные порядковые или номинальные) переменные [2].

Количественные данные предполагают, что переменная принимает некоторое числовое значение. Это позволяет определять интервал, который отделяет одну единицу от другой, а также упорядочивать единицы. Непрерывные числовые данные могут принимать любые значения, а числовые данные (дискретные) принимают определенные значения. Одним из примеров количественных данных является представление возраста двумя типами: в виде непрерывной переменной – указывается точный возраст пациента, и в виде дискретной переменной – указывается только количество полных лет (43,3 года и 43 года; 19,9 лет и 20 лет).

Существует три вида качественных данных: категориальные, номинальные и порядковые.

Для описания состояния объекта необходимо использовать категориальные данные. Объекту присваивается номер, соответствующий его категории. Определяющим условием для использования категориальных данных является отношение одного объекта только к одной возможной категории для одного критерия.

Номинальные данные – это такой вид качественных данных, которые отражают условные коды неизмеряемых категорий (например, коды диагнозов). Они используются в том случае, если категории неупорядоченные. Числа используются для обозначения состояния объекта. Например, по полу: 1 – женский, 2 – мужской.

Порядковые (ранговые, ординарные) данные – это такой вид качественных данных, который представляет условную степень выраженности какого-либо признака.

Основным отличием порядковых данных от дискретных количественных является отсутствие пропорциональной шкалы для измерения выраженности признака.

Описательная статистика одна из основных составляющих анализа данных. Важная задача описательной статистики – это предоставление сжатой и концентрированной характеристики изучаемого явления в графическом и числовом виде. На рисунке 1.1 изображена диаграмма, содержащая основные виды статистических данных.

Показатели описательной статистики можно разбить на несколько групп. Они необходимы для представления и анализа результатов всей выборки, экспериментальной и контрольной группы.

О нормальности распределения признака сообщают указание в представлении данных меры центральной тенденции (среднее, медиана, мода). При ассиметричном распределении эти показатели не совпадают, а при нормальном распределении совпадают.

Рисунок 1.1 – Виды статистических данных

Мода (Mo) – это наиболее частое значение в выборке, или среднее значение класса с наибольшей частотой [2]. Мода используется для того, чтобы дать общее представление о распределении.

Медиана (Me, Md) соответствует центральному значению в последовательном ряду всех полученных значений или среднему значению наиболее часто встречающихся значений выборки [2].

Среднее арифметическое (M) – это показатель центральной тенденции, полученной делением суммы всех значений данных на число этих данных [2]. Среднее арифметическое используется для представления количественных переменных с нормальным распределением. Среднее значение, как мера центральной тенденции в описательной статистике количественных данных, имеет одно из двух представлений. Первое в виде «$M \pm S$», или как в зарубежной традиции $M(S)$, где M – среднее, а S – стандартное отклонение. При нормальном распределении в диапазон $M \pm S$ укладывается порядка 70% всех значений признака.

Второе представление результатов – в виде «$M \pm m$», где m – стандартная ошибка среднего, определяемая следующим образом:

$$m = {}^{s}/_{\sqrt{n}}. \quad (1.1)$$

Данная форма представления данных не является информативной. Так как объектом наблюдения может быть сложная система, значительно различающаяся по своим свойствам, то это определяет отсутствие истинного значения параметра. Действительно, определяется не точное значение, а диапазон, в который укладывается большинство значений исследуемого признака, то есть ширина распределения. Поэтому оптимальным описанием ширины распределения принимается представление 95% доверительного интервала с указанием нижней (5%) и верхней (95%) границы.

1.3 Корреляционный и регрессионный анализ

Исследование между переменными изучается на этапе анализа данных. Корреляционный анализ, который устанавливает факт наличия или отсутствия зависимости между переменными, а также регрессионный анализ, и часто факторный анализ. Методы поиска зависимостей между переменными представлены на рисунке 1.2.

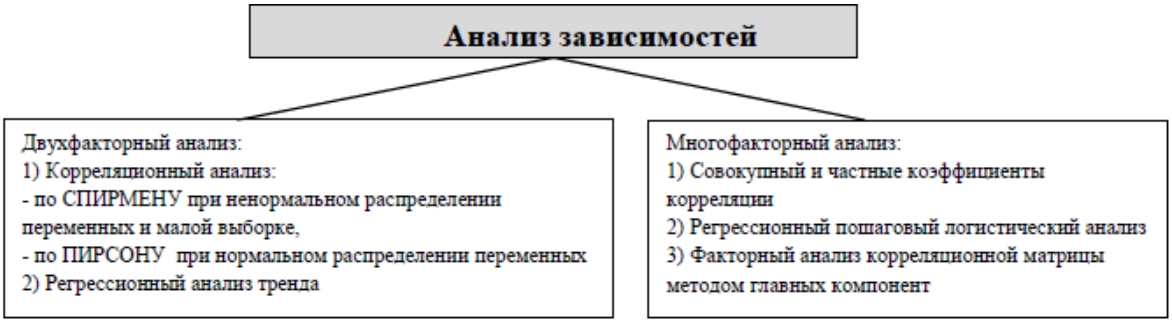

Рисунок 1.2 – Статистические методы поиска зависимостей между переменными

1.3.1 Корреляционный анализ

Корреляция – это взаимосвязь между двумя или более переменными [4]. Выявление наличия или отсутствия взаимосвязи – это основная цель данного анализа. Коэффициент линейной корреляции Пирсона r используется в том

случае, когда имеются две переменных, значения которых измерены в шкале отношений. Данный коэффициент принимает значения от -1 до $+1$, причем если значение находится ближе к 1, то это говорит о наличии сильной взаимосвязи, а если ближе к 0, то слабой. Отрицательный коэффициент означает наличие противоположной связи: чем выше значение одной переменной, тем ниже значение другой соответственно.

Термин «линейный» свидетельствует о том, что исследуется наличие линейной связи между переменными.

1.3.2 Регрессионный анализ

Регрессионный анализ говорит о наличии зависимостей между независимой переменной и одной или несколькими зависимыми переменными и позволяет определить вид этой связи, а также дает возможность для прогнозирования значения одной (зависимой) переменной, отталкиваясь от значения другой (независимой) переменной. Независимые переменные – это регрессоры или предикторы, а зависимые переменные называются критериальными [13].

1.3.3 Бинарная логистическая регрессия

Зависимость дихотомических от независимых переменных можно узнать с помощью бинарной логистической регрессии. В случае с дихотомическими переменными речь идет о некотором событии, которое может произойти или не произойти. Бинарная логистическая регрессия рассчитывает вероятность наступления события в зависимости от значений независимых переменных с выводом коэффициентов регрессии для каждой такой переменной и ее статистической значимости.

Вероятность наступления бинарного события рассчитывается по формуле:

$$F(z) = P(Y = 1|X) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}, \quad (1.2)$$

где $z = b_1X_1 + b_2X_2 + \ ... + b_nX_n + a;$

X_i – значения независимых переменных;

b_i – коэффициенты, расчет которых является задачей бинарной логистической регрессии;

a – константа полученного регрессионного уравнения [22, с. 20].

Можно сделать предположение о том, что событие не наступит в том случае, если полученная вероятность меньшее 0,5, в противном случае – наступление события.

Следует не интерпретировать коэффициенты в полученной регрессии как эффект от изменения X. Для того, чтобы это верно трактовать необходимо найти производную логистической функции по параметру X и вычислить предельный эффект при конкретном значении переменной X.

Экспоненты коэффициентов логистической регрессии с учетом 95% доверительного интервала используются как отношения шансов в качестве оценки вероятности наступления изучаемого бинарного события по представляемой переменной в совокупности всех представленных статистически значимых переменных [22].

От выборки зависит точность результатов расчета логистической регрессии. Именно поэтому созданная модель требует проверки ее адекватности.

Один из простых методов оценки адекватности модели является сравнение полученных данных с предварительно используемыми исходами и проверка этой модели на исходных данных. Оценка выражается как процент наблюдений, верно предсказанных с помощью модели регрессии.

Статистика Вальда, использующая распределение хи-квадрат и представляющая собой квадрат отношения соответствующего коэффициента к его стандартной ошибке проводится для проверки значимости отличия коэффициентов от 0.

Также статистика Вальда является критерием проверки значимости коэффициентов регрессии. Это свойство основано на свойстве

асимптотически нормальных свойствах оценок максимального правдоподобия и при этом используется следующая формула:

$$W = \beta * \frac{1}{Var(\beta)} * \beta, \quad (1.3)$$

где β обозначает оценку параметра;

$Var(\beta)$ соответствует асимптотическому значению дисперсии оценки параметра [5].

Вычисление статистики Вальда показано во многих модулях (где используется метод максимального правдоподобия), например, в модулях нелинейное оценивание, временные ряды.

Функция подобия оценивает качество приближения регрессионной модели к гипотетически реальной. Мерой правдоподобия служит отрицательное удвоенное значение логарифма этой функции ($-2LL$). В качестве начального значения для $-2LL$ применяется значение, которое получается из регрессионной модели, содержащей только константы. Эта величина имеет распределение Хи-квадрат Пирсона и показывает уровень согласованности модели регрессии со всеми независимыми переменными [22, с. 21].

1.3.4 Мультиномиальная логистическая регрессия

Следующий метод является одним из вариантов логистической регрессии. В нём зависимая переменная имеет больше двух категорий, а не является дихотомической. При бинарной логистической регрессии независимая переменная может иметь интервальную шкалу. А мультиномиальная логистическая регрессия подходит только для категориальных независимых переменных, причем имеет значение, относятся ли они к шкале наименований или к порядковой шкале.

Для построения необходимо сформировать n недублированных логитов для $n + 1$ возможных значений независимой переменной, причем одна

категория используется как эталонная, ее коэффициенты принимаются равными 0:

$$g1 = ln\frac{p1}{pn} = b_{10} + b_{11} + \cdots + b_{1(n-2)},$$
$$g2 = ln\frac{p2}{pn} = b_{20} + b_{21} + \cdots + b_{2(n-2)}, \quad (1.4)$$
$$gn = 0.$$

Нахождение коэффициентов $b_{10}, b_{11}, b_{20}, \text{и } b_{21}$ является основной задачей мультиномиальной логистической регрессии. Первая цифра индекса указывает на номер логита, а вторая на порядковый номер коэффициента в данном логите, причем цифра 0 на второй позиции индекса означает константу, за которой далее следует ровно столько коэффициентов, сколько независимых переменных (факторов) взято в рассмотрение. Коэффициентам последней (эталонной) категории присваивается значение 0.

Получив значение для не дублирующих логитов, можно рассчитать значения дублирующихся логитов, используя правила вычисления логарифма:

$$ln\frac{p1}{p2} = ln\frac{p1}{pn} - ln\frac{p2}{pn}. \quad (1.5)$$

Для каждой i-ой категории зависимых переменных эта вероятность может быть вычислена по следующей формуле:

$$p(i - te\ Kategorie) = \frac{\exp(g_i)}{\sum_{k=1}^{n}\exp(g_k)}. \quad (1.6)$$

В случае наличия лишь одной независимой переменной проведение расчета с применением столь громоздкого метода является достаточно бессмысленным – все соотношения могут быть выяснены проще, при помощи таблиц сопряженности [15, с. 69].

1.3.5 Регрессия Кокса

Регрессия Кокса (модель пропорциональных рисков) – математическое представление и графическое построение в виде коэффициентов регрессионного уравнения, экспонент коэффициентов (отношения шансов)

риска наступления события как функции, зависящей от времени, и оценки влияния каждой из независимых переменных на этот риск.

Риск наступления события измеряет правдоподобие наступления события в ближайшем будущем для тех, кто еще находится в группе риска. Риск наступления события равен предельному значению условий вероятности наступления события во временном промежутке $[t, t + dt]$ для объектов, еще оставшихся в группе риска на момент времени t, деленному на длину временного интервала dt.

Метод Кокса основан на следующих положениях. Логарифмическая линейность. Все объясняющие переменные влияют линейно на логарифм функции риска наступления события.

Независимость объясняющих переменных. Все объясняющие переменные независимы. В случае присутствия взаимного влияния некоторых регрессоров, в модель должны быть дополнительно включены функции их взаимодействия.

Пропорциональность рисков. Риски наступления события для любых двух объектов пропорциональны, и коэффициент пропорциональности не зависит от времени.

На основании этих предположений выбрана функциональная форма модели: регрессия Кокса предполагает, что риск наступления события для i-того индивида имеет вид:

$$lnh_i = lnh_0(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \qquad (1.7)$$

где $h_0(t)$ – «базовый» риск, общий для всех индивидов;

$X_1, \dots, X_p$ – независимые переменные, регрессоры;

$\beta_1, \dots, \beta_p$ – соответствующие коэффициенты.

Базовый риск $h_0(t)$ – риск наступления события для объекта из референтной группы (для которого все независимые переменные равны нулю).

Коэффициенты $\beta_1, \dots, \beta_p$ отражают влияние каждой из независимых переменных (регрессоров) на функцию риска: при увеличении X_j на единицу

и фиксированных значениях остальных регрессоров, риск наступления события возрастает в e^{β_j} раз.

Метод Кокса не рассматривает зависимость риска от времени. Последнее происходит из предположения о пропорциональности рисков. Чтобы ослабить это предположение, используются ковариванты, зависящие от времени.

1.4 Снижение размерности

Статистически не значимые связи имеются во многих исследованиях. В медицинских исследованиях существует большое количество связей, поэтому невозможно сделать точный прогноз, сколько будет значимых для цели исследования. От всех возможных связей из оказывается 5–25%.

1.4.1 Факторный анализ

Факторный анализ – это процедура, с помощью которой большое количество переменных сводят к меньшему числу независимых влияющих величин, называемых факторами (факторными комплексами, компонентами). При этом в один фактор объединяются переменные, сильно коррелирующие между собой. Переменные из разных факторов слабо коррелируют между собой. Таким образом, целью факторного анализа является нахождение таких комплексных факторов, которые как можно более полно объясняют наблюдаемые связи между переменными, имеющимися в наличии.

Сильная зависимость между двумя разными переменными дает понять об избыточности двух пунктов исследования. Ее между переменными можно обнаружить с помощью диаграммы рассеяния. Полученная линия регрессии путем аппроксимации (подгонки) дает графическое представление зависимости. Переменная будет включить в себя наиболее существенные черты обеих переменных, если определить новую переменную на основе линии регрессии, изображенной на этой диаграмме. Новый фактор в

действительности является линейной комбинацией двух исходных переменных.

Выделение главных компонентов в факторном анализе проводится по диаграмме рассеяния изучаемых переменных. Процедура выделения главных компонентов подобна вращению, доводящему до максимума исходного пространства переменных. Например, на диаграмме рассеяния вы можете рассматривать линию регрессии как ось X, повернув ее так, что она совпадает с прямой регрессии. Этот тип вращения называется вращением, максимизирующим дисперсию.

После нахождения линии, для которой дисперсия максимальна, вокруг нее остается некоторый разброс данных, затем происходит повтор процедуры. В анализе главных компонентов именно так и делается: после того, как первый фактор выделен, то есть после того, как первая линия проведена, определяется следующая линия, максимизирующая остаточную вариацию (разброс данных вокруг первой прямой), и т. д. Таким образом, факторы последовательно выделяются один за другим. Так как каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, то факторы оказываются независимыми друг от друга. Иными словами, некоррелированными или ортогональными.

Критерий Кайзера и критерий каменистой осыпи являются рекомендациями, на основании которых принимается решение об остановке процедуры выделения.

Критерий Кайзера. Сначала можно отобрать только факторы с собственными значениями большими 1. Если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser, 1960), и является наиболее широко используемым.

Критерий каменистой осыпи. Данный критерий является графическим методом, впервые предложенным Кэттелем (Cattell, 1966). Собственные значения (факторную нагрузку) обычно изображают в виде простого графика

(рисунок 1.3). Компонентополагающим числом на графике является место, где убывание собственных значений слева направо представляет собою крутой склон, расстояние между точками примерно равно единице и собственное значение компонента больше 1. Незначимые компоненты находятся далее на максимально замедленной части кривой, расстояние между точками меньше единицы и собственное значение компонента меньше единицы, так называемый «щебень».

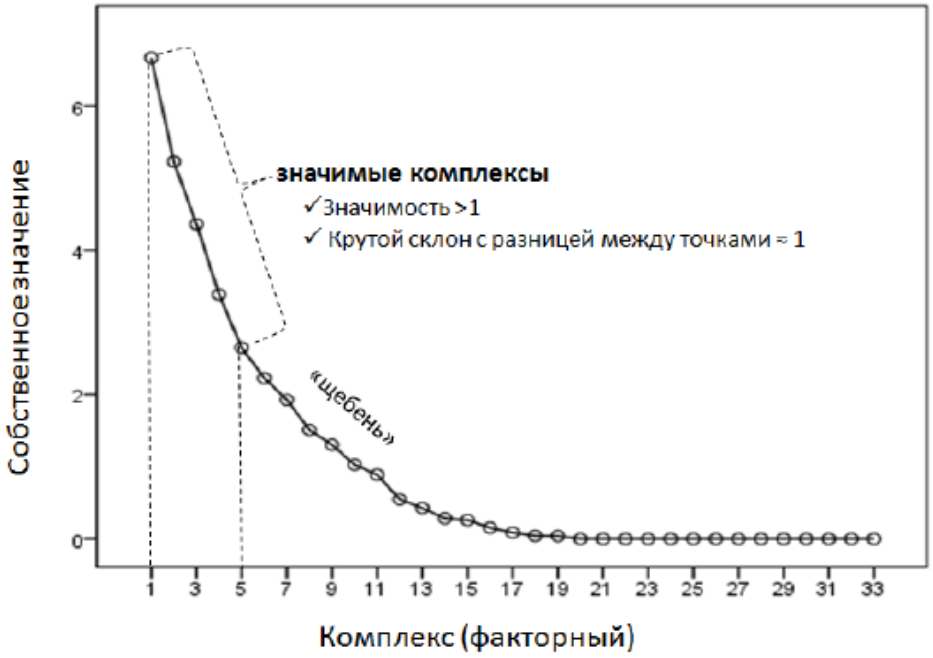

Рисунок 1.3 – Точечная диаграмма значимости факторных моделей

1.5 Выбор среды разработки и языка программирования

На данный момент специалисты МАУЗ ГКБ № 2 хранят информацию о пациентах в специальной разработанной информационной системе. Данная система написана на языке программирования JavaScript с использованием технологии NodeJS, поэтому для разработки модуля необходимо использовать именно этот язык программирования.

В информационной системе МАУЗ ГКБ № 2 используется СУБД PostgreSQL, поэтому для создания модуля статистического анализа потока пациентов необходимо также использовать библиотеки PostgreSQL node-postgres. Среда разработки Visual Studio Code.

1.6 Выводы по разделу

После анализа предметной области был выбран язык программирования JavaScript, для хранения данных будет использована система PostgreSQL и среда разработки Visual Studio Code. Также уточнена цель: разработать модуль статистического анализа потока пациентов в информационной системе МАУЗ ГКБ № 2.

Для достижения данной цели необходимо решить следующие задачи: разработать математическую модель; разработать алгоритм работы программы; выполнить программную реализацию; протестировать программу на экспериментальных данных.

2 МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СТАТИСТИЧЕСКОГО АНАЛИЗА ПОТОКА ПАЦИЕНТОВ МАУЗ ГКБ № 2

2.1 Дисперсионный анализ для сравнения нескольких групп

Исследование значимости различия между средними (для групп или переменных) – это главная цель дисперсионного анализа. При помощи разбиения суммы квадратов на компоненты проводится проверка. Общая дисперсия разбивается на части, одна связана с различием средних значений, а вторая обусловлена случайной ошибкой. Для анализа статистической значимости различия между средними значениями используется последняя компонента дисперсии.

Статистические методы (критерии значимости) используют для описания данных и для оценки статистической значимости результатов проведенных исследований. Методы статистической обработки представлены на рисунке 2.1. Алгоритм построения данных критериев:

1) формируется нулевая гипотеза – предположение о том, что исследуемые факторы не влияют на исследуемую величину и полученные различия случайны;

2) вычисляется вероятность получения наблюдаемых различий при условии справедливости нулевой гипотезы;

3) при малой полученной вероятности (пункт 2) отвергается нулевая гипотеза, и делается вывод: результаты эксперимента статистически значимы.

То есть, необходимо решить вопрос о случайности выявленных различий, от этого зависит принятие решения о том, является ли выявленные различия свидетельством различного состояния и/или свидетельством эффекта от вмешательства. Количественную характеристику случайности представляет теория вероятностей в виде p-значения. Чем это значение больше, тем больше вероятность отсутствия различий в пользу нулевой гипотезы, и чем оно меньше, тем больше вероятность наличия различий в пользу альтернативной гипотезы.

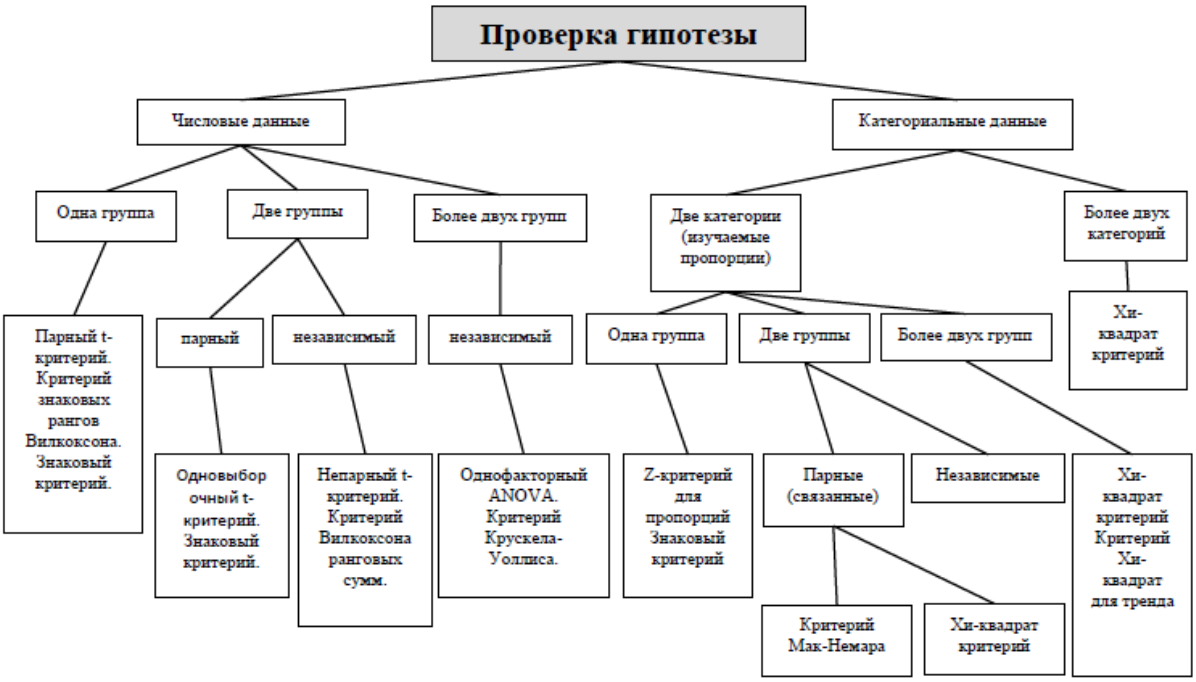

Рисунок 2.1 – Методология индуктивной статистической обработки исследования

Для выявления различных критериев оценки прохождения того или иного медицинского осмотра пациентами проводилось несколько исследований. Чтобы понять основные направления движения пациентов был проведен анализ потока, который позволил разделить посещения по цели визита. Таким образом, выделены четыре составляющие: здоровый поток, острые больные, больные на повторный прием, профпоток. Данные потоки представлены на рисунке 2.2.

Пациенты часто сталкиваются с различными проблемами во время прохождения медицинского осмотра и наблюдение за четырьмя потоками позволит понять и определить наиболее острые проблемы.

Для разделения данных потоков поликлиники часто делают различные входы в медорганизацию. Это позволяет также изолировать заболевших людей от здоровых.

Кроме новых входов, можно поменять расположение кабинетов. В каждом потоке пациенты затрачивают много времени на лишние перемещения, ведь зачастую приходится ходить по разным этажам. Чтобы

реорганизовать пространство, выявляют, куда может пойти пациент из одного потока, и стараются сократить передвижения.

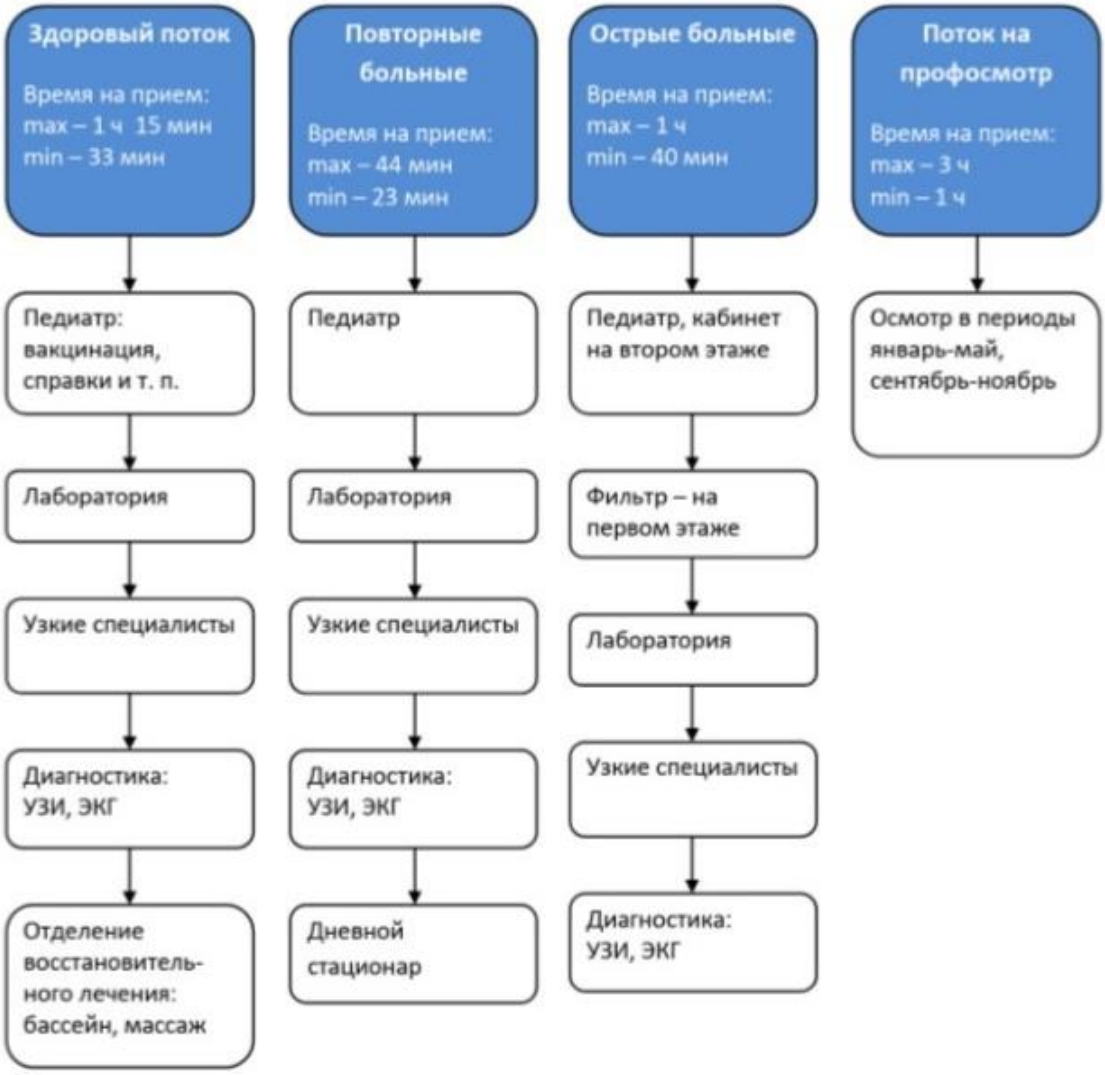

Рисунок 2.2 – Разделение пациентов на четыре потока

Исследование было произведено на основании данных потока на медицинский осмотр. Для того, чтобы понять различия по времени прохождения той или иной медицинской услуги было произведено разделение по полу (мужской и женский пол). Известно, что женщины в среднем проходят медицинские осмотры дольше. Для проверки данной гипотезы определим мужчин в первую группу (рисунок 2.3), а женщин во вторую группу (рисунок 2.4). В группах было по 36 больных.

Средняя длительность приема пациентов составила для первой группы 4,51 минут (стандартное отклонение 1,98 мин.), для второй группы 6,28 минут (стандартное отклонение 2,54 мин.).

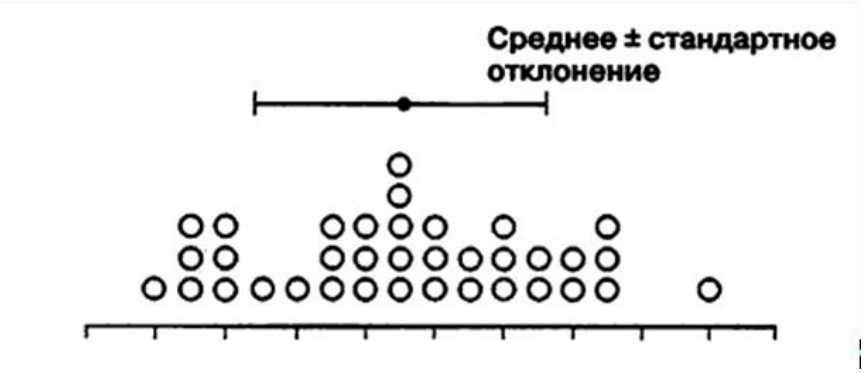

Рисунок 2.3 – Результат длительности медицинского осмотра в 1 группе (каждый пациент обозначен кружком, положение кружка – время прохождения осмотра), n_1=36, $\overline{x_1} = 4.51$, s_1=1.98

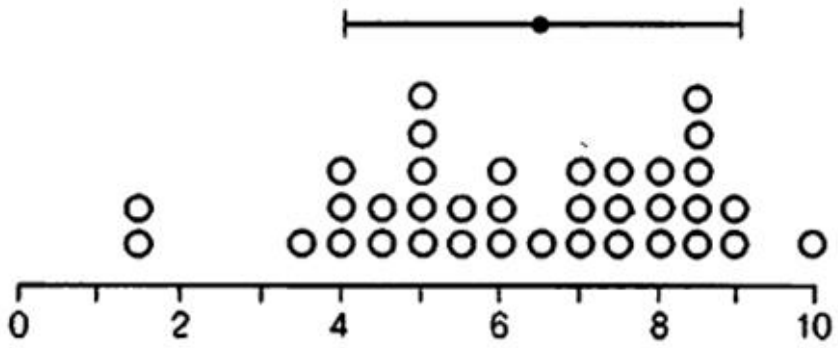

Рисунок 2.4 – Результат длительности медицинского осмотра во 2 группе, $n_1 = 36,\ \overline{x_2} = 6.28,\ s_2 = 2.54$

С помощью дисперсионного анализа проверим гипотезу $H0$: можно ли считать эти различия случайными.

Вычислим сначала внутригрупповую дисперсию как среднюю дисперсий обеих групп:

$$s^2_{\text{вну}} = \frac{1}{2}(s_1^2 + s_2^2) = \frac{1}{2}((1{,}98)^2 + (2{,}54)^2) = 5{,}19. \quad (2.1)$$

Вычислим межгрупповую дисперсию. Среднее двух выборочных средних равно

$$\bar{x} = \frac{1}{2}(\overline{x_1} + \overline{x_2}) = \frac{1}{2}(4.51 + 6.28) = 5.40, \quad (2.2)$$

следовательно, стандартное отклонение равно

$$s_{\bar{x}} = \sqrt{\frac{(\overline{x_1} - \bar{x})^2 + (\overline{x_2} - \bar{x})^2}{m - 1}} = \tag{2.3}$$

$$= \sqrt{\frac{(4.51 - 5.40)^2 + (6.28 - 5.40)^2}{2 - 1}} = 1.25$$

и межгрупповая дисперсия равна

$$s^2_{\text{меж}} = n s^2_{\bar{x}} = 36 \cdot (1.25)^2 = 56.25. \tag{2.4}$$

Эти два вида дисперсий нужны для того, чтобы с их помощью делать расчет статистического критерия F, который как раз и позволяет дать научно обоснованный ответ на вопрос: являются ли различия между группами достоверными и насколько достоверными.

Теперь вычислим критерий F. Его суть заключается в том, что он сравнивает две дисперсии: межгрупповую и внутригрупповую, поэтому их соотношение называют F-отношением:

$$F = \frac{\text{Дисперсия совокупности, оцененная по выборочным средним}}{\text{Дисперсия совокупности, оцененная по выборочным дисперсиям}} = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}}, \tag{2.5}$$

$$F = \frac{56{,}25}{5{,}19} = 10{,}84.$$

Критическое значение F однозначно определяется уровнем значимости (максимально приемлемая вероятность отвергнуть верную нулевую гипотезу, обозначается α, обычно $\alpha = 0{,}05$) и еще двумя параметрами: внутригрупповое число степеней свободы и межгрупповое (обозначается ν). Межгрупповое число степеней свободы ($\nu_{\text{меж}}$) вычисляется как разность между числом групп и единицы, то есть$\nu_{\text{меж}} = m - 1$. Внутригрупповое число степеней свободы – это произведение числа групп (m) на численность каждой из группы (n) минус единица: $\nu_{\text{вну}} = m(n - 1)$.

Теперь найдем межгрупповое и внутригрупповое число степеней свободы $\nu_{\text{меж}} = m - 1 = 2 - 1 = 1$, $\nu_{\text{вну}} = m(n - 1) = 2(36 - 1) = 70$. В таблице критических значений F определим, чему равно F (для $\alpha = 0{,}05$

шрифт обычный и для $\alpha = 0{,}01$ шрифт жирный). На пересечении первого столбца и строки «70» находится число 7,01 (жирный шрифт), то есть при уровне значимости 0,01 критическое значение $F = 7{,}01$.

Значит, ответ на наш вопрос (можно ли считать различия в длительности медицинского осмотра случайными): вероятность этого весьма мала, меньше 1%. Пациенты мужского пола находились в больнице меньше, чем женщины, различия эти статистически значимы.

Известно, что в пожилом возрасте анатомо-физиологические системы человека претерпевают более или менее значительные изменения. По мере старения меняются социальное положение человека и образ жизни, ухудшаются самочувствие и состояние здоровья. Человек с трудом приспосабливается к возрастным ограничениям. Следовательно люди пожилого возраста проходят гораздо чаще медицинские осмотры по сравнению с людьми молодого возраста и тратят на это больше времени. Необходимо узнать влияет ли возраст пациентов на время прохождения медицинского осмотра. В эксперимент вошли 78 пациентов, разделенных на 3 группы по 26 человек в каждой. Первая группа (контрольная группа А) состояла из пациентов возрастом от 45 до 60 лет. Во вторую группу (группа В) входили пациенты возрастом от 31 года до 45 лет. Пациенты третьей группы (группа С) люди возраста от 18 до 30 лет.

В группе А среднее число длительности осмотра у врача составило 11,5 минут, у В – 10,1, а у С – 9,1. Проверим, можно ли эти различия отнести за счет случайности.

Дисперсия совокупности по среднему выборочных дисперсий составляет:

$$s^2_{\text{вну}} = \frac{1}{3}(s_1^2 + s_2^2 + s_3^2) = \frac{1}{3}((1.3)^2 + (2.1)^2 + (2.4)^2) = 3.95, \qquad (2.6)$$

где s_1^2, s_2^2 и s_3^2 – это выборочные оценки дисперсии в группах. Дисперсия внутри каждой группы вычисляется относительно среднего для группы.

Для оценки дисперсии по разбросу выборочных средних необходимо оценить стандартную ошибку среднего, для чего вычислить стандартное отклонение среднего 3 выборок. Среднее трех средних величин равно

$$\bar{x} = \frac{1}{3}(\overline{x_1} + \overline{x_2} + \overline{x_3}) = \frac{1}{3}(11.5 + 10.1 + 9.1) = 10.2. \quad (2.7)$$

Следовательно, оценка стандартной ошибки:

$$s_{\bar{x}} = \sqrt{\frac{(\overline{x_1} - \bar{x})^2 + (\overline{x_2} - \bar{x})^2 + (\overline{x_3} - \bar{x})^2}{m - 1}} = \quad (2.8)$$

$$= \sqrt{\frac{(11.5 - 10.2)^2 + (10.1 - 10.2)^2 + (9.1 - 10.2)^2}{3 - 1}} = 1.2.$$

Объем выборки $n = 26$, значит, оценка дисперсии по разбросу средних дает величину:

$$s^2_{\text{меж}} = n \cdot s_{\bar{x}} = 26 \cdot (1.2)^2 = 37.44 \quad (2.9)$$

Если верна нулевая гипотеза, то межгрупповая и внутригрупповая дисперсии являются оценками одной и той же дисперсии, они приближенно равны. Используя это, вычислим критерий F:

$$F = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}}. \quad (2.10)$$

После вычисления F получаем, что $F = \frac{37.44}{3.95} = 9.48$.

В нашем случае число степеней свободы $\nu_{\text{меж}} = m - 1 = 3 - 1 = 2$, $\nu_{\text{вну}} = m(n - 1) = 3 \cdot (26 - 1) = 75$. По таблице находим критическое значение $F = 4{,}90$ при 1% уровне значимости. Следовательно, различия между группами статистически значимы: вероятность случайно получить такие различия не превышает 1%.

2.2 Сравнение двух групп: критерий Стьюдента с поправкой Бонферрони

Благодаря дисперсионному анализу можно проверять значимость различий нескольких групп, но часто нужно проводить сравнения только двух

групп. В этом случае применяется критерий Стьюдента – частный случай дисперсионного анализа.

Существует три правила использования данного критерия:

1) критерий Стьюдента используется для проверки гипотезы о различии средних только двух групп;

2) если эксперимент предполагает большее число групп, то необходимо воспользоваться дисперсионным анализом;

3) если с помощью критерия Стьюдента проверялись различия между несколькими группами, то верный уровень значимости можно получить, умножив уровень значимости (введенным ранее α) на число возможных сравнений.

Несмотря на то, что критерий предназначен для сравнения двух групп, на практике он широко используется для оценки различий большего числа групп благодаря попарного их сравнения. При этом применяется эффект множественных сравнений. Суть которого заключается в том, что при частом применении критерия вероятность ошибочно найти различия там, где их нет, возрастает.

Следует воспользоваться дисперсионным анализом, если количество исследуемых групп больше двух. Но он проверяет лишь гипотезу о равенстве всех средних. Если гипотеза не подтверждается, то нельзя выяснить точно какая группа отличается от других.

Методы множественного сравнения позволяют это сделать. Простейшим из этих методов сравнения является введение поправки Бонферрони.

С помощью дисперсионного анализа выяснилось, что время медицинского осмотра пациентов в группах A (контрольная), B и C статистически значимы. Однако межгрупповые различия не были определены. Используя критерий Стьюдента с поправкой Бонферрони, сравним три группы A, B и C.

Внутригрупповая оценка дисперсии $s^2_{\text{вну}} = 3{,}95$. Число групп $m = 3$, численность каждой группы $n = 26$. Значит число степеней свободы $\nu = m(n-1) = 3 \cdot (26-1) = 75$. Проведем попарное сравнение трех групп.

Для формализации рассмотрим отношение:

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}. \quad (2.11)$$

Это отношение будет близко к нулю для двух случайных выборок, извлеченных из одной нормально распределенной совокупности. Чем меньше t, тем больше возрастает вероятность нулевой гипотезы, а чем больше t, тем больше причин отвергнуть нулевую гипотезу и считать различия статистически значимыми.

Точность выборочной оценки среднего характеризуется стандартной ошибкой среднего:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \quad (2.12)$$

где n – объем выборки, а σ – стандартное отклонение совокупности, из которой извлечена выборка.

Дисперсия суммы двух случайно извлеченных значений равна сумме дисперсий совокупностей, из которых они извлечены. Отсюда выведем формулу для стандартной ошибки среднего. Пусть случайным образом извлекли n значений из совокупности, имеющей стандартное отклонение σ. Выборочное среднее равно

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), \quad (2.13)$$

$$n\bar{x} = x_1 + x_2 + \cdots + x_n. \quad (2.14)$$

Так как дисперсия каждого из x_i равна σ^2, дисперсия величины $n\bar{x}$ составит

$$\sigma^2_{n\bar{x}} = \sigma^2 + \cdots + \sigma^2 = n\sigma^2, \quad (2.15)$$

а стандартное отклонение

$$\sigma_{n\bar{x}} = \sqrt{n}\sigma. \quad (2.16)$$

Стандартное отклонение среднего $\bar{x}$ (тождественно равного $n\bar{x}/n$):

$$\sigma_{\bar{x}} = \frac{\sigma_{n\bar{x}}}{n} = \sqrt{n}\frac{\sigma}{n} = \frac{\sigma}{\sqrt{n}}. \quad (2.17)$$

Следовательно, получили формулу для стандартной ошибки среднего.

На основе доказанного можно сделать вывод, что дисперсия разности двух случайно извлеченных значений равна сумме дисперсий совокупностей, из которых они извлечены:

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2. \quad (2.18)$$

Заменим в данной формуле дисперсии на выборочные оценки, чтобы оценить дисперсию разности членов двух совокупностей:

$$s_{x-y}^2 = s_x^2 + s_y^2. \quad (2.19)$$

Данная формула пригодна для оценки стандартной ошибки разности выборочных средних, так как это стандартное отклонение совокупности средних значений всех выборок объемом n. Значит,

$$s_{\bar{x}-\bar{y}}^2 = s_{\bar{x}}^2 + s_{\bar{y}}^2, \quad (2.20)$$

а искомая стандартная ошибка разности средних

$$s_{\bar{x}-\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}. \quad (2.21)$$

Воспользовавшись выведенными формулами, можно найти отношение t. Так как

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}, \quad (2.22)$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_{\bar{x}1}^2 + s_{\bar{x}2}^2}}.$$

Если выразить ошибку среднего через выборочное стандартное отклонение, то получится иная запись формулы

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} \quad (2.23)$$

где n – это объем выборки.

Выборочные дисперсии (s_1^2 и s_2^2) дают один и тот же результат, если обе выборки извлечены из одной совокупности. Они являются оценками одной и той же дисперсии σ^2, поэтому их следует заменить на объединенную оценку дисперсии. Объединенная оценка дисперсии для выборок одинакового объема вычисляется по следующей формуле:

$$s^2 = \frac{s_1^2 + s_2^2}{2}. \tag{2.24}$$

Следовательно, меняется вычисление значения t, получаем формулу

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n} + \frac{s^2}{n}}}. \tag{2.25}$$

При одинаковом объёме выборок обе формулы дают одинаковое значение при вычислении t.

Сравнивая контрольную группу А с группой В, имеем:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{2\frac{s_{\text{вну}}^2}{n}}} = \frac{10{,}1 - 11{,}5}{\sqrt{\frac{2 * 3{,}95}{26}}} = -2{,}54 \tag{2.26}$$

при сравнении контрольной группы и группы С:

$$t = \frac{\bar{x}_3 - \bar{x}_1}{\sqrt{2\frac{s_{\text{вну}}^2}{n}}} = \frac{9{,}1 - 11{,}5}{\sqrt{\frac{2 * 3{,}95}{26}}} = -4{,}35, \tag{2.27}$$

и при сравнении второй группы В с третьей группой С:

$$t = \frac{\bar{x}_2 - \bar{x}_3}{\sqrt{2\frac{s_{\text{вну}}^2}{n}}} = \frac{10{,}1 - 9{,}1}{\sqrt{\frac{2 * 3{,}95}{26}}} = 1{,}81 \tag{2.28}$$

Для того, чтобы сделать вывод из полученных результатов, необходимо рассмотреть неравенство Бонферрони:

$$\alpha' < k\alpha, \tag{2.29}$$

где α' – вероятность хотя бы один раз выявить ошибочно несоответствия.

Другими словами, α' является верным уровнем значимости много раз применённого критерия. Из приведенного выше неравенства следует, что если возникнет необходимость обеспечить вероятность ошибки α', то каждое

сравнение должно иметь уровень значимости α'/k, что и является поправкой Бонферрони.

После трех сравнений групп A, B и C, применив поправку Бонферрони, находим уровень значимости в каждом сравнении должен быть $\alpha' = 0{,}05/3 \approx 0{,}017$. По таблице критических значений t находим (так как в таблице нет значения для $\alpha' = 0{,}017$, то можно взять ближайшее меньшее значение), что при $\nu = 75$ (степень свободы) критическое значение составляет примерно 2,45.

Таким образом, можно сделать вывод, что и у пациентов группы B, и у пациентов группы C время прохождения медицинского осмотра ниже, чем в контрольной группе, но при этом не отличается друг от друга у групп B и C.

Важно отметить то, что поправка Бонферрони хорошо работает для небольшого числа сравнений. Если это число превышает 8, то метод становится слишком «строгим», и даже очень большие различия приходится признавать статистически незначимыми (способность критерия выявлять различия называется чувствительностью).

Оба метода: дисперсионный анализ и критерий Стьюдента с поправкой Бонферрони, позволили оценить вероятность нулевой гипотезы, то есть предположения об отсутствии эффекта экспериментального воздействия, во всех исследованиях. Вероятность нулевой гипотезы оценивалась с помощью критериев значимости – F, t. Гипотеза отклонялась, если значение критерия превышало критическое. Данное отклонение наблюдалось в первом и во втором эксперименте дисперсионного анализа. Также справедливо утверждалось, что найдены статистически значимые различия. Если значение критерия оказывалось меньше критического, то делался вывод об отсутствии статистически значимых различий.

2.3 Метод группировок в статистике. Ряды распределения

Распределение единиц наблюдения по группам по одному или нескольким признакам называется статистической группировкой. Эти

признаки называются группировочными. В зависимости от задач исследования строят типологические, структурные и аналитические группировки.

Типологическая группировка представляет собой распределение единиц наблюдения качественно неоднородной совокупности по социально-экономическим типам, классам, качественно однородным группам.

При структурной группировке разделение единиц однородной совокупности на группы происходит с целью выявления ее структуры по одному из признаков.

С помощью аналитической группировки определяют наличие связи между признаками и ее направление. Один из признаков является результативным, а другой – факторным. Результативный признак меняется под воздействием факторного признака.

Таблица 2.1 – Распределение количества пациентов у врачей различных профилей

Услуга	Кол-во пациентов, чел
Терапевт	3178
Нарколог	3168
Психиатр	3153
Офтальмолог	1445
Оториноларинголог	2237
Дерматовенеролог	2079
Невролог	1446
Стоматолог	1897
Акушер-гинеколог	1718

В зависимости от количества признаков, по которым проводится группировка, различают простые и сложные группировки. Если группировка проводится по одному признаку, то она называется простой. Если единицы совокупности группируются сразу по двум или более признакам, то такая

группировка называется сложной. При этом внутри групп, образованных по одному признаку, единицы совокупности подразделяются на подгруппы по другому признаку. Примером сложной группировки является группировка пациентов по двум признакам – полу и возрасту. Ее результаты представлены в таблице 2.2.

Результаты группировки собранных статистических данных, как правило, представляются в виде рядов распределения. Ряд распределения – это упорядоченное распределение единиц совокупности на группы по изучаемому признаку [14].

Таблица 2.2 – Группировка пациентов по полу и возрасту

Возраст	Мужчины	Женщины	Итого
18-24	1317	1336	2653
25-31	2891	4092	6983
32-38	4952	5521	10473
39-45	5903	5410	11313
46-52	4903	6213	11116
53-59	4398	5033	9431
60-66	1892	3238	5130

Ряды распределения делятся на атрибутивные и вариационные, в зависимости от признака, положенного в основу группировки. Если признак качественный, то ряд распределения называется атрибутивным. Если признак, по которому строится ряд распределения, количественный, то ряд называется вариационным.

Вариационный ряд распределения всегда состоит из двух частей: вариант и соответствующих им частот. Вариантой называется значение, которое может принимать признак у единиц совокупности, частотой – количество единиц наблюдения, обладающих данным значением признака. Сумма частот всегда равна объему совокупности.

В результате статистического наблюдения получены следующие данные о среднем времени прохождении медицинского осмотра пациентами (сек).

Таблица 2.3 – Среднее время прохождения медицинского осмотра пациентами

310	545	616	718	982	576	755	851	674	610
392	809	551	478	565	370	492	552	422	532
424	463	477	601	428	605	538	839	319	381
460	508	572	330	452	405	607	638	345	301
598	584	551	364	446	358	337	517	892	749

Просматривать такой массив данных крайне неудобно, кроме того, не видно закономерностей изменения показателя. Построим интервальный ряд распределения.

Для начала определим число интервалов по формуле Стёрджеса:

$$n = 1 + 3{,}322lgN, \tag{2.30}$$

где n – число интервалов;

N – объём совокупности (число единиц наблюдения).

Таким образом, получим интервальный ряд распределения времени прохождения медицинского осмотра

$$n = 1 + 3{,}322lgN = 1 + 3{,}322lg = 6{,}6. \tag{2.31}$$

Дробное число, характеризующее количество интервалов, желательно округлять в большую сторону. Получим число интервалов $n = 7$.

Определим величину интервалов по формуле:

$$i = \frac{x_{max} - x_{min}}{n}, \tag{2.32}$$

где x_{max} – максимальное значение признака;

x_{min} – минимальное значение признака.

$$i = \frac{982 - 301}{7} = 97{,}3. \tag{2.33}$$

Интервалы вариационного ряда наглядны, если их границы имеют «круглые» значения, поэтому округлим величину интервала 97,3 до 97,0.

Определим границы интервалов.

Интервалы, как правило, записывают таким образом, чтобы верхняя граница одного интервала являлась одновременно нижней границей следующего интервала. Так, для нашего примера получим: 301–398; 398–495; 495–592; 592–689; 689–786; 786–883; 883–982.

По исходным данным построим ранжированный ряд. Для этого запишем в порядке возрастания значения, которые принимает признак (таблица 2.4).

Таблица 2.4 – Ранжированный ряд

301	392	478	565	638
310	405	492	572	674
319	422	508	576	718
330	424	517	584	749
337	428	532	598	755
345	446	538	601	809
358	452	545	605	839
364	460	551	607	851
370	463	551	610	892
381	477	552	616	982

Подсчитаем частоты. При подсчете частот может возникнуть ситуация, когда значение признака попадет на границу какого-либо интервала. В таком случае можно руководствоваться правилом: данная единица приписывается к тому интервалу, для которого ее значение является верхней границей. Так, значение 5,0 будет относиться ко второму интервалу.

Результаты группировки, оформим в таблице 2.5.

В последней графе таблицы представлены накопленные частоты, которые получают путем последовательного суммирования частот, начиная с первой. Накопленная частота, например, 34, показывает, что у 34 пациентов время прохождения одного специалиста не превышает 592 секунды (верхняя граница соответствующего интервала).

Таблица 2.5 Результаты группировки

Промежуток времени, сек	Кол-во человек, ед. (частоты)	Накопленные частоты
301-398	11	11
398-495	11	22
495-592	12	34
592-689	8	42
689-786	3	45
786-883	3	48
883-982	2	50
Итого:	50	-

2.4 Выводы по разделу

После проведения анализа предметной области были выбраны математические методы для проведения статистического анализа потока пациентов в МАУЗ ГКБ № 2, метод оценки статистической значимости различий (дисперсионный анализ), метод сравнения двух групп: критерий Стьюдента с поправкой Бонферрони.А также с помощью метода группировок и с помощью рядов распределения рассчитаны вероятности нахождения анализов в различных состояниях для каждого пациента.

Дисперсионный анализ и критерий Стьюдента с поправкой Бонферрони, позволили оценить вероятность нулевой гипотезы, то есть предположения об отсутствии эффекта экспериментального воздействия, во всех исследованиях. Вероятность нулевой гипотезы оценивалась с помощью критериев значимости – F, t. Гипотеза отклонялась, если значение критерия превышало критическое. Данное отклонение наблюдалось в первом и во втором эксперименте дисперсионного анализа. Также справедливо утверждалось, что найдены статистически значимые различия.

3 ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МОДУЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ПРОХОЖДЕНИЯ МЕДИЦИНСКОГО ОСМОТРА ПАЦИЕНТАМИ МАУЗ ГКБ № 2

3.1 База данных пациентов МАУЗ ГКБ № 2

Для разработки данного модуля была использована информация из существующей базы данных МАУЗ ГКБ № 2 (рисунок 3.1).

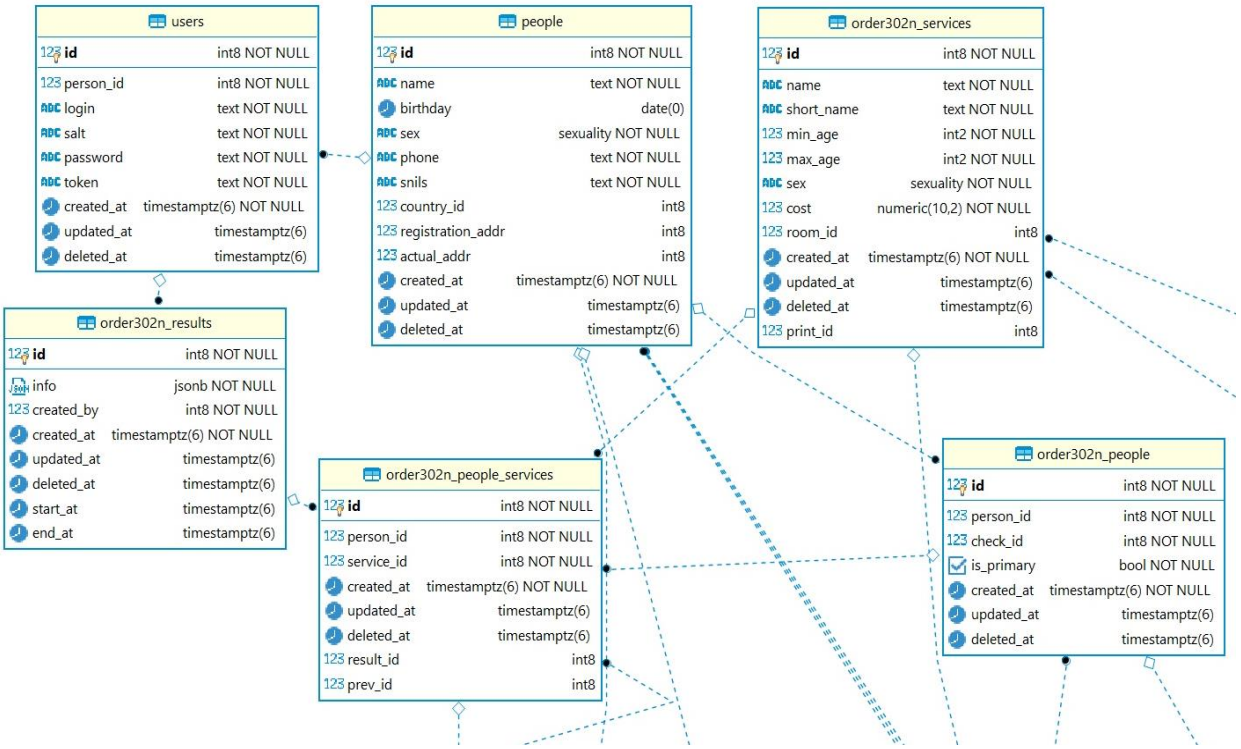

Рисунок 3.1 – База данных МАУЗ ГКБ № 2

Рассмотрим таблицы, представленные в базе данных подробнее.

Таблица «users» содержит информацию о пользователях системы. Поля таблицы описаны в таблице 3.1.

Таблица 3.1 – Описание таблицы «users»

Наименование	Тип	Комментарий
id	int8	Идентификатор
person_id	int8	Идентификатор человека
login	text	Логин
salt	text	Соль для пароля
password	text	Хэш пароля

Продолжение таблицы 3.2 – Описание таблицы «users»

Наименование	Тип	Комментарий
token	text	Ключ авторизации
created_at	timestamptz(6)	Дата и время создания
updated_at	timestamptz(6)	Дата и время обновления
deleted_at	timestamptz(6)	Дата и время удаления

Таблица «people» хранит персональные данные пациентов и сотрудников (таблица 3.2). Она будет использоваться модулем для получения ФИО, даты рождения и пола пациента.

Таблица 3.2 – Описание таблицы «people»

Наименование	Тип	Комментарий
id	int8	Идентификатор
name	text	ФИО пациента
birthday	date(0)	Дата рождения
sex	sexuality	Пол
phone	text	Номер телефона
snils	text	Номер СНИЛС
country_id	int8	Идентификатор страны
registration_addr	int8	Адрес регистрации
actual_addr	int8	Фактический адрес
created_at	timestamptz(6)	Дата и время создания
updated_at	timestamptz(6)	Дата и время обновления
deleted_at	timestamptz(6)	Дата и время удаления

Таблица «order302n_services» хранит список всех услуг (таблица 3.3). Она содержит данные о названии, фильтрах, применяемых при добавлении к пациентам, стоимости услуги, идентификаторы кабинета и печатной формы. Разрабатываемая программа получает из этой таблицы названия услуг.

Таблица 3.3 – Описание таблицы «order302n_services»

Наименование	Тип	Комментарий
id	int8	Идентификатор
name	text	Название услуги
short_name	text	Краткое название услуги
min_age	int2	Минимальный возраст
max_age	int2	Максимальный возраст
sex	sexuality	Пол
cost	numeric(10,2)	Стоимость
room_id	int8	Идентификатор кабинета
created_at	timestamptz(6)	Дата и время создания
updated_at	timestamptz(6)	Дата и время обновления
deleted_at	timestamptz(6)	Дата и время удаления
print_id	int8	Идентификатор печатной формы

Таблица «order302n_people» описывает связь между конкретным медосмотром и пациентом, и содержит информацию о том, является ли этот медосмотр для пациента первичным (таблица 3.4). Эта таблица используется модулем для выделения пациентов из всего множества людей в базе.

Таблица 3.4 – Описание таблицы «order302n_people»

Наименование	Тип	Комментарий
id	int8	Идентификатор
person_id	int8	Идентификатор человека
check_id	int8	Идентификатор медосмотра
is_primary	bool	Является ли осмотр первичным
created_at	timestamptz(6)	Дата и время создания
updated_at	timestamptz(6)	Дата и время обновления
deleted_at	timestamptz(6)	Дата и время удаления

Таблица «order302n_people_services» связывает пациента с услугами, которые ему необходимо пройти (таблица 3.5). Кроме того, эта таблица хранит идентификатор результата осмотра или обследования. Из этой таблицы модуль получает список услуг для каждого пациента.

Таблица 3.5 – Описание таблицы «order302n_people_services»

Наименование	Тип	Комментарий
id	int8	Идентификатор
person_id	int8	Идентификатор человека
service_id	int8	Идентификатор услуги
created_at	timestamptz(6)	Дата и время создания
updated_id	timestamptz(6)	Дата и время обновления
deleted_id	timestamptz(6)	Дата и время удаления
result_id	int8	Идентификатор результатов осмотра или обследования

Таблица «order302n_results» содержит данные об осмотре или обследовании пациентов (таблица 3.6).

Таблица 3.6 – Описание таблицы «order302n_results»

Наименование	Тип	Комментарий
id	int8	Идентификатор
info	jsonb	Значения полей форм осмотров и обследований
created_by	int8	Идентификатор пользователя-создателя
created_at	timestamptz(6)	Дата и время создания
updated_at	timestamptz(6)	Дата и время обновления
deleted_at	timestamptz(6)	Дата и время удаления
start_at	timestamptz(6)	Время начала приёма
end_at	timestamptz(6)	Время окончания приёма

Для работы возьмём из таблицы order302n_results столбцы start_at и end_at, из таблицы people – sex и birthday, из таблицы order302n_services столбцы id и name. Выполним запрос, представленный на рисунке 3.2.

```
SELECT onr.start_at, onr.end_at, p.birthday, p.sex, ons.id, ons.name
FROM order302n_results onr, order302n_people_services onps,
 order302n_services ons, people p, order302n_people onp
WHERE onr.id = onps.result_id AND onps.service_id = ons.id
 AND onps.person_id = onp.id AND onp.person_id = p.id
```

Рисунок 3.2 – Форма аутентификации

В качестве кэширующей таблицы для работы модуля создадим следующую таблицу, с помощью запроса, представленного на рисунке 3.3.

```
create table if not exists order302n_stats (
id BIGSERIAL PRIMARY key,
service_id BIGINT NOT NULL REFERENCES order302n_services(id),
age SMALLINT not null,
sex sexuality not null,
dur smallint not null,
created_at TIMESTAMPTZ NOT NULL DEFAULT NOW(),
updated_at TIMESTAMPTZ,
deleted_at TIMESTAMPTZ
)
```

Рисунок 3.3 – Создание кэширующей таблицы

В ней будет храниться продолжительность осмотра или обследования (таблица 3.7).

Таблица 3.7 – Описание таблицы «order302n_stats»

Наименование	Тип	Комментарий
id	int8	Идентификатор
service_id	int8	Идентификатор услуги
age	int2	Возраст
sex	sexuality	Пол
dur	int2	Продолжительность осмотра
deleted_at	timestamptz(6)	Дата и время удаления
created_at	timestamptz(6)	Дата и время создания
updated_at	timestamptz(6)	Дата и время обновления

Таблица примет вид, изображённый на рисунке 3.4.

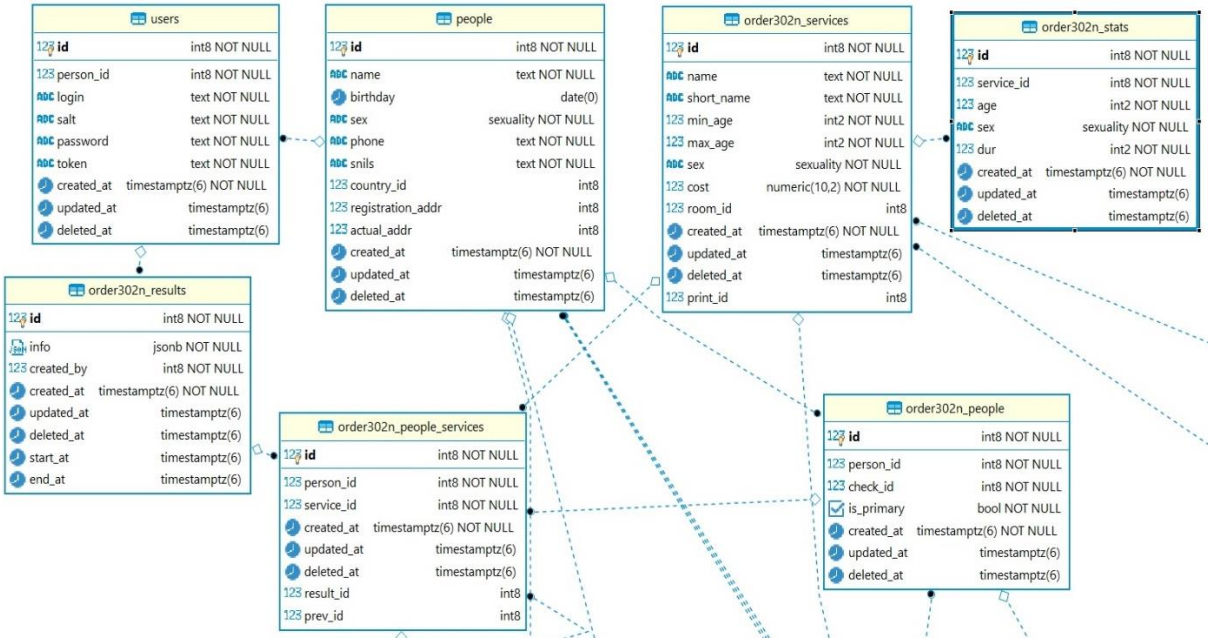

Рисунок 3.4 – Данные о времени прохождения осмотра в базе данных

3.2 Построение алгоритма разработки модуля

Для решения задачи создания модуля статистического анализа пациентов МАУЗ ГКБ №2, необходимо выделить отдельные модули, которые должны реализовывать определенную часть функций и выполнять следующие действия:

1) собирать информацию о пациентах;

2) собирать информацию о услугах;

3) группировать полученные данные;

4) организовывать хранение данных;

5) организовывать представление данных в виде таблиц и графиков.

Общий алгоритм работы программы можно представить в виде алгоритма, изображенного на рисунке 3.5.

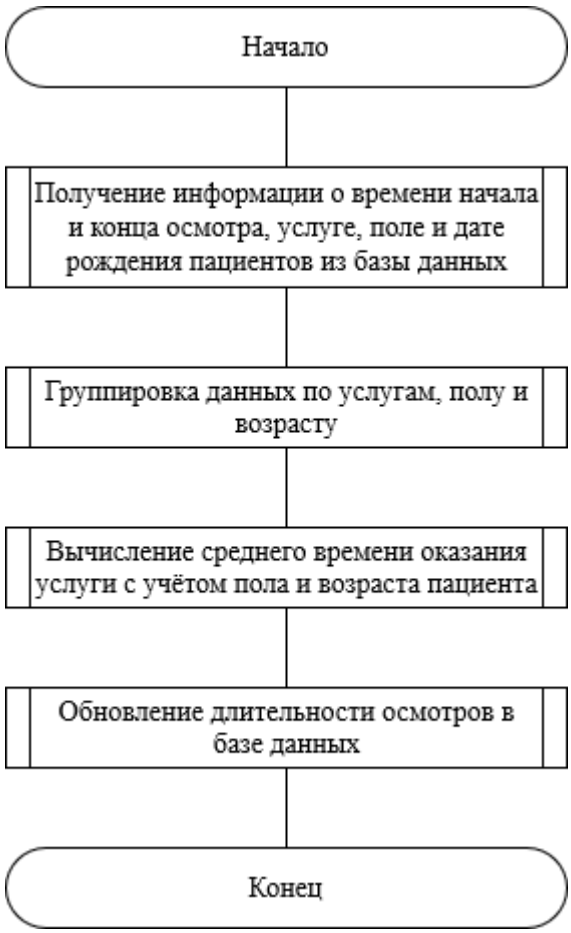

Рисунок 3.5 – Общий алгоритм работы программы

3.2.1 Модуль авторизации в системе

Рассмотрим модуль авторизации в системе. Данный модуль, предназначенный для авторизации пользователя в системе, выполняется при заходе неавторизованного пользователя на сайт. Затем неавторизованный пользователь проходит авторизацию. Далее по предоставленным в запросе данным определяется роль пользователя. Существует три роли пользователя: администратор, руководитель отдела, врач профпатолог. После чего выполняется соответствующий роли модуль.

На рисунке 3.6 приведена схема алгоритма работы данного модуля.

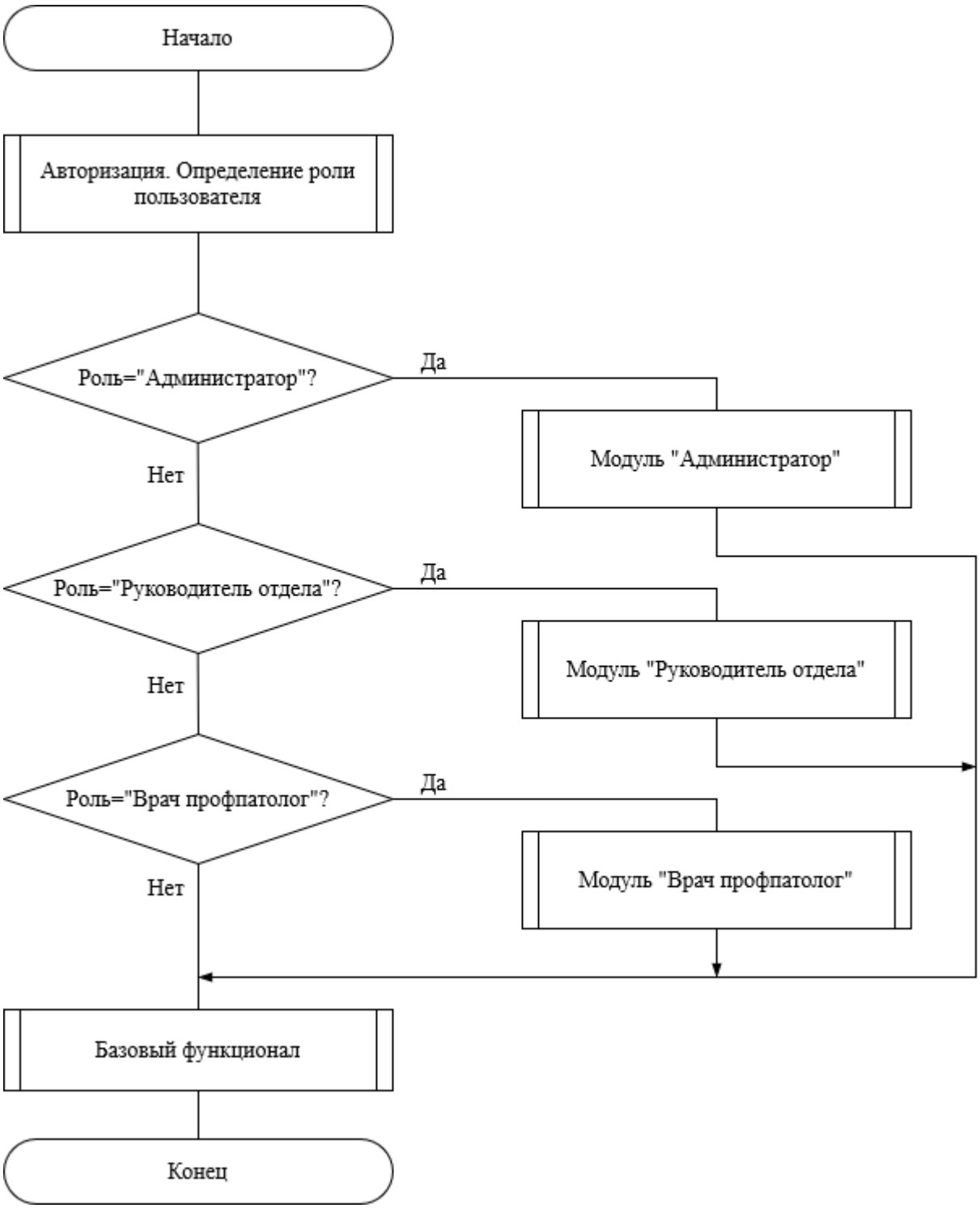

Рисунок 3.6 – Схема алгоритма модуля авторизации

3.2.2 Алгоритм группировки данных

Рассмотрим алгоритм группировки данных, представленный на рисунке 3.7. В теле алгоритма группировки данных находится два цикла. Цикл по данным осмотров, в ходе которого выполняется группировка времени осмотров пациентов одного возраста, и цикл по сгруппированным данным, в

ходе которого происходит вычисление среднего времени осмотра для каждой группы пациентов.

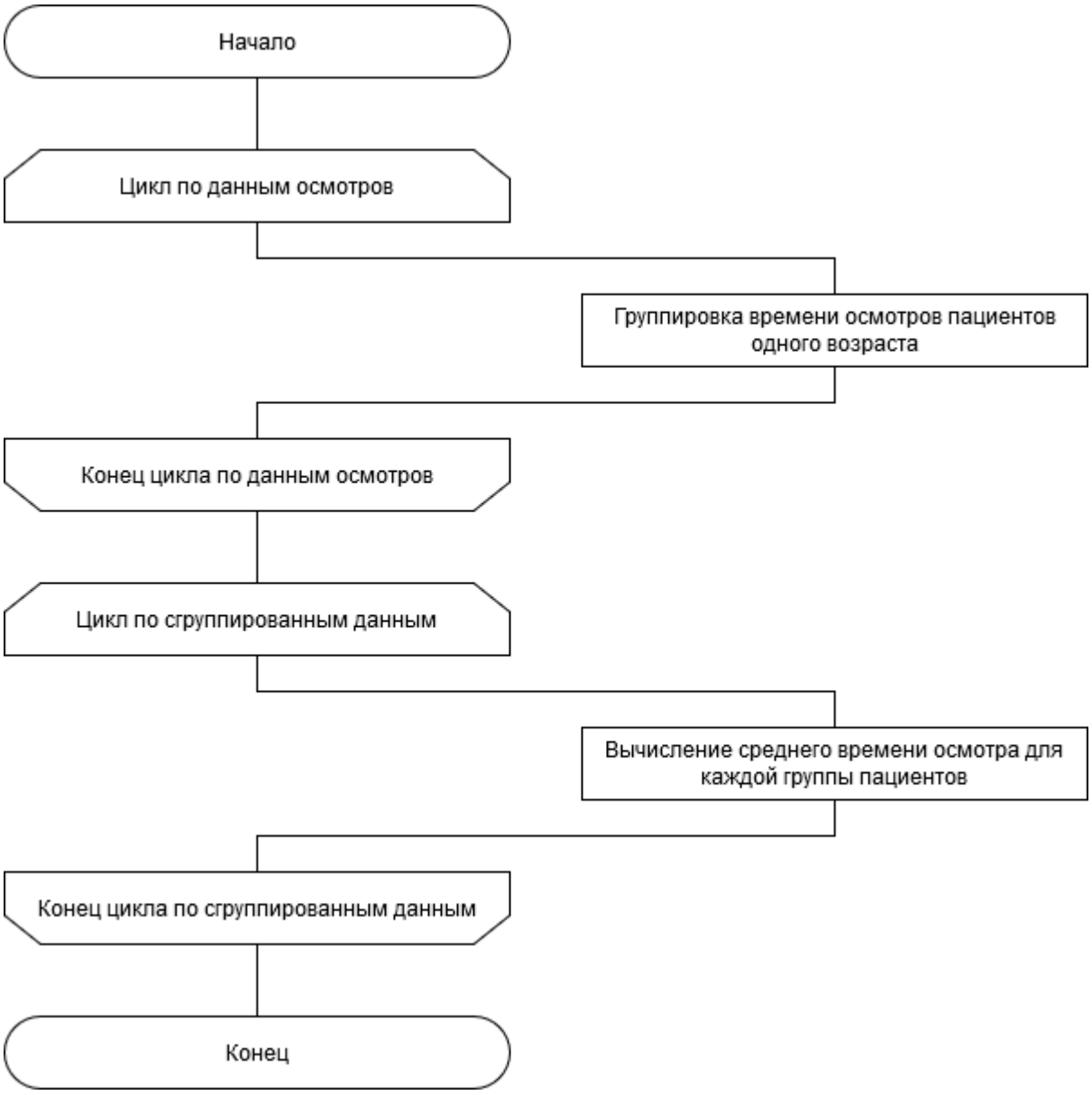

Рисунок 3.7 – Схема алгоритма группировки данных

3.2.3 Схема алгоритма генерации расширенного отчёта

В алгоритме генерации расширенного отчёта, представленном на рисунке 3.7, сперва получается подробная информация о медицинских осмотрах, затем определяются минимальный и максимальный возраста в выборке. Далее в файле формата Excel происходит разделение колонок по полу и возрасту пациентов. В теле алгоритма находится цикл по услугам и вложенный цикл по диапазону возрастов.

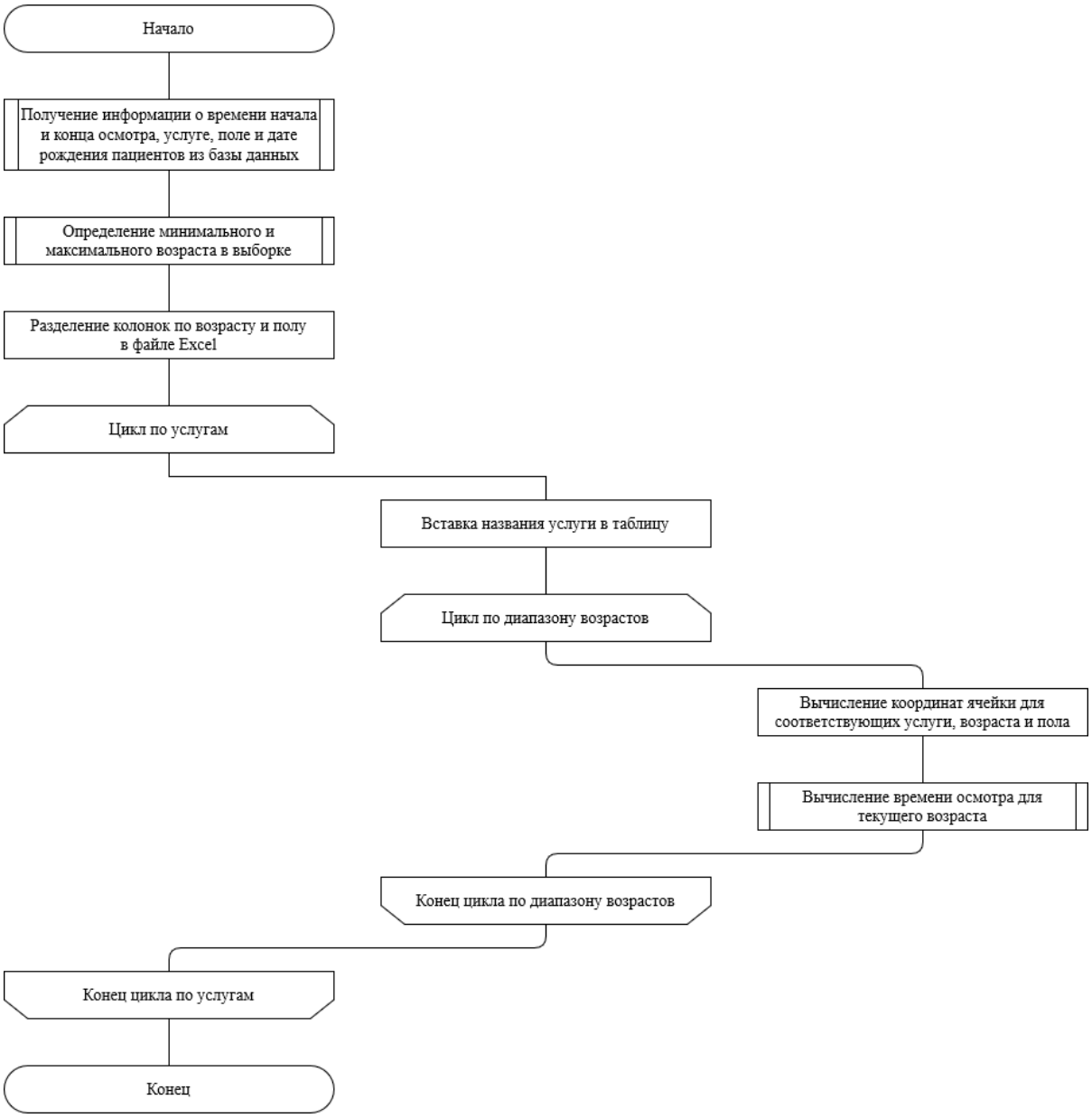

Рисунок 3.8 – Схема алгоритма генерации расширенного отчёта

3.3 Разработка пользовательского интерфейса

В ходе данной работы был разработан пользовательский интерфейс. Экран авторизации представлен на рисунке 3.9. Для того, чтобы получить доступ к функциям системы, необходимо ввести имя пользователя и пароль.

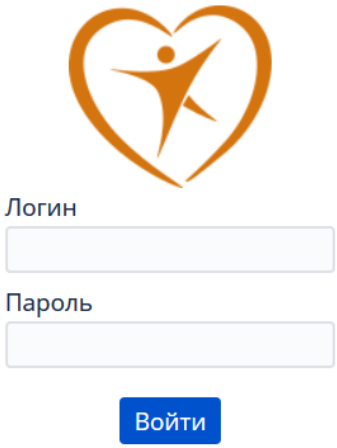

Рисунок 3.9 – Форма аутентификации

В случае успешной аутентификации происходит загрузка модуля, соответствующего роли пользователя в системе. Интерфейс модуля персонализирован для каждой роли и включает в себя только необходимые каждой роли функции.

Модуль «Статистика» (рисунок 3.10) предназначен для просмотра результатов статистического анализа потока пациентов.

Рисунок 3.10 – Модуль «Статистика»

В правом верхнем углу располагаются кнопки: «Отчёт по услугам», «Развернутый отчёт» и «Печать».

При нажатии на кнопку «Отчёт по услугам» открывается список всех медицинских услуг, которые на данный момент существуют в МАУЗ ГКБ № 2 (рисунок 3.11).

Для создания новой услуги необходимо нажать на кнопку «Новый» в правом верхнем углу.

Услуги

Новый

Название	Краткое	Пол	Мин возраст	Макс возраст	Кабинет
HBs-Ag	HBs-Ag	М/Ж	0	100	каб.504
Акушер-гинеколог	Акушер-гинеколог	М/Ж	0	100	3 этаж. Смотреть табл. м/о
АЛК или КП в моче	АЛК или КП в моче	М/Ж	0	100	каб.101
Аллерголог	Аллерголог	М/Ж	0	100	каб.214
АЛТ	АЛТ	М/Ж	0	100	каб.504
Анализ кала на яйца гельминтов	Анализ кала на яйца гельминтов	М/Ж	0	100	3 этаж. Смотреть табл. м/о
анти-HBc-Ig	анти-HBc-Ig	М/Ж	0	100	каб.504
АСТ	АСТ	М/Ж	0	100	каб.504
Аудиометрия	Аудиометрия	М/Ж	0	100	каб.112
Базофильная зернистость эритроцитов	Базофильная зернистость эритроцитов	М/Ж	0	100	каб.504
бак посев кала	бак посев кала	М/Ж	0	100	каб.216
Бактериологическое исследование ЖКТ - кал на возбудителей дизентерии	бак посев кала	М/Ж	0	100	каб.216
Билирубин	Билирубин	М/Ж	0	100	каб.504
Биомикроскопия переднего отрезка глаза	Биомикроскопия переднего отрезка глаза	М/Ж	0	100	каб.519/524
Биомикроскопия сред глаза	Биомикроскопия сред глаза	М/Ж	0	100	каб.519/524
ВИЧ	ВИЧ	М/Ж	0	100	каб.504
Время кровотечения	Время кровотечения	М/Ж	0	100	каб.103
ГГТП	ГГТП	М/Ж	0	100	каб.504
Глюкоза	Глюкоза	М/Ж	0	100	3 этаж. Смотреть табл. м/о

Рисунок 3.11 – Модуль «Услуги»

После выбора услуги загрузится статистический анализ потока пациентов, представленный в виде диаграмм (рисунок 3.12)

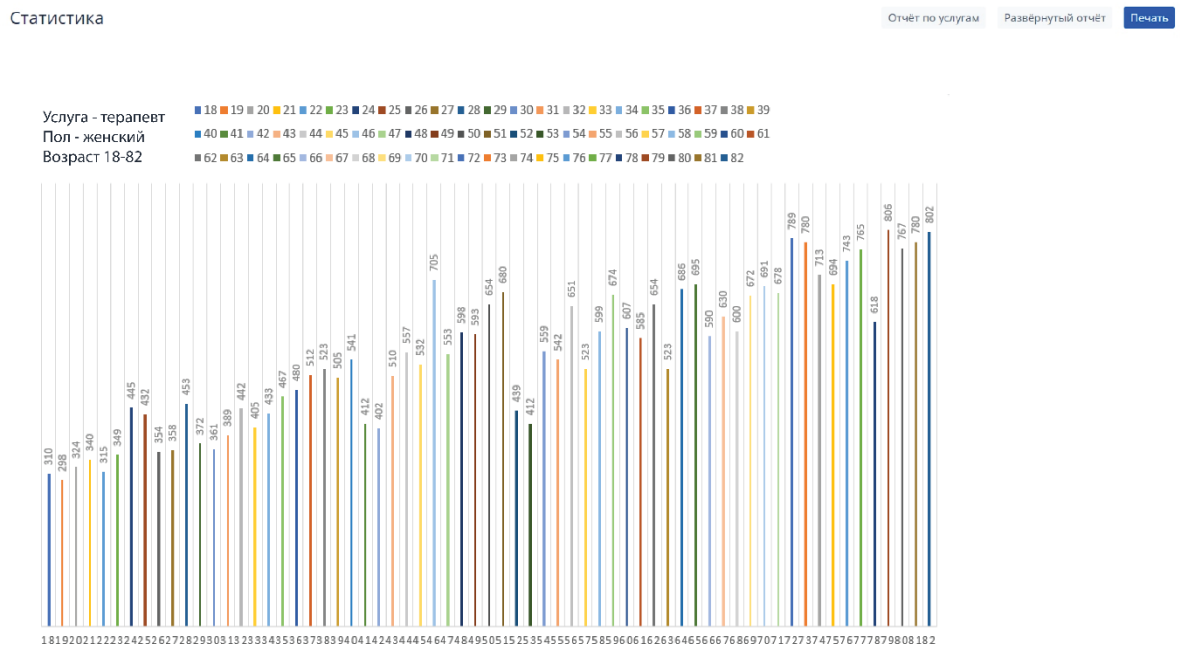

Рисунок 3.12 – Диаграмма отчёта

При нажатии на кнопку «Развёрнутый отчёт» пользователю предоставляется файл отчёта в формате Excel (рисунок 3.13).

		20		21		22		23		24		25	
		М	Ж	М	Ж	М	Ж	М	Ж	М	Ж	М	Ж
3	Терапевт	495,25	502	500,6667	551,1	500,1765	492,2857	518,4118	509,2	518,9565	512,0625	510	500,4118
4	Нарколог	223,5	224,3333	192,5556	193,1	219,1765	222,0476	205,4118	219,4	206,087	197,5	198,7	192,4118
5	Психиатр	324,75	319,6667	351	324,9	332,8235	328,6667	353,4706	345,9	348	331,125	315,3	346,7647
6	Флюорография	277,75	274	280	285,8	260,2353	295,0952	285,7059	274,85	275,5	275,125	264,45	262
7	Общий анализ крови	246,75	219	256,4444	240,3	249,2941	277,4286	279,5882	298,05	275,3478	285,6875	262,85	252,7059
8	Общий анализ мочи	82,75	113,6667	89,22222	98,6	92,70588	96,80952	72,94118	87,25	104,4348	108,3125	94	99,76471
9	ЭКГ	775	736	729,1111	718,4	756,1176	754,5238	780,0588	740,4	745,2174	779,4375	762,1	753,5294
10	Глюкоза	253,25	277,6667	257,2222	292,9	277,5882	278	283	262,9	258,6957	288,1875	275,1	296,4706
11	Холестерин	280,25	298	273,4444	276,3	259,4118	272,3333	275,9412	283,25	254,9565	271,875	263,6	278,6471
12	Офтальмолог	377	308,5	335,3333	304	345,375	333,5556	344,1	305,5556	350,4615	347,6667	340	314,5
13	Оториноларинголог	387	439,6667	459,1429	450,8	421,6	464,4737	433,75	458,8947	435,1765	420,3333	452,7857	428,9375
14	Дерматовенеролог	365	407,5	414,8333	446	443,7778	460,5882	460,4286	452	460,9375	450,25	442,2727	444,75
15	Невролог	503	569,5	589	584	579,25	597,6667	554,5	599,5556	568	566	559,1429	578,25
16	Стоматолог	705,5	744	738,4	755,875	744,25	759,0667	745,1667	750	757,5	740,6364	760,6364	738,25
17	Акушер-гинеколог		872		920,9		919,8571		920		921,6875		937,3529
18	Острота зрения	470	514	436,8	514	424,3333	479,125	449,4444	451,375	467,8333	452,5	462,6667	429
19	Поля зрения					490		450,5		437		476	
20	Аудиометрия	830,5	836,5	784,5	856,5	834,3333	816,3333	820,1667	791,6	813,9	867	767	723,5
21	Исследование функции вестибулярного аппарата		580	606	578	648,6	700,5	633,8571	654,6667	638,9091	586	647	631
22	Исследование крови на сифилис	314,5	382	287,8	328,375	316	315,5333	355,1667	346,6	322,9286	324	328,7273	337,5625
23	Исследования на гельминтозы	380	284,5	305,8	343,375	356,2857	332,1333	330,6667	341,8667	327,7143	319,7273	367,6364	338,5
24	Соскоб на энтеробиоз	193,5	133,5	191,2	167,75	221,4286	205,4	256	209,8	215,9286	222,0909	217,5455	205,1875
25	Мазки на гонорею	247	150	148	235,375	167	234,2727	213	198,8333		213,7	225,4286	201,4167
26	Хирург		593	521,6	461	473	527,5	486,7143	541	517	490	539,1667	507,5

Рисунок 3.13 – Развёрнутый отчёт

3.4 Проверка на экспериментальных данных

Для проверки работы программы была использована выборка данных о времени терапевтического осмотра среди пациентов женского пола и мужского пола, возрастом от 18 до 82 лет. Количество пациентов, которые проходили медицинский осмотр у врача терапевта за промежуток времени с февраля по май составило 58693 (рисунок 3.14).

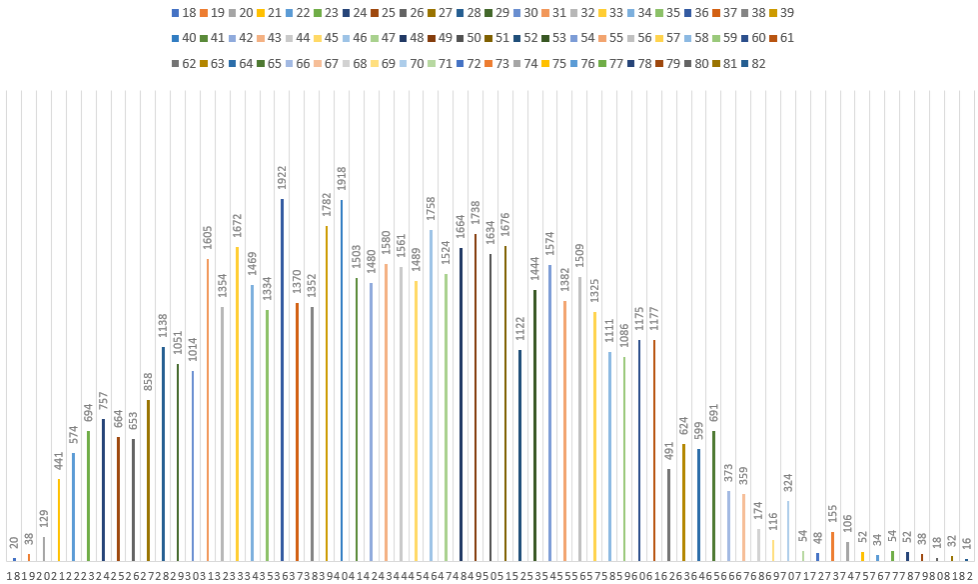

Рисунок 3.14 – Количество пациентов

По окончании работы программы можно увидеть результаты статистического анализа (рисунок 3.15).

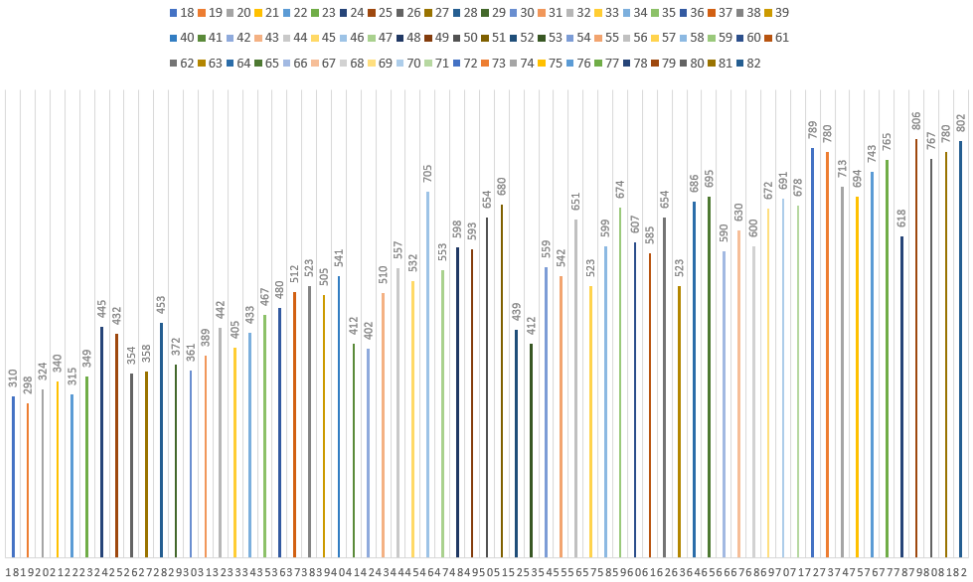

Рисунок 3.15 – Результаты статистического анализа

3.5 Выводы по разделу

В данном разделе были представлены алгоритмы и диаграммы, связанные с разработкой модуля статистического анализа потока пациентов.

Была продемонстрирована работа с базой данных, в частности была создана кэширующая таблица для работы модуля.

Был разработан пользовательский интерфейс программы и проведена проверка работы программы на экспериментальных данных.

ЗАКЛЮЧЕНИЕ

После проведения анализа предметной области были выбраны математические методы для проведения статистического анализа потока пациентов в МАУЗ ГКБ № 2, метод оценки статистической значимости различий (дисперсионный анализ), метод сравнения двух групп: критерий Стьюдента с поправкой Бонферрони. На основании данных методов были проведены эксперименты. А также с помощью метода группировок и с помощью рядов распределения рассчитаны вероятности нахождения анализов в различных состояниях для каждого пациента.

Дисперсионный анализ и критерий Стьюдента с поправкой Бонферрони, позволили оценить вероятность нулевой гипотезы, то есть предположения об отсутствии эффекта экспериментального воздействия, во всех исследованиях. Вероятность нулевой гипотезы оценивалась с помощью критериев значимости – F, t. Гипотеза отклонялась, если значение критерия превышало критическое. Данное отклонение наблюдалось в первом и во втором эксперименте дисперсионного анализа. Также справедливо утверждалось, что найдены статистически значимые различия. Если значение критерия оказывалось меньше критического, то делался вывод об отсутствии статистически значимых различий.

Для разработки модуля использовался язык программирования JavaScript с использованием технологии NodeJS. В информационной системе МАУЗ ГКБ № 2 используется СУБД PostgreSQL, поэтому для создания модуля статистического анализа потока пациентов использовалась библиотека PostgreSQL node-postgres. Среда разработки Visual Studio Code. Для хранения результатов работы модуля статистического анализа потока пациентов была создана кеширующая таблица.

Использование модуля анализа потока пациентов в информационной системе МАУЗ ГКБ № 2 позволяет совершенствовать рабочий процесс специалистов, автоматизирует прохождение медицинских осмотров,

облегчает работу медицинского персонала. Для успешного освоения модуля анализа потока пациентов в медицинской информационной системе необходимо иметь навыки работы с персональным компьютером.

В результате работы получен законченный программный продукт, который внедрен и используется в работе МАУЗ ГКБ №2.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Бахвалов, Н.С. Численные методы / Н.С. Бахвалов – М.: Наука, 1975. – 632 с.

2. Гмурман, В.Е. Теория вероятностей и математическая статистики: учебное пособие / В.Е. Гмурман – М.: Высшая школа, 2003. – 479 с.

3. Дружинина, И. В. Информационные технологии в профессиональной деятельности средних медицинских работников: учебное пособие / И.В. Дружинина. – 5-е изд., стер. – Санкт-Петербург: Лань, 2020. – 112 с. – ISBN 978-5-8114-5208-8. – Текст: электронный // Лань: электронно-библиотечная система. – URL: https://e.lanbook.com/book/136189 (дата обращения: 27.04.2020). – Режим доступа: для авториз. пользователей.

4. Зубов, Н.Н. Математические методы и модели в фармацевтической науке и практике / Н.Н. Зубов, С.З. Умаров, С.А. Бунин. – СПб.: Изд-во Политехи, ун-та, 2008. – 249 с.

5. Избачков, Ю. Информационные системы / Ю. Избачков, В. Петров. – СПб.: Питер, 2005.

6. Кант, В.И. Математические методы и моделирование в здравоохранении. – М.: Медицина, 1987. – 224 с.

7. Кирьянов, Б.Ф. Математические модели в здравоохранении: учебное пособие / Б.Ф. Кирьянов, М.С. Токмачов – НовГУ им. Ярослава Мудрого. – Великий Новгород, 2009. – 279 с.

8. Ланг, Т.А. Описание статистики в медицине. Руководство для авторов, редакторов и рецензентов / Т.А. Ланг, М. Сесик – М.: Практическая медицина. – 2011. – 477с.

9. Леонов, В.П. Применение разведочного анализа для оценки исходных данных / Леонов В.П. // Электронный журнал Биометрика. – 2005. Режим доступа: http://www.biometrica.tomsk.ru/urolitiaz.htm/ свободный. – Загл. с экрана. — Яз. рус.

10. Мандель, И.Д. Кластерный анализ / И.Д. Мандель. – М.: Финансы и статистика, 1988. – 176с.

11. Кочетов, А.Г. Методы статистической обработки медицинских данных: Методические рекомендации для ораторов и аспирантов медицинских учебных заведений, научных работников / А.Г. Кочетов, О.В. Лянг., В.П. Масенко и др. – М.: РКНПК, 2012. – 42 с.

12. Обмачевская, С.Н. Медицинская информатика. Курс лекций: учебное пособие / С.Н. Обмачевская. – 2-е изд., стер. – Санкт-Петербург: Лань, 2019. – 184 с. – ISBN 978-5-8114-4524-0. – Текст: электронный // Лань: электронно-библиотечная система. – URL: https://e.lanbook.com/book/121989 (дата обращения: 21.03.2020). – Режим доступа: для авториз. пользователей.

13. Реброва, О.Ю. Статистический анализ медицинских данных: Применение пакета прикладных программ STATISTICA / О.Ю. Реброва – М.: МедиаСфера, 2006 (III и IV изд.). – 312 с.

14. Сайт «Статистика» [Электронный ресурс]: Первичный анализ данных. – Режим доступа: http://www.statistica.ru/glossary/general/pervichnyy-analiz-dannykh/, свободный – Загл. с экрана. – Яз. рус.

15. Современный учебник JavaScript [Электронный ресурс] URL: https://learn.javascript.ru/ (дата обращения: 10.03.2020).

16. Центральный Javascript-ресурс [Электронный ресурс]: учебник с примерами скриптов, форум, книги и др. / Кантор И. – Электрон. дан. – М., 2007 – Режим доступа: http://javascript.ru/, свободный. – Загл. с экрана. – Яз. рус.

17. Шмойлова, Р. А. Теория статистики: учебник / Р.А. Шмойлова, В.Г. Минашкин, Н.А. Садовникова. – 5-е изд. – Москва: Финансы и статистика, 2014. – 656 с. – ISBN 978-5-279-03295-2. – Текст: электронный // Лань: электронно-библиотечная система. – URL: https://e.lanbook.com/book/53873 (дата обращения: 20.02.2020). – Режим доступа: для авториз. пользователей.

18. Banks J. Interpreting Simulation Software Checklists. – OR/MS Today, 1996, Num. 23. – P. 74–78.

19. Brown, L.D. Interval estimation for a binomial proportion / L.D. Brown, T.T. Cai, A. Dasgupta // Statistical science. – 2001. – №2. – P. 101–133.

20. Day.js 2kB JavaScript date utility library [Электронный ресурс] URL: https://day.js.org/en/ (дата обращения: 23.04.2020).

21. exceljs: Excel Workbook Manager [Электронный ресурс] URL: https://github.com/exceljs/exceljs/ (дата обращения: 07.05.2020).

22. Frappe Charts – Quick Start [Электронный ресурс] URL: https://frappe.io/charts/docs/ (дата обращения: 11.03.2020).

23. Garcia-Perez, M.A. On the confidence interval for the binomial parameter/ M.A. Garcia-Perez // Quality and quantity. – 2005. – N 39. – P. 467–481.

24. Kaplan, R.M. New health promotion Indicators: the general health policy model / R.M. Kaplan // Health Promotion. – V. 3, № 1, 1996. – P. 35–49.

25. MySQL.RU: Одобрено лучшими российскими программистами [Электронный ресурс] – Электрон. дан. – Режим доступа: http://www.mysql.ru/docs/mysql-man-4.0-ru/introduction.html#what-is, свободный. – Загл. с экрана. – Яз. рус.

26. Tandy, R.D. Technical Note: The Initial Stages of Statistical Data Analysis / R.D. Tandy // Journal of Athletic Training. – 1998. – P.69–71.

27. Welcome | node postgres [Электронный ресурс] URL: https://node-postgres.com/ (дата обращения: 13.04.2020).

28. Welcome to The Apache Software Foundation [Электронный ресурс] / The Apache Software Foundation – Электрон. дан. – 2012 – Режим доступа: http://www.apache.org/, свободный. – Загл. с экрана. – Яз. англ.

29. Wilson, E.B. Probable inference, the law of succession, and statistical inference / E.B. Wilson // Journal of American Statistical Association. – 1927. – №22. – P. 209–212.

30. World Health Organization: Statistics of World Health Organization, 2001. – P.19–21.

ПРИЛОЖЕНИЕ
Текст программы

```
async function getCheckDurations() {
    const {rows} = await client.query(`
        SELECT onr.start_at, onr.end_at, p.birthday, p.sex, ons.id, ons.name
        FROM order302n_results onr, order302n_people_services onps,
order302n_services ons, people p, order302n_people onp
        WHERE onr.id = onps.result_id AND onps.service_id = ons.id AND
onps.person_id = onp.id AND onp.person_id = p.id
    `);

    const stats = new Map();

    for (let r of rows) {
        const service_id = r.id;
        let service = stats.get(service_id);

        if (!service) {
            service = {
                name: r.name,
                genders: new Map([
                    ['Ж', []],
                    ['М', []]
                ])
            };
            stats.set(service_id, service);
        }

        const sex = r.sex;
        let gender = service.genders.get(sex);

        const age = dayjs().diff(r.birthday, 'year');
        const dur = r.end_at - r.start_at;
        gender.push({age, dur});
    }

    return stats;
}

function minMaxAge(rows) {
    let min_age = 100, max_age = 0;
```

```
    for (let service of rows.values()) {
        for (let gender of service.genders.values()) {
            for (let person of gender.values()) {
                if (person.age < min_age) {
                    min_age = person.age;
                }
                if (person.age > max_age) {
                    max_age = person.age;
                }
            }
        }
    }

    return [min_age, max_age];
}

function computeDurationForAge(arr) {
    const temp = new Map();
    for (let el of arr) {
        let set = []
        if (temp.has(el.age)) {
            set = temp.get(el.age);
        } else {
            temp.set(el.age, set);
        }
        set.push(el.dur);
    }

    const ageDur = new Map();
    for (let [age, durs] of temp.entries()) {
        if (durs.length === 0) {
            ageDur.set(age, 0);
        } else {
            ageDur.set(age, durs.reduce((acc, x) => acc + x) / durs.length);
        }
    }
    return ageDur;
}

async function generateFullReport(filename, rows) {
    const wb = new Excel.Workbook();
    const ws = wb.addWorksheet("Развёрнутый отчёт");
```

```
const borderThin = {
    top: {style: "thin"},
    left: {style: "thin"},
    bottom: {style: "thin"},
    right: {style: "thin"},
};

const alignCenter = {
    vertical: 'middle',
    horizontal: 'center',
    wrapText: true
};

let [min_age, max_age] = minMaxAge(rows);

for (let age = min_age; age <= max_age; age++) {
    const i = (age - min_age) * 2 + 2;
    ws.mergeCells(1, i, 1, i + 1);

    let cell = ws.getCell(1, i);
    cell.value = age;
    cell.alignment = alignCenter;
    cell.border = borderThin;

    cell = ws.getCell(2, i);
    cell.value = 'M';
    cell.alignment = alignCenter;
    cell.border = borderThin;

    cell = ws.getCell(2, i + 1);
    cell.value = '𝕂';
    cell.alignment = alignCenter;
    cell.border = borderThin;
}

let rowNum = 3;
for (let [service_id, service] of rows.entries()) {
    let cell = ws.getCell(rowNum, 1);
    cell.value = service.name;
    for (let [gender, ages] of service.genders.entries()) {
        const durations = computeDurationForAge(ages);
```

```
            for (let age = min_age; age <= max_age; age++) {
                const i = (age - min_age) * 2 + 2;
                const offset = gender === 'M' ? 0 : 1;
                const val = predictDuration(durations, age);

                cell = ws.getCell(rowNum, i + offset);
                cell.value = val.dur;
                if (val.predicted) {
                    cell.style.font = {bold: true}
                }
            }
        }
        rowNum++;
    }

    return await wb.xlsx.writeFile(`${filename}.xlsx`);
}
```