САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

**КАФЕДРА УПРАВЛЕНИЯ МЕДИКО-БИОЛОГИЧЕСКИМИ СИСТЕМАМИ**

**Большаков Иван Павлович**

**Выпускная квалификационная работа бакалавра**

# Оценка классического течения инфаркта миокарда при наличии заболевания легких

Направление 010300

Фундаментальная информатика и информационные технологии

Научный руководитель,

кандидат физ.-мат. наук,

доцент

Платонов А.В.

Санкт-Петербург

2016

# Оглавление

# Введение

Первое место по числу смертей на планете занимают заболевания сердца и сосудов. С разными вариациями ишемической болезни сердца сталкиваются каждый год миллионы людей. Инфаркт миокарда (ИМ) – самый распространенный вариант данного заболевания.

ИМ является болезнью, которая приводит к таким последствиям, как инвалидность, кардинальная смена образа жизни и даже смерти. Проблема высокой смертности (примерно 30 процентов заболевших), к сожалению, не обходит стороной и развитые страны. Частота ИМ зависит от возраста и половой принадлежности человека: к примеру, заболевание встречается в 5 раз чаще у мужчин, чем у женщин, а 80% случаев ИМ приходится на возраст от 40 до 65 лет [1].

В последние годы при наблюдении клинической картины заболевания обнаружилось учащение случаев смертельных исходов у пациентов молодого и среднего возраста, особенно у мужской части населения [2,3]. Осложненные формы болезни гораздо чаще наблюдаются у молодых заболевших [4]. Нередко эти осложнения приводят к появлению хронической сердечной недостаточности, которая ведет к потере трудоспособности, риску инвалидности и смертельным исходам после перенесенной болезни. Примерно половина молодых людей с данным заболеванием умирает до госпитализации. Все это говорит о том, что ИМ протекает у молодых пациентов «злокачественно» [5].

Это ставит проблему профилактики и лечения ИМ, а также выявления ранних стадий сердечной недостаточности в разряд приоритетных для здравоохранения и военно-медицинской службы, особенно у мужчин молодого и среднего возраста. В данном вопросе может помочь аппарат математической статистики.

Все чаще и чаще в настоящее время специалисты в области статистики принимают участие в процессе планирования и анализа результатов клинической картины пациентов. Возрастает роль анализа в связи с тем, что в случае многих болезней жизнь и здоровье человека зависит от скорости и качества поставленного диагноза, от быстрого и надлежащего процесса лечения. В связи с этим, нельзя пренебрегать таким мощным средством математического аппарата, как статистические методы, особенно в случае смертельно опасных для людей заболеваний, таких как ИМ.

Проводя статистические исследования, можно увидеть, какие факторы (характеристики пациента, его привычки, перенесенные заболевания) способствуют развитию сердечно-сосудистой недостаточности, какое лечение наиболее эффективно, какие факторы ведут к летальному исходу и от чего зависит тот или иной тип осложнений. Все эти критерии, учитывая серьезность болезни, очень важны и актуальны. В данном направлении уже были получены некоторые положительные результаты [6, 7]. К примеру, было выяснено, что активный образ жизни и отказ от курения ведет к значительному снижению количества острых форм ИМ. Поэтому продолжать статистические исследования в указанной области необходимо.

В данной работе для выявления статистических зависимостей и закономерностей были применены дисперсионный и дискриминантный анализ, а также непараметрические методы сравнения. Все виды анализа были реализованы в статистическом пакете *STATISTICA*. Кроме этого, была написана программа на языке *C#,* реализующая прогнозирование осложнений болезни с помощью классифицирующих функций, полученных на этапе дискриминантного анализа.

# Постановка задачи

Дана медицинская база данных, содержащая информацию о 1040 пациентах, перенесших инфаркт миокарда. Каждая запись содержит факторы различного типа: качественные, количественные, порядковые. Все факторы пронумерованы и имеют название.

В связи с критической ситуацией течения болезни и частого возникновения осложнений среди молодых пациентов мужского пола, необходимо выделить в базе данных только мужчин до 60 лет.

Для того чтобы решить вопрос о связи заболеваний легких и инфаркта миокарда необходимо разбить выделенных мужчин на три группы: с заболеванием легких (фактор № 99 признак 2), с заболеванием легких (фактор № 99 признак 3) и без заболевания легких (фактор № 99 признак 1).

После группировки были получены следующие результаты:

- общее число мужчин до 60 лет в медицинской базе: 533 человека;
- число людей в группе без заболевания легких (99_1): 411 человек;
- число людей в группе с заболеванием легких (99_2): 5 человек;
- число людей в группе с заболеванием легких (99_3): 117 человек;

В связи с малой численностью группы (99_2) и невозможностью корректно провести какой-либо вид анализа в данной группе, оптимально будет объединить группы (99_2) и (99_3) в одну группу – группу людей с заболеванием легких в принципе (будем обозначать ее в дальнейшем 99_23). Ее численность составляет 122 человека.

Далее в каждой из образовавшихся двух групп необходимо выделить наиболее значимые количественные факторы (всего 59 штук), влияющие на:

1. параметр «исход» (фактор № 10 в базе; выжил – 1; умер – 2);

2. параметр «возникновение осложнений» (фактор № 20 в базе; 1 - неосложненный ИМ, 2 - осложненный ИМ).

Для наглядности список количественных переменных и некоторые сведения о них (общее число, среднее значение, минимум, максимум и дисперсия) для пациентов без заболевания легких представлены ниже в таблице 1.

| Variable | Valid N | Mean | Minimum | Maximum | Std.Dev. |
|---|---|---|---|---|---|
| ИМТ | 395 | 27,628 | 19,0311 | 44,983 | 3,7252 |
| ППТ | 395 | 2,860 | 1,6148 | 51,195 | 5,5052 |
| АДмакс. | 411 | 171,861 | 120,0000 | 260,000 | 29,2821 |
| АДмин. | 411 | 101,350 | 10,0000 | 140,000 | 12,7953 |
| АДсист | 408 | 137,772 | 0,0000 | 240,000 | 30,4004 |
| АДдиаст | 408 | 85,571 | 0,0000 | 140,000 | 19,1767 |
| ЧСС | 378 | 76,516 | 0,0000 | 188,000 | 19,3595 |
| ОХ | 298 | 5,765 | 3,1000 | 10,320 | 1,1915 |
| ТГ | 172 | 2,637 | 0,5400 | 9,900 | 1,8268 |
| ЛПНП | 171 | 4,191 | 2,2000 | 8,620 | 1,0534 |
| ЛПОНП | 171 | 0,780 | 0,3000 | 2,640 | 0,3207 |
| ЛПВП | 171 | 0,926 | 0,3400 | 2,100 | 0,2646 |
| КА | 159 | 5,344 | 2,1800 | 12,600 | 1,9376 |
| ОХ/ЛПВП | 171 | 6,707 | 3,3810 | 15,176 | 2,0115 |
| ПИ | 250 | 85,578 | 47,0000 | 120,000 | 12,1942 |
| Creatin | 288 | 0,100 | 0,0600 | 0,222 | 0,0206 |
| Gluc | 340 | 6,209 | 3,2000 | 18,900 | 2,0453 |
| Na | 250 | 138,415 | 120,0000 | 156,000 | 5,5066 |
| K | 255 | 4,443 | 3,1000 | 5,900 | 0,5112 |
| Ca | 98 | 2,314 | 1,0300 | 2,960 | 0,3531 |
| Mg | 10 | 0,842 | 0,5700 | 1,050 | 0,1985 |
| Cl | 149 | 101,368 | 90,0000 | 118,000 | 4,5510 |
| Мочев. к-та | 6 | 173,320 | 0,2200 | 529,000 | 253,5927 |
| ЛПНП/ЛПВП | 171 | 4,865 | 1,9048 | 12,676 | 1,8283 |
| ДА1 | 354 | 33,421 | 22,0000 | 44,600 | 3,7509 |
| ЛП1 | 329 | 40,743 | 27,0000 | 58,700 | 5,0746 |
| КДР1 | 334 | 5,326 | 3,5000 | 7,900 | 0,7570 |
| КСР1 | 334 | 3,918 | 2,0000 | 6,500 | 0,9314 |
| МЖП1 | 374 | 12,651 | 5,0000 | 21,000 | 2,8824 |
| ЗС1 | 374 | 11,655 | 6,0000 | 20,000 | 2,6779 |

| Variable | Valid N | Mean | Minimum | Maximum | Std.Dev. |
|---|---|---|---|---|---|
| КДРпж1 | 29 | 24,103 | 15,0000 | 56,000 | 7,1880 |
| ПП1 | 1 | 40,000 | 40,0000 | 40,000 | |
| КДО1 | 334 | 142,242 | 50,8686 | 335,080 | 49,1226 |
| КСО1 | 334 | 73,479 | 12,7273 | 215,997 | 42,9867 |
| ФВ1 | 336 | 49,268 | 15,0000 | 84,930 | 15,5996 |
| ФУ1 | 335 | 25,882 | 7,0000 | 54,167 | 9,8431 |
| VE1 | 222 | 0,504 | 0,2000 | 0,840 | 0,1423 |
| VA1 | 222 | 0,647 | 0,2700 | 1,000 | 0,1433 |
| VE/VA1 | 222 | 0,796 | 0,2500 | 1,400 | 0,2187 |
| IVRT1 | 214 | 98,458 | 55,0000 | 164,000 | 23,5899 |
| RR (DT)1 | 379 | 0,841 | 0,4286 | 1,579 | 0,1811 |
| УО1 | 371 | 66,450 | 14,0543 | 136,901 | 25,2157 |
| УИ1 | 371 | 32,751 | 6,7791 | 69,768 | 12,6616 |
| СВ1 | 360 | 4,909 | 0,9658 | 13,670 | 2,0946 |
| СИ1 | 360 | 2,414 | 0,4867 | 7,050 | 1,0607 |
| Рпа1 | 386 | 32,836 | 11,9000 | 127,372 | 16,1650 |
| ЧСС1 | 379 | 74,674 | 38,0000 | 140,000 | 16,4108 |
| Рсрлао(Адср)1 | 391 | 104,380 | 36,6667 | 166,667 | 19,9873 |
| ОПСС1 | 359 | 2017,738 | 419,0000 | 6084,000 | 993,2507 |
| ОЛС1 | 359 | 533,015 | 99,4151 | 2921,419 | 397,1509 |
| ИСПСН1 | 407 | 58,219 | 6,0000 | 100,000 | 21,0869 |
| RR -->DT1 | 379 | 0,841 | 0,4286 | 1,579 | 0,1811 |
| АДсист1 | 391 | 139,634 | 50,0000 | 240,000 | 28,7995 |
| АДдиаст1 | 391 | 86,921 | 30,0000 | 140,000 | 17,4472 |
| МО1 | 361 | 4,893 | 0,9300 | 13,673 | 2,0990 |
| ИндНор1 | 411 | 8,159 | 1,5400 | 17,160 | 3,0448 |
| ОТС1 | 330 | 0,474 | 0,2203 | 0,921 | 0,1386 |
| МЛЖ1 | 330 | 319,949 | 123,5737 | 609,337 | 96,2707 |
| ИММЛЖ1 | 330 | 157,212 | 58,7954 | 334,478 | 44,9669 |

Табл.1. Количественные переменные и сведения о них. Группа без заболевания легких

Обобщая вышесказанное, задача состоит в выявлении тех количественных факторов, которые оказывают наибольшее влияние на исход и осложненные формы болезни при инфаркте миокарда. Это позволит определить причины смертельных случаев данной болезни и тяжелых осложнений, которые часто к ним приводят.

Кроме этого, необходимо провести данный анализ в двух группах пациентов, тем самым выяснив разницу клинического течения болезни у больных с заболеванием легких и без подобного заболевания.

По результатам анализа, в каждой из групп желательно построить функцию прогнозирования исхода и осложнений для поступивших пациентов, чтобы врачи могли в кратчайшие сроки поставить правильный диагноз и назначить качественное лечение.

Данную задачу можно решить различными способами (например, используя аппарат нейронных систем, систем нечеткого вывода и т.д.). В связи с тем, что параметры «Исход» и «Осложнения» (зависимые переменные) имеют качественный формат, а независимые переменные, которые необходимо проанализировать – количественный формат, то в данной работе для решения поставленной задачи рассмотрены такие средства математической статистики, как дисперсионный и дискриминантный анализ.

# Обзор литературы

Среди болезней системы кровообращения лидирующую позицию по числу смертей занимает ишемическая болезнь сердца (ИБС) и ее острая форма – инфаркт миокарда (ИМ) [8].

В связи с серьезностью подобного рода болезни, и высокой актуальностью ее анализа, ведутся многочисленные исследования. Например, к настоящему времени выяснено, что такие факторы, как употребление алкоголя, наркотических веществ, курение, физическая активность, пищевые привычки, проживание в загрязненных районах и некоторые другие влияют на возникновение ИМ [9]. Также, в международном исследовании было выяснено, что двумя наиболее важными из них во всех регионах мира являются курение и ненормальное соотношение аполипопротеинов В и А-1 [10].

В России за период с 1990 г. по настоящее время наибольшие рост смертельных случаев от сердечно-сосудистых заболеваний зарегистрирован среди мужчин 20-60 лет [11]. Кроме этого, у людей молодого возраста отмечено внезапное начало болезни ИМ. Часто это обуславливает позднюю госпитализацию пациентов и позднее начало лечения, тем самым являясь причиной осложнений болезни [12]. Другие исследования последних лет также показали, что болезнь ИМ с каждым годом «омолаживается» и увеличение смертности происходит, главным образом, среди мужчин молодого и среднего возраста [3, 13]. Мужской пол является доказанным фактором риска заболевания. При исследовании пациентов до 40 лет доля мужчин среди них составила 89%-100% [14].

В исследованиях на предмет возникновения ИМ у мужчин молодого возраста наиболее значимыми факторами оказались курение, артериальная гипертензия, сахарный диабет, наследственная отягощенность ИБС [15]. Также было доказано, что малоподвижный образ жизни и курение прямо

пропорционально увеличивает риск смерти среди пациентов мужского пола [16].

В результате учета выявленных факторов риска, больничная смертность от ИМ существенно снизилась с 1994 г. по 2006 г. Однако, среди мужчин до 55 лет это снижение было наименьшим – 33,3% [17]. В настоящее время, несмотря на возросшие возможности диагностики и лечения сердечно-сосудистых заболеваний, уменьшить заболеваемость и смертность среди таких пациентов пока не удается [1, 11].

Таким образом, несмотря на большое число исследований в данной области, количество осложненных форм ИМ и смертность среди мужчин молодого и среднего возраста остаются высокими, поэтому исследовать факторы риска в данном возрастном сегменте мужской части населения необходимо. Кроме этого, необходима детализация факторов в зависимости от заболевания легких; также недостаточно исследована роль факторов в развитии осложнений ИМ. Все это говорит о высокой актуальности данной работы, где для нахождения зависимостей используются статистические методы. Сами дисперсионный и дискриминантный методы, используемые в работе, описаны в [18, 19]. В [20, 21] описываются правила и способы математико-статистической обработки данных. База [22] содержит примеры применения статистических методов к медицине, которые показывают возможности пакета *STATISTICA.*

# Глава 1. Краткое описание математического аппарата

## 1.1 Дисперсионный анализ

### 1.1.1 Однофакторный дисперсионный анализ

Целью дисперсионного анализа является ответ на вопрос: «Как связаны между собой количественные и качественные переменные?». Качественные переменные выступают в роли признаков-факторов, а количественные – в роли признаков-результатов. В зависимости от количества таких признаков выделяют различные виды дисперсионного анализа. В данной работе будем рассматривать однофакторный дисперсионный анализ (одна зависимая и одна независимая переменная) [18,23,24,25].

В данном виде анализа с помощью сравнения дисперсий проверяется значимость различия между средними значениями двух или более подгрупп общей выборки. Сначала формулируется нулевая гипотеза. Она предполагает, что все средние значения равны между собой. Это значит, что исследуемый фактор не оказывает никакого влияния на исследуемую величину. В таком случае разброс данных внутри подгрупп должен быть не меньше, чем разброс данных между этими подгруппами.

Итак, пусть имеется выборка величины $n$. Задан фактор, который может принимать $m$ значений. Тогда вся выборка разобьется на $m$ подгрупп, соответствующих разным значениям фактора. Пусть $n_k$ – величина $k$-й подгруппы, $k = 1, \dots, m$ $(n_1 + \cdots + n_k = n)$; $x_{ik}$ – $i$-е значение в $k$-й подгруппе выборки, $i = 1, \dots, n_k$; $k = 1, \dots, m$. Найдем выборочное среднее в каждой подгруппе

$$\overline{x_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ik} \ , \quad k = 1, \dots, m \ ,$$

и общее среднее

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k} x_{ik} \ .$$

Основное тождество дисперсионного анализа имеет следующий вид:

$$D_{\text{общ}} = D_{\text{меж}} + D_{\text{вн}},$$

где $D_{\text{общ}} -$ общая дисперсия, $D_{\text{меж}} -$ межгрупповая дисперсия, $D_{\text{вн}} -$ внутригрупповая дисперсия. Расчет дисперсий проводится по следующим формулам:

$$D_{\text{общ}} = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2 = \sum_{k=1}^{m}\sum_{i=1}^{n_k} x_{ik}^2 - n\bar{x}^2,$$

$$D_{\text{меж}} = \sum_{k=1}^{m} n_k(\overline{x_k} - \bar{x})^2 = \sum_{k=1}^{m} n_k\overline{x_k}^2 - n\bar{x}^2 \ ,$$

$$D_{\text{вн}} = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(x_{ik} - \overline{x_k})^2 = \sum_{k=1}^{m}\sum_{i=1}^{n_k} x_{ik}^2 - \sum_{k=1}^{m} n_k\overline{x_k}^2 .$$

Построим корреляционное отношение $\eta^2$ по формуле:

$$\eta^2 = \frac{D_{\text{меж}}}{D_{\text{общ}}}.$$

Его значение находится в пределах от 0 до 1. Близость данной величины к 0 говорит об отсутствии влияния независимой переменной на зависимую, в то время как близость к 1 указывает на наличие такого влияния.

Для проверки гипотезы о равенстве средних пользуемся критерием Фишера [24]. Находим значение *F*-статистики:

$$F = \frac{D_{\text{меж}}}{D_{\text{вн}}}\frac{n-m}{m-1},$$

и если вычисленное значение этой статистики будет больше, чем критическое табличное значение $F_{\alpha,m-1,n-m}$ ,(см., например, [23, с. 238]). то тогда есть основания считать, что независимый фактор оказывает влияние на разброс средних значений (нулевая гипотеза отклоняется), в противном случае, влияние одной переменной на другую не находит подтверждения. Здесь $\alpha -$ уровень значимости. Степень свободы для межгрупповой дисперсии составляет $m - 1$, а для внутригрупповой дисперсии $n - m$. Табличное критическое значение $F_{\alpha,m-1,n-m}$ определяется с помощью количества степеней свободы и соответствующего уровня значимости (по умолчанию 5%).

Для возможности проведения дисперсионного анализа должны выполниться следующие условия:

1) результирующая переменная внутри каждой группы имеет нормальное распределение (при незначительных отклонениях от нормального распределения дисперсионный анализ все же может быть применен);

2) подгруппы имеют равные дисперсии;

3) $m \geq 2$, т.е. число подгрупп должно быть не менее двух;

4) $n_i \geq 2$, $i = 1, \ldots, k$, т.е. число значений в каждой подгруппе должно быть не менее двух.

### 1.1.2 Проверка распределения выборки на нормальность

Среди методов для проверки распределения выборки на нормальность можно выделить такие, как критерий Шапиро-Уилка, $t$ – тест, критерий Колмогорова-Смирнова и т.д. (см., например, [26]). Также можно построить гистограмму и визуально убедиться в нормальности распределения. В пакете *STATISTICA* это достаточно просто сделать: для выбранной переменной

имеется возможность построить гистограмму и применить указанные выше критерии. Глядя на гистограмму, не всегда легко сказать, что данные распределены нормально, но, если объем выборки достаточно велик, форма выборочного распределения приближается к нормальной («почти» нормальное распределение), даже если распределение исследуемых переменных не является нормальным. Этот важный принцип основывается на центральной предельной теореме (см., например, [21, с. 238]).

В пакете *STATISTICA* присутствует улучшенная версия алгоритма Шапиро-Уилка, которая позволяет его применять при больших выборках (до 2000 наблюдений). Отметим, что данный критерий является самым эффективным методом проверки на нормальность распределения, поэтому рассмотрим именно его.

Критерий Шапиро-Уилка основан на оптимальной несмещённой оценке дисперсии к её обычной оценке методом максимального правдоподобия. Статистика критерия имеет следующий вид:

$$W = \frac{1}{D}\left[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right]^2,$$

где $n -$ общее число наблюдений, $x_i - i$-е значение в выборке, $i = 1, \dots, n$,

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

т.е. $\bar{x}$ – выборочное среднее, $D -$ выборочная дисперсия. Коэффициенты $a_{n-i+1}, , i = 1, \dots, n,$ можно посмотреть в статистических таблицах (см., например, [26, с.236]).

Критические значения $W(\alpha)$ статистики находятся также из таблицы (см. [26]). Если вычисленное значение статистики меньше критического, то нулевая гипотеза отклоняется на уровне значимости $\alpha$. Уровень

согласованности с нулевой гипотезой о нормальности распределения можно получить по формуле:

$$z = \gamma + \eta \ln\left(\frac{W - \varepsilon}{1 - W}\right),$$

где коэффициенты $\gamma, \eta, \varepsilon$ также можно взять из статистических таблиц (см. [26, с.239]).

Чем ближе вычисляемая статистика $W$ к 1, тем меньше вероятность ошибочно принять гипотезу о нормальности распределения.

Стоит сказать, что классическая реализация критерия Шапиро-Уилка имеет ограниченную применимость. При $n$>100 использование таблицы с коэффициентами $a_{n-i+1}, i = 1, \dots, n,$ становится неудобным. В связи с этим была предложена модификация данного критерия [27]. Чтобы применять критерий Шапиро-Уилка без помощи таблиц, была выведена полезная аппроксимация. Для уровня значимости α = 0,05 предлагается статистика

$$W_1 = \left(1 - \frac{0{,}6695}{n^{0{,}6518}}\right)\frac{D}{B},$$

где

$$B = \left\{\sum_{j=1}^{m} a_j \left(x_{n-j} - x_j\right)\right\}^2; \qquad m = \left[\frac{n}{2}\right]; \qquad a_0 = \frac{0{,}899}{(n - 2{,}4)^{0{,}4162}} - 0{,}02;$$

$$a_j = a_0\left[z + \frac{1483}{(3 - z)^{10{,}845}} + \frac{71{,}6 * 10^{-10}}{(1{,}1 - z)^{8{,}26}}\right]; \qquad z = \frac{n - 2j + 1}{n - 0{,}5}.$$

Если $W_1 < 1$, то нулевая гипотеза отклоняется.

### 1.1.3. Проверка на равенство дисперсий

В данной работе использовался один из критериев на проверку равенства дисперсий – критерий Левена, который является аналогом *t*-критерия Стьюдента, критерия Бартлетта, критерия Фишера и других [26, 28].

Критерий Левена проверяет $m$ выборок на наличие равных дисперсий. Данный критерий отличает малая чувствительность к отклонениям от нормального распределения.

Нулевая гипотеза формулирует равенство дисперсий у $m$ выборок:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 ,$$

.

а конкурирующая с ней гипотеза:

$$H_1: \sigma_i^2 \neq \sigma_j^2,$$

предполагает, что по крайней мере для одной пары индексов $i, j \in \{1, \dots, k\}$ условие не выполняется.

Пусть $n_i$ – количество наблюдений в $i$ -ой выборке, $n = \sum_{i=1}^{m} n_i$ – общее количество наблюдений, $x_{ij}$ – $j$-е наблюдение в $i$-ой выборке. Критерий Левена имеет следующего вида статистику:

$$W = \frac{n-m}{m-1} \frac{\sum_{i=1}^{m} n_i (\overline{z_i} - \bar{z})^2}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} n_i (z_{ij} - \overline{z_i})^2},$$

где $z_{ij}$ – один из вариантов:

1. $z_{ij} = \left| x_{ij} - \overline{x_i} \right|$ ($\overline{x_i}$ – среднее значение по $i$-ой выборке);
2. $z_{ij} = \left| x_{ij} - \widetilde{x_i} \right|$ ($\widetilde{x_i}$ – медиана в $i$-ой выборке);
3. $z_{ij} = \left| x_{ij} - \overline{x_i'} \right|$ ($\overline{x_i'}$ – усечённое среднее в $i$-ой выборке);

$\overline{z_i}$ – среднее $z_{ij}$ по $i$-й выборке, $\bar{z}$ – общее среднее $z_{ij}$ по всем группам.

В каждом из трех указанных вариантов выбора величин $z_{ij}$ устойчивость критерия Левена рассмотрена с разных углов. Случай с выборочными средними присутствовал изначально в работе Левена. Позже, Браун и Форсайт ввели в данный критерий использование усеченных средних и выборочных медиан [29]. При введении усеченного среднего, критерий устойчив к отклонениям в сторону распределения Коши, а в варианте с выборочными медианами − к асимметрии закона. В случае симметричных и умеренно отличающихся распределений критерий имеет наибольшую мощность при использовании выборочного среднего.

Нулевая гипотеза критерия отклоняется, если

$$W > F_{\alpha,m-1,n-m},$$

где $F_{\alpha,m-1,n-m}$ − табличное критическое значение $F$-распределения с $m-1$ и $n-m$ степенями свободы и уровнем значимости $\alpha$ (см., например, [23, с. 238]).

## 1.2 Непараметрические методы сравнения

Непараметрические критерии применяются в тех случаях, когда вид распределения неизвестен или предположения о нормальности распределения и равенстве дисперсий весьма неточны. Они основываются на использовании рангов и частот и не включают в расчет параметры вероятностного распределения.

В данной работе был выбран один из самых известных и самых распространенных методов непараметрического сравнения двух выборок – $U$- критерий Манна-Уитни [30].

Критерий Манна-Уитни применяется для сравнения выборок объема $n_1$ и $n_2$ и проверяет гипотезу $H_0$ о том, что выборки имеют равные

средние и медианы, то есть что выборки получены из однородных генеральных совокупностей.

Статистика *U*-критерия выражается следующим образом. Все значения из обеих выборок расположим в виде вариационного ряда (порядке возрастания). Каждому элементу ряда присвоим ранг, который определяется номером данного элемента в ряду.

В случае совпадения по величине нескольких наблюдений, ранг каждого будет равен среднему арифметическому их номеров. Последний элемент в составленном ряду должен иметь ранг $n_1 + n_2$.

Пусть $R_1$ — сумма рангов первой выборки, $R_2$ — сумма рангов второй выборки. Вычислим значения :

$$w_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 ,$$

$$w_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 .$$

После подсчета, должно выполняться условие

$$w_1 + w_2 = n_1 n_2.$$

За выборочное значение статистики $U$ критерия принимаем наименьшее из $w_1$, $w_2$. Полученное значение *U*-критерия сравниваем с табличным критическим значением $U$ при заданной численности групп и соответствующим уровнем значимости. Если полученное значение $U$ не больше критического, то нулевая гипотеза отклоняется, и статистическая значимость различий в группах принимается.

Для возможности применения критерия должны выполниться условия:

1. в каждой из выборок должно быть не менее 3 наблюдений (допускается, чтобы в одной выборок было 2 значения, но во второй тогда не менее 5);
2. совпадающих по значениям наблюдений должно быть как можно меньше.

## 1.3 Дискриминантный анализ

Дискриминантный анализ – это статистический метод, позволяющий изучать различия между несколькими группами. Эта теория объединяет в себе несколько статистических процедур: методы классификации наблюдений по группам и методы интерпретации межгрупповых различий [19, 25, 31].

Интерпретация результатов дает ответ на вопрос « Можно ли отличить один класс от другого, используя данный набор характеристик?»; говорит о том, насколько хорошо эти характеристики могут провести различия и выявить наиболее информативные из них.

Методы классификации позволяют получить функции, которые помогают определить, к какой группе относится объект. Эти функции, зависящие от значений характеристик, называются дискриминантными (ДФ). А характеристики, применяемые для отличия одной группы от другой, называются дискриминантными переменными (ДП). Характеристики, находящиеся в линейной зависимости с другими, не являются информативными и исключаются.

Дискриминантный анализ используют для создания «модели», позволяющей лучше всего определить к какой совокупности принадлежит элемент. В рамках этой теории используют два термина: «в модели» - для описания параметров, которые используются для классификации; и «вне модели» – для описания неиспользуемых параметров.

Дискриминантный анализ можно применять при выполнении следующего ряда предположений:

1. наблюдаемые величины (измеряемые характеристики объекта) имеют нормальное распределение (следует заметить, что умеренные отклонения от этого предположения не являются критическими);

2. дисперсии и ковариации наблюдаемых переменных в разных классах однородны (отличие между классами имеется только в средних); умеренные отклонения от этого предположения также допустимы;

3. в модели дискриминантного анализа должно быть:

а) не менее двух классов;

б) по крайней мере, два объекта в каждом классе;

в) любое число дискриминантных переменных при условии, что оно не превосходит число объектов за вычетом двух;

Также стоит упомянуть о значениях толерантности. Толерантность является мерой избыточности переменной. Значение толерантности переменной вычисляется как $(1 - R^2)$, где $R^2$ – множественная корреляция переменной с остальными. Если переменная почти полностью избыточна (т.е. значение толерантности для неё приближается к нулю), то матрица задачи является плохо обусловленной. В случае, когда данное предположение верно, дискриминантный анализ не применим. Чтобы избежать ситуации с плохо обусловленной матрицей, нужно использовать пошаговый дискриминантный анализ. Переменные, со значением толерантности ниже установленного уровня, включаться в модель не будут.

### 1.3.2 Пошаговый дискриминантный анализ

Пошаговый дискриминантный анализ – метод анализа зависимостей, в котором переменные в модель вводятся последовательно. На каждом шаге изучаются все переменные, из них выбирается самая значимая (вносящая наибольший вклад в дискриминацию между группами). Эта переменная включается в модель на данном шаге, а на следующем происходит поиск новой переменной для включения среди оставшихся вне модели [19, 25].

Пусть имеется $p$ переменных, $g$ групп и $n_k$ объектов в группе $k$; $n$ – общее количество наблюдений. Основной целью дискриминации является нахождение такой линейной комбинации переменных, которая бы оптимально разделила рассматриваемые группы:

$$f_{km} = u_0 + u_1 x_{1km} + u_2 x_{2km} + \cdots + u_p x_{pkm}, \qquad (1.3.2.1)$$

$$m = 1, \dots, n_k\,, \qquad k = 1, \dots, g,$$

где $f_{km}$ – значение линейной комбинации для $m$-го объекта в группе $k$; $x_{ikm}$ – значение $i$-ой ДП для $m$-го объекта в группе $k$; $u_i -$ неизвестные коэффициенты, выбираемые таким образом, чтобы центры кластеров групп максимально возможно отличались друг от друга. Такая линейная комбинация ДП называется канонической дискриминантной функцией (КДФ). Такие функции должны быть некоррелированы между собой. Общее количество КДФ не должно превышать числа ДП и, по крайней мере, должно быть на 1 меньше количества групп.

Соотношение (1.3.2.1) задает математическое преобразование $p$-мерного пространства ДП в $q$-мерное пространство КДФ (где $q$ – максимальное число функций).

Для получения коэффициентов КДФ используется статистический критерий различения групп. Чем меньше рассеивание объектов группы

относительно их центроида и больше расстояние между центрами кластеров, тем лучше будет происходить классификация. Наилучшая КДФ $f$ для дискриминации данных находится через максимизацию отношения межгрупповой дисперсии к внутригрупповой.

Оценим информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп. Для этого вычисляется матрица попарных произведений и сумм квадратов $T = \{t_{ij}\}$. Она характеризует расположение объектов в пространстве, определяемом переменными. Элементы этой матрицы находятся по следующей формуле:

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \overline{x_i})(x_{jkm} - \overline{x_j}), \qquad (1.3.2.2)$$

$$\overline{x_i} = \frac{1}{n} \sum_{k=1}^{g} n_i \overline{x_{ik}}\,, \qquad i = 1, \dots p,$$

$$\overline{x_{ik}} = \frac{1}{n_i} \sum_{m=1}^{n_k} x_{ikm}\,, \qquad i = 1, \dots p, \qquad k = 1, \dots g, \qquad m = 1, \dots, n_k.$$

Выражение в скобках в формуле (1.3.2.2) – это отклонения значений переменных от общего среднего. Элементы, стоящие на диагонали матрицы $T$ определяют сумму квадратов отклонений дисперсий ДП от общего среднего для этих ДП. Остальные элементы матрицы – это суммы произведений отклонения по одной переменной на отклонение по другой. Если разделить матрицу $T$ на $(n-1)$, то получим ковариационную матрицу.

Для измерения степени разброса объектов внутри групп рассмотрим матрицу $W = \{w_{ij}\}$, которая отличается от $T = \{t_{ij}\}$ только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных. Элементы матрицы $W$ определяются, как:

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \overline{x_{ik}})\left(x_{jkm} - \overline{x_{jk}}\right), \qquad i,j = 1, \dots p.$$

Если центроиды групп окажутся равными, то элементы матриц $W$ и $T$ совпадут. Если же центроиды групп различны, то разница $B = T - W$ будет определять межгрупповую сумму квадратов и попарных произведений. Элементы матрицы $B = \{b_{ij}\}$ определяются как $b_{ij} = t_{ij} - w_{ij}$.

Матрицы $W$ и $B$ содержат информацию о зависимостях как между группами, так и внутри групп. Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах. Решим следующего вида систему:

$$\begin{aligned}
\sum_{k=1}^{q} b_{1k} v_k &= \sum_{k=1}^{q} \lambda_k w_{1k} v_k \\
\sum_{k=1}^{q} b_{2k} v_k &= \sum_{k=1}^{q} \lambda_k w_{2k} v_k \qquad (1.3.2.3) \\
&\dots \\
\sum_{k=1}^{q} b_{pk} v_k &= \sum_{k=1}^{q} \lambda_k w_{pk} v_k
\end{aligned}$$

где $\lambda_k$ – собственное число $B$, $v_k$ – собственные вектора $B$, $k = 1, \dots, q$.

Для построения КДФ система (1.3.2.3) решается относительно $\lambda_k$ и $v_k$. Количество получаемых нетривиальной решений этой системы уравнений равно $q$. Каждое из таких решений имеет свое значение $\lambda_k$и свой вектор $v_k = (v_{k1}, \dots, v_{kp})$ и определяет одну КДФ. Компоненты $v_{k1}, \dots, v_{kp}$ вектора $v_k, k = 1, \dots, q$ , используются для вычисления $q$ дискриминантных функций (1.3.2.1) коэффициенты каждой из которых имеют вид:

$$u_i = v_{ki}\sqrt{n-g},$$

$$u_o = -\sum_{i=1}^{p} u_i \overline{x_i},$$

$$i = 1, \dots, p, \qquad k = 1, \dots, q.$$

Коэффициенты $u_i$ приводят значения дискриминантной функции $f_{km}$ к стандартной форме. Это означает, что соответствующие дискриминантные значения по совокупности наблюдений будут иметь нулевое среднее и единичное внутригрупповое стандартное отклонение.

Такие функции будут находиться на каждом шаге. На первом шаге в модель будет введена та переменная, КДФ которой будет более значимой. Для этого проверяется нулевая гипотеза о равенстве центроидов во всех группах. Если дискриминантная функция статистически значимая, то эта гипотеза должна быть отвергнута. Эта гипотеза проверяется с помощью статистики *F*-включение:

$$F = \frac{(g-1)}{(n-p-g+1)} \sum_{k=1}^{g} n_k (\overline{X_k} - \bar{X})^T C^{-1} (\overline{X_k} - \bar{X}),$$

где $\overline{X_k} = (\overline{x_{1k}}, \dots, \overline{x_{pk}})$ , $\bar{X} = (\overline{x_1}, \dots, \overline{x_p})$, $C = \{c_{ij}\}$, $c_{ij=} \frac{t_{ij}}{(n-1)}$; а также с помощью коэффициента лямбда Уилкса:

$$L = \frac{\det W}{\det T},$$

значения которого будут находиться в интервале от 0 (полная дискриминация) до 1 (нет дискриминации). На каждом следующем шаге будут высчитываться КДФ с использованием переменных из модели с каждой не из неё. С какой переменной не из модели лямбда Уилкса будет меньше, а *F*- включение больше, та и будет включена в модель на этом шаге.

После того, как вид дискриминантных функций определяется, происходит процедура классификации.

### 1.3.3 Классификация

Классификация – это процесс принятия решения: указанный объект «принадлежит к» или «очень похож» на данную группу. Такое решение принимается на основе информации, содержащейся в дискриминантных переменных [19].

Для классификации в пошаговом дискриминантном анализе с включением применяется линейная комбинация, максимизирующая различия между классами, и вместе с этим минимизирующая дисперсию внутри классов. Такая линейная комбинация для каждой группы, называется классифицирующей функцией:

$$h_k = d_{k0} + d_{k1}x_1 + d_{k2}x_2 + \cdots + d_{kp}x_p \quad , k = 1, \dots, g,$$

где $x_i -$ значение $i$-й переменной, $h_k$ – значение функции для группы $k$; $d_{ki}$ – коэффициенты регрессии, которые необходимо найти; $d_{k0}$ – постоянный член. Объект относится к группе с наибольшим значением $h_k$. Коэффициенты для классифицирующих функций находятся по следующим формулам:

$$d_{ki} = (n - g)\sum_{j=1}^{p} m_{ij}\overline{x_{jk}}, \quad k = 1, \dots, g, \quad i = 1, \dots, p,$$

где $m_{ij}$ – элементы матрицы $W^{-1}$. Постоянный член $d_{k0}$ находится по формуле:

$$d_{k0} = -0.5 \sum_{j=1}^{p} d_{kj} \overline{x_{jk}}, \quad k = 1, \dots, g.$$

### 1.3.3.1 Расстояние Махалонобиса

Более наглядным способом классификации является измерение расстояний между объектом и каждым из центроидов групп, чтобы затем отнести объект в ближайшую группу. Однако, когда переменныеизмерены в разных единицах, коррелированы, имеют различные стандартные отклонения, сложно определить понятие «расстояния». Индийский статистик Махалонобис предложил обобщенную меру расстояния, которая устраняет эти трудности [19]. Она записывается в следующей форме:

$$D^2(X|G_k) = (n-g) \sum_{i=1}^{p} \sum_{j=1}^{p} m_{ij} \, (x_i - \overline{x_{ik}})\big(x_j - \overline{x_{jk}}\big),$$

где $D^2(X|G_k)$ – квадрат расстояния от точки $X = (x_1, \dots, x_p)$ (данный объект) до центроида $G_k$ класса $k$; $k = 1, \dots, g$. После вычисления $D^2$ для каждого класса классифицируем объект в группу с наименьшим $D^2$. Это будет группа, профиль которой по дискриминантным переменным больше похож на профиль данного объекта.

Если предположить, что каждый объект должен относиться к одной из групп, то можно вычислить вероятность его принадлежности к каждому из классов:

$$p(G_k|X) = \frac{\pi_k \exp(-0{,}5 \, D^2(X|G_k))}{\sum_{r=1}^{g} \pi_r \exp(-0{,}5 \, D^2(X|G_k))}, \quad k = 1, \dots, g.$$

Сумма этих вероятностей, часто называемых апостериорными, по всем группам равна 1. Объект принадлежит к той группе, для которой апостериорная вероятность максимальна, что эквивалентно использованию наименьшего расстояния. В данной формуле под символом $\pi_k$ понимается априорная вероятность, т.е. вероятность принадлежности объекта $k$–ому классу до учета экспериментальных данных.

Для корректного применения функции классификации стоит использовать две выборки: анализируемую, которая используется для вычисления функции и проверочную – для проверки результатов расчета на основании первой выборки. Такую процедуру называют кросс-проверкой.

# Глава 2. Решение поставленной задачи

## 2.1 Однофакторный дисперсионный анализ и непараметрические методы сравнения для фактора «Исход»

Рассмотрим влияние количественных факторов, перечисленных в постановке задачи на фактор № 10 (Исход) в группе 99_1 (без заболевания легких).

Для начала определим количественные переменные, которые не могут быть рассмотрены в дисперсионном анализе, исходя из ограничений, представленных в параграфе 1.1.1. Рассмотрим число наблюдений в каждой из двух групп для всех количественных переменных (проверим четвертое ограничение в 1.1.1).Например, в таблице 2.1.1 можно видеть число наблюдений для факторов № 178 (Ca),№ 179 (Mg):

| variable | Valid N Group 1 | Valid N Group 2 |
|---|---|---|
| Ca | 98 | 0 |
| Mg | 10 | 0 |

Табл. 2.1.1. Число наблюдений по уровню калия и магния для двух групп

Из данной таблицы видно, что в одной из групп число наблюдений меньше двойки, поэтому такие переменные мы рассматривать для дисперсионного анализа (ANOVA) не можем. Полный список неподходящих переменных по четвертому ограничению: Ca, Mg, Мочев. к-та, КДРпж1, ПП1, VE1,VA1, VE/VA1, IVRT1.

Для оставшихся количественных факторов продолжим исследование, которое подробно рассмотрим на примере влияния переменной «ИМТ» (на переменную «Исход»).

Проведем сначала проверку на нормальность распределения переменной «ИМТ» с помощью критерия Шапиро-Уилка и построения гистограммы (см. рисунок. 2.1.1) и выведем график, который изображает зависимость ожидаемых нормальных частот значений признака от их реальных частот. Очевидно, что если между наблюдаемым и ожидаемым распределениями нет никакой разницы, точки на этом графике выстроятся строго вдоль прямой (рисунок 2.1.2):

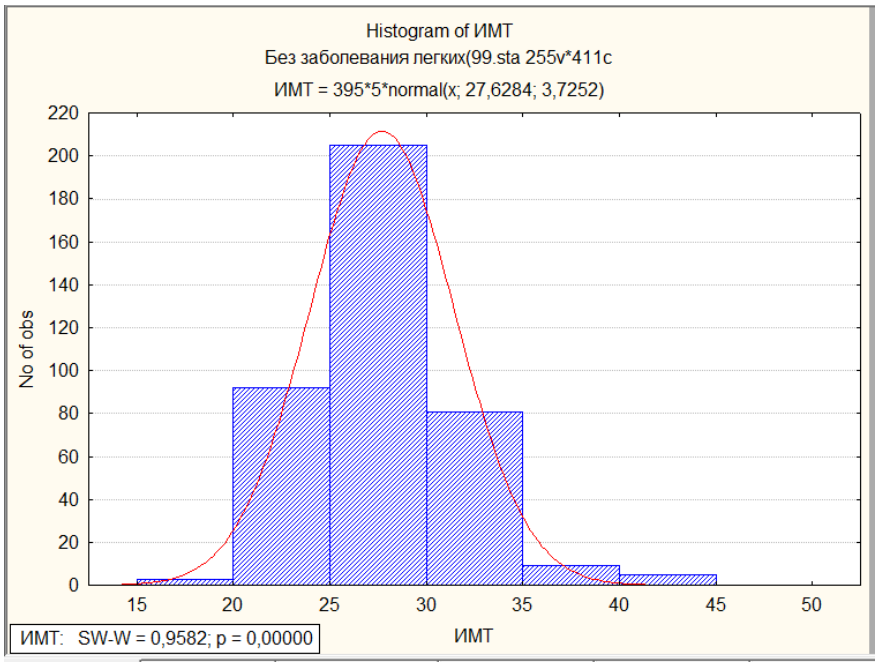


Рис. 2.1.1. Гистограмма для фактора «ИМТ»

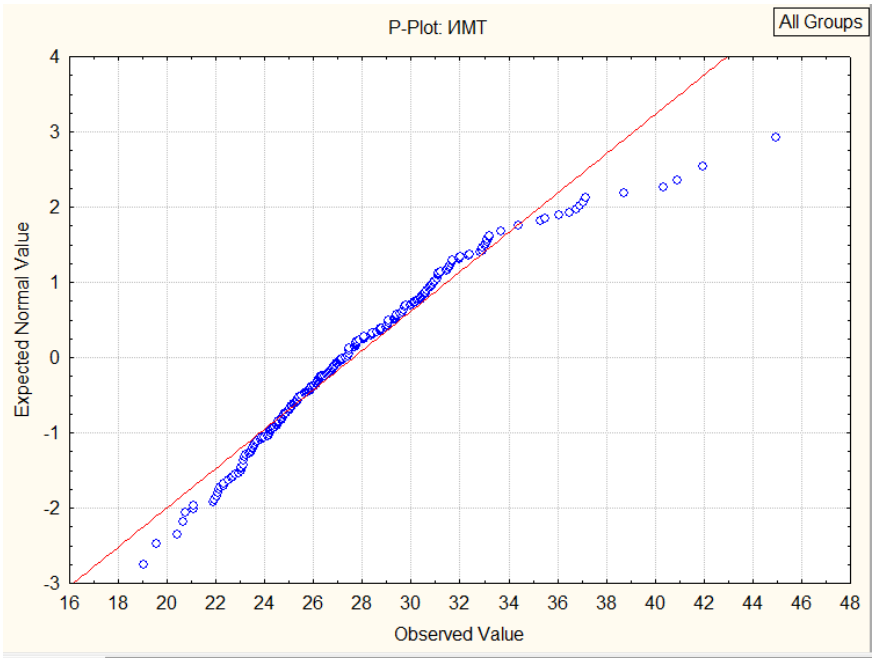


Рис. 2.1.2. График ожидаемых и реальных частот. Фактор «ИМТ»

Статистика Шапиро-Уилка *W* близка к 1, но достигнутый уровень согласия с нулевой гипотезой (распределение нормальное) *p*=0,00000. Таким образом, нельзя с уверенностью утверждать, что распределение нормальное. Но, как уже было отмечено ранее, дисперсионный анализ устойчив к небольшим отклонениям от стандартных предположений.

Кроме того, на рисунке 2.1.2 точки достаточно плотно выстраиваются вдоль теоретически ожидаемой прямой, что еще раз подтверждает нормальность распределения данных.

Проверим теперь гипотезу о равенстве дисперсий с помощью теста Левена. Получим следующие результаты (см. таблицу 2.1.2):

Levene's Test for Homogeneity of Variances (Без заболевания легких(99_1).sta)
Effect: "Исход"
Degrees of freedom for all F's: 1, 393

| | MS Effect | MS Error | F | p |
|---|---|---|---|---|
| ИМТ | 14,35158 | 5,740935 | 2,499868 | 0,114660 |

Табл.2.1.2. Проверка на равенство дисперсий. Тест Левена. Переменная «ИМТ»

*MS Error*, *MS Effect* – средние значения суммы квадратов, *F* – выборочное значение *F*-статистики, *p* – вычисленный уровень согласованности с нулевой гипотезой.

Как видно из таблицы 2.1.2, гипотеза о равенстве дисперсий принимается на уровне значимости *p*=0,114660 (это больше чем 0,05, поэтому считаем, что проверку на равенство дисперсий данная переменная прошла).

Проведем однофакторный дисперсионный анализ для переменной «ИМТ» (см. таблицу 2.1.3):

Univariate Tests of Significance for ИМТ (Без заболевания легких(99_1).sta)
Sigma-restricted parameterization
Effective hypothesis decomposition

| Effect | SS | Degr. of Freedom | MS | F | p |
|---|---|---|---|---|---|
| Intercept | 22088,67 | 1 | 22088,67 | 1592,042 | 0,000000 |
| Исход | 14,94 | 1 | 14,94 | 1,077 | 0,300071 |
| Error | 5452,65 | 393 | 13,87 | | |

Табл.2.1.3. ANOVA для «ИМТ»

Из таблицы 2.1.3 можно сделать вывод, что средние двух выборок не отличаются, значение уровня значимости $p$=0,30071, поэтому можно сказать, что ИМТ не влияет исход. На рисунке 2.1.3 также можно видеть предполагаемую разницу между средними:

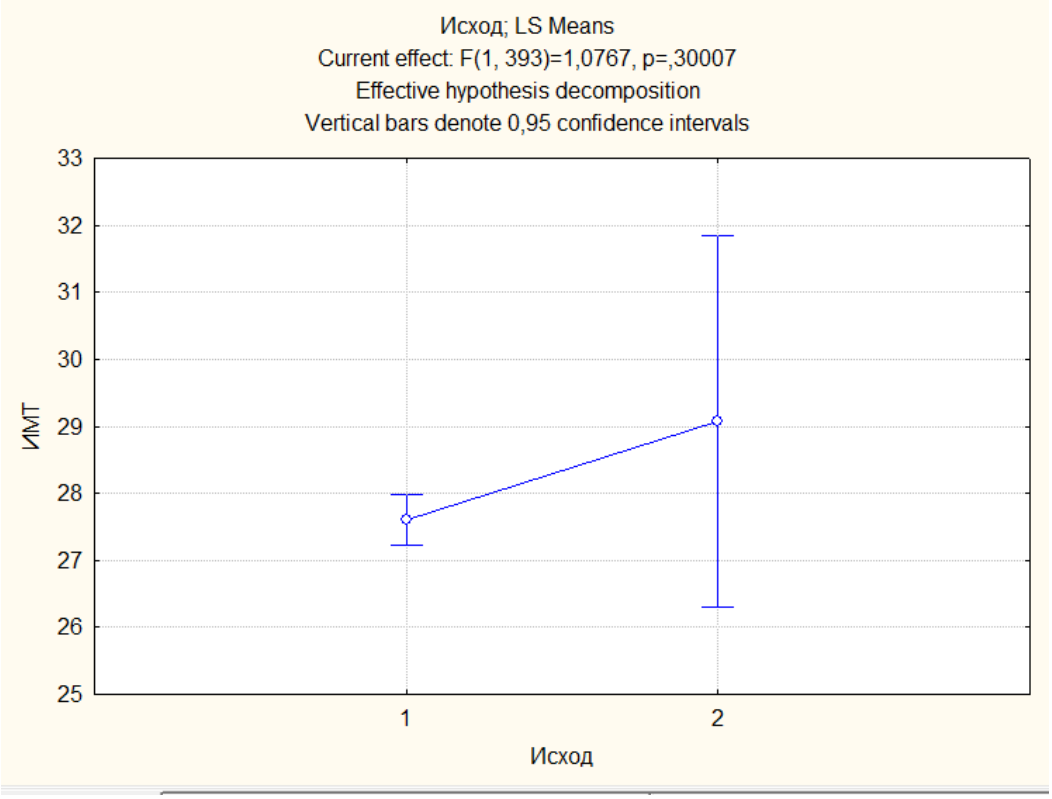


Рис.2.1.3. Разница между средними. «Исход» и «ИМТ»

Для тех переменных, для которых не выполняется ограничение на нормальность распределения и на равенство дисперсий, требуется перепроверить данные с помощью непараметрических методов сравнения. В качестве такого метода был выбран критерий Манна-Уитни.

Проводя исследование над переменными, которые не прошли ограничения ANOVA, получаем результаты теста Манна-Уитни – таблицу 2.1.4 и сводную таблицу 2.1.5:

| Variable | Wald-Wolfowitz Runs Test (Без заболевания легких(99_1).sta)<br>By variable Исход<br>Marked tests are significant at p <,05000<br>Valid N Group 1 | Valid N Group 2 | Mean Group 1 | Mean Group 2 | Z | p-level | Z adjstd | p-level | No. of Runs | No. of ties |
|---|---|---|---|---|---|---|---|---|---|---|
| **ИМТ** | 388 | 7 | 27,6023 | 29,0762 | -5,62381 | 0,000000 | 4,874346 | 0,000001 | 11 | 1 |
| АДсист | 385 | 23 | 139,3403 | 111,5217 | -2,54241 | 0,011010 | 2,307299 | 0,021039 | 39 | 31 |
| АДдиаст | 385 | 23 | 86,7870 | 65,2174 | -4,42328 | 0,000010 | 4,188174 | 0,000028 | 35 | 29 |
| ЧСС | 355 | 23 | 76,2901 | 80,0000 | -1,91038 | 0,056085 | 1,683014 | 0,092373 | 40 | 32 |
| ОХ/ЛПВП | 169 | 2 | 6,7250 | 5,1772 | -7,45336 | 0,000000 | 5,545386 | 0,000000 | 3 | 0 |
| Gluc | 335 | 5 | 6,1703 | 8,7800 | 0,28991 | 0,771885 | -0,695786 | 0,486564 | 11 | 6 |
| Рла1 | 380 | 6 | 32,5493 | 50,9950 | -3,14825 | 0,001643 | 2,280235 | 0,022594 | 11 | 0 |
| ЧСС1 | 373 | 6 | 74,2724 | 99,6667 | -4,83523 | 0,000001 | 3,974880 | 0,000070 | 10 | 6 |
| АДдиаст1 | 384 | 7 | 86,8906 | 88,5714 | 0,37385 | 0,708517 | -0,371942 | 0,709937 | 15 | 11 |

Табл.2.1.4. Результаты теста Манна-Уитни

| Количественные переменные | Проверка на нормализацию Параметр W | Проверка на нормализацию Параметр p | Равенство дисперсий Параметр p | ANOVA Параметр p | Вывод | Окончательный вердикт(с тестом Манна-Уитни) |
|---|---|---|---|---|---|---|
| Адсист(110) | 0,9715 | 0,00000 | 0,001152 | 0,000017 | Влияет(-) | Влияет |
| Аддиаст(111) | 0,9354 | 0,0000 | 0,00003 | 0,000000 | Влияет(-) | Влияет |
| ЛПНП(168) | 0,9635 | 0,0002 | 0,068283 | 0,043747 | Влияет | Влияет |
| Creatin(174) | 0,892 | 0,0000 | 0,445324 | 0,048265 | Влияет | Влияет |
| ЛП1(186) | 0,9606 | 0,00000 | 0,707051 | 0,032927 | Влияет | Влияет |
| ФВ1(195) | 0,9877 | 0,0062 | 0,989862 | 0,005891 | Влияет | Влияет |
| ФУ1(197) | 0,9772 | 0,00004 | 0,625693 | 0,015053 | Влияет | Влияет |
| RR(DT)1(202) | 0,9738 | 0,00000 | 0,121391 | 0,022222 | Влияет | Влияет |
| ИСПСН1(213) | 0,9844 | 0,0002 | 0,45239 | 0,000000 | Влияет | Влияет |
| RR-->DT1(216) | 0,9738 | 0,00000 | 0,121391 | 0,022222 | Влияет | Влияет |
| ИндНор1(222) | 0,9887 | 0,0028 | 0,079319 | 0,000002 | Влияет | Влияет |
| ИМТ(8) | 0,9582 | 0,00000 | 0,11466 | 0,3000071 | Не влияет(-) | Не влияет |
| ППТ(9) | 0,1545 | 0,0000 | 0,445476 | 0,758415 | Не влияет | Не влияет |
| АДмакс.(77) | 0,9246 | 0,0000 | 0,559476 | 0,053985 | Не влияет | Не влияет |
| АДмин.(78) | 0,8493 | 0,0000 | 0,274438 | 0,306373 | Не влияет | Не влияет |
| ЧСС(140) | 0,9426 | 0,0000 | 0,000027 | 0,373831 | Не влияет(-) | Не влияет |
| ОХ(166) | 0,9818 | 0,0008 | 0,059331 | 0,92609 | Не влияет | Не влияет |
| ТГ(167) | 0,8813 | 0,0000 | 0,084485 | 0,116491 | Не влияет | Не влияет |
| ЛПОНП(169) | 0,8141 | 0,0000 | 0,172522 | 0,427047 | Не влияет | Не влияет |
| ЛПВП(170) | 0,9162 | 0,00000 | 0,130928 | 0,466165 | Не влияет | Не влияет |
| КА(171) | 0,9496 | 0,00002 | 0,050806 | 0,642975 | Не влияет | Не влияет |
| ОХ/ЛПВП(172) | 0,9421 | 0,00000 | 0,072341 | 0,280625 | Не влияет(-) | Не влияет |
| ПИ(173) | 0,9815 | 0,0025 | 0,067647 | 0,96089 | Не влияет | Не влияет |
| Gluc(175) | 0,7936 | 0,0000 | 0,00001 | 0,004472 | Влияет(+) | Не влияет |
| Na(176) | 0,9675 | 0,00002 | 0,570231 | 0,414481 | Не влияет | Не влияет |
| K(177) | 0,99 | 0,0766 | 0,765666 | 0,828852 | Не влияет | Не влияет |
| Cl(180) | 0,9685 | 0,0017 | 0,859319 | 0,72784 | Не влияет | Не влияет |
| ЛПНП/ЛПВП(182) | 0,9345 | 0,00000 | 0,079217 | 0,261477 | Не влияет | Не влияет |
| ДА1(185) | 0,9896 | 0,0131 | 0,789704 | 0,082279 | Не влияет | Не влияет |
| КДР1(187) | 0,9715 | 0,00000 | 0,19831 | 0,58486 | Не влияет | Не влияет |
| КСР1(188) | 0,9739 | 0,00001 | 0,67305 | 0,24802 | Не влияет | Не влияет |
| МЖП1(189) | 0,9904 | 0,0152 | 0,978769 | 0,617414 | Не влияет | Не влияет |
| ЗС1(190) | 0,9843 | 0,0004 | 0,477146 | 0,675505 | Не влияет | Не влияет |
| КДО1(193) | 0,9223 | 0,0000 | 0,235871 | 0,771921 | Не влияет | Не влияет |
| КСО1(194) | 0,8959 | 0,0000 | 0,579589 | 0,321322 | Не влияет | Не влияет |
| УО1(203) | 0,9874 | 0,0027 | 0,97444 | 0,368944 | Не влияет | Не влияет |
| УИ1(204) | 0,9865 | 0,0016 | 0,592751 | 0,175015 | Не влияет | Не влияет |
| СВ1(205) | 0,9268 | 0,0000 | 0,189085 | 0,845663 | Не влияет | Не влияет |
| СИ1(206) | 0,9137 | 0,0000 | 0,11411 | 0,78244 | Не влияет | Не влияет |
| Рла1(207) | 0,8097 | 0,0000 | 0,000663 | 0,005405 | Влияет(+) | Не влияет |
| ЧСС1(208) | 0,9566 | 0,00000 | 0,000011 | 0,000151 | Влияет(+) | Не влияет |
| Рсрлао(Адср)1(210) | 0,9854 | 0,0005 | 0,005579 | 0,598121 | Не влияет | Не влияет |
| ОПСС1(211) | 0,8679 | 0,0000 | 0,459 | 0,467617 | Не влияет | Не влияет |
| ОЛС1(212) | 0,7372 | 0,0000 | 0,23605 | 0,356858 | Не влияет | Не влияет |
| Адсист1(217) | 0,9791 | 0,00002 | 0,033167 | 0,446604 | Не влияет | Не влияет |
| Аддиаст1(219) | 0,9608 | 0,00000 | 0,007849 | 0,800953 | Не влияет(-) | Не влияет |
| МО1(220) | 0,9275 | 0,0000 | 0,191611 | 0,830659 | Не влияет | Не влияет |
| ОТС1(223) | 0,9669 | 0,00000 | 0,170674 | 0,360987 | Не влияет | Не влияет |
| МЛЖ(224) | 0,9773 | 0,00005 | 0,427641 | 0,971914 | Не влияет | Не влияет |
| ИММЛЖ1(225) | 0,9835 | 0,0008 | 0,309921 | 0,598218 | Не влияет | Не влияет |

Табл.2.1.5. Сводная таблица. ANOVA + тест Манна Уитни

В столбце «Вывод» скобки, содержащие «+» или «−», стоят в тех ячейках, которые соответствуют переменным, не прошедшим ограничения на равенство дисперсий. Знак «+» означает что тест Манна-Уитни изменил «решение» дисперсионного анализа («−» не изменил). В столбце «Окончательный вердикт» сначала выведены те переменные, которые влияют на «Исход». Их оказалось 11.

Аналогично рассмотрим результаты влияния количественных факторов на фактор № 10 (Исход) в группе 99_23 (с заболеванием легких).

Только 7 переменных удовлетворили условию 4 из пункта 1.1.1.
Такими переменными оказались АДмакс, АДмин, АДсист, АДдиаст, ЧСС, ИСПСН1,ИндНор1. Все они удовлетворяют требованиям нормального распределения и равенства дисперсий.

В итоге из них оказала влияние на «Исход» лишь одна: ИСПСН1.

## 2.2 Однофакторный дисперсионный анализ и непараметрические методы сравнения для фактора «Осложнения»

Рассмотрим влияние количественных факторов на фактор № 20 (Осложнения) в группе 99_1 (без заболевания легких).

Проводя аналогичные рассуждения и исследования как в пункте 2.1, получаем следующие результаты (табл.2.2.1):

| Количественные переменные | Проверка | ANOVA, проверка н | Влияние теста Манна-Уитни | Окончательный вывод |
|---|---|---|---|---|
| Адсист(110) | нет | да | да | да |
| Аддиаст(111) | нет | да | да | да |
| ЛПОНП(169) | нет | да | да | да |
| ЛП1(186) | нет | да | да | да |
| КДР1(187) | нет | да | да | да |
| КСР1(188) | нет | да | да | да |
| КДО1(193) | нет | да | да | да |
| КСО1(194) | нет | да | да | да |
| ФУ1(197) | нет | да | да | да |
| УО1(203) | нет | да | да | да |
| УИ1(204) | нет | да | да | да |
| ОЛС1(212) | нет | да | да | да |
| ИндНор1(222) | нет | да | да | да |
| ППТ(9) | нет | нет | нет | нет |
| ЧСС(140) | нет | да | нет | нет |
| КА(171) | нет | да | нет | нет |
| ОХ/ЛПВП(172) | нет | да | нет | нет |
| ПИ(173) | нет | да | нет | нет |
| Са(178) | нет | да | нет | нет |
| VA1(199) | нет | да | нет | нет |
| IVRT1(201) | нет | да | нет | нет |
| СВ1(205) | нет | да | нет | нет |
| СИ1(206) | нет | да | нет | нет |
| Рла1(207) | нет | да | нет | нет |
| ЧСС1(208) | нет | да | нет | нет |
| ОПСС1(211) | нет | да | нет | нет |
| Аддиаст1(219) | нет | да | нет | нет |
| МО1(220) | нет | да | нет | нет |
| МЛЖ(224) | нет | да | нет | нет |
| ИММЛЖ1(225) | нет | да | нет | нет |
| ИМТ(8) | да | нет | | нет |
| АДмакс.(77) | да | нет | | нет |
| АДмин.(78) | да | нет | | нет |
| ОХ(166) | да | нет | | нет |
| ТГ(167) | да | нет | | нет |
| ЛПНП(168) | да | нет | | нет |
| ЛПВП(170) | да | нет | | нет |
| Creatin(174) | да | нет | | нет |
| Gluc(175) | да | нет | | нет |
| Na(176) | да | нет | | нет |
| K(177) | да | нет | | нет |
| Mg(179) | да | нет | | нет |
| Cl(180) | да | нет | | нет |
| Мочев. к-та(181) | да | нет | | нет |
| ЛПНП/ЛПВП(182) | да | нет | | нет |
| ДА1(185) | да | нет | | нет |
| МЖП1(189) | да | нет | | нет |
| ЗС1(190) | да | нет | | нет |
| КДРпж1(191) | да | нет | | нет |
| ФВ1(195) | да | нет | | нет |
| VE1(198) | да | нет | | нет |
| VE/VA1(200) | да | нет | | нет |
| RR(DT)1(202) | да | нет | | нет |
| Рсрлао(Адср)1(210) | да | нет | | нет |
| ИСПСН1(213) | да | нет | | нет |
| RR-->DT1(216) | да | нет | | нет |
| Адсист1(217) | да | нет | | нет |
| ОТС1(223) | да | нет | | нет |

Табл. 2.2.1. Сводная таблица для фактора «Осложнения»

В данной таблице столбец «Влияние теста Манна-Уитни» содержит значения только для тех переменных, которые не удовлетворили условию равенства дисперсий. Значение в этом столбце соответствует результату теста.

В столбце «Окончательный вывод» мы можем видеть 13 влияющих на «Осложнения» количественных переменных.

Аналогично рассмотрим влияние количественных факторов на фактор № 20 (Осложнения) в группе 99_23 (с заболеванием легких).

В итоге получаем, что в данной группе на параметр № 20 (Осложнения) оказали влияние 18 факторов: Адсист (110), Аддиаст (111), ЧСС (140), Creatin (174), КДР1 (187), КСР1 (188), КДО1 (193), КСО1 (194), ИСПСН1 (213), ЛПОНП (169), ЛП1 (186), ФВ1 (195), ФУ1 (197), VE1 (198), VA1 (199), IVRT1 (201), ИндНор1 (222), ОТС1 (223). В скобках указаны номера переменных в исходной базе данных пациентов.

## 2.3 Дискриминантный анализ для фактора «Осложнения»

Рассмотрим применение дискриминантного анализа к поставленной задаче. Необходимо проверить, какие переменные являются наиболее значимыми при прогнозировании осложнений после инфаркта миокарда в двух случаях: у пациентов с заболеваниями легких и без подобного заболевания.

Результаты, полученные после применения дисперсионного анализа (ANOVA) и список факторов, влияющих на параметр «Осложнения» (№ 21) можно найти в параграфе 2.2. Именно их будем рассматривать в данном виде анализа. Предположения, необходимые для применения дискриминантного анализа отдельно проверять не будем, так как идентичные проверки были выполнены на этапе ANOVA.

Сначала рассмотрим группу 99_1 (пациенты без заболевания легких).

Занесем во вкладку дискриминантного анализа в программном пакете *STATISTICA* все 13 факторов, выявленные на предыдущем этапе работы. В силу условий, высказанных в пункте 2.1, будем использовать пошаговый дискриминантный анализ с включением (Forward).

Установленные параметры для значений толерантности и *F*-включения можно увидеть на рисунке 2.3.1:

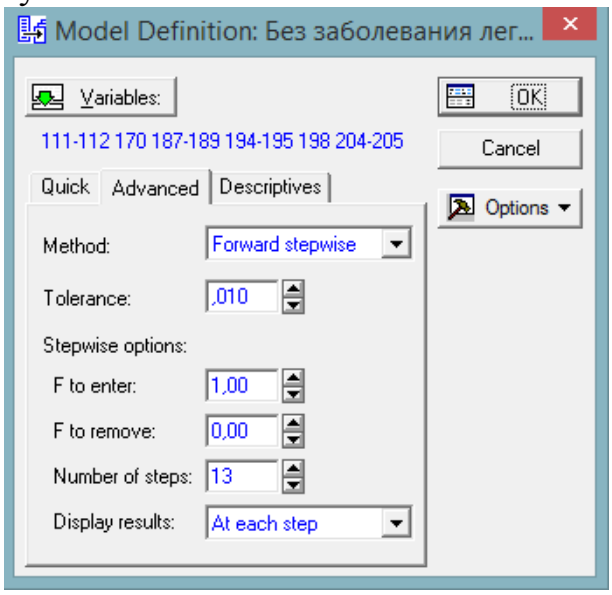


Рис. 2.3.1. STATISTICA. Дискриминантный анализ. Установка параметров

Дополнительным условием было обозначено, чтобы результаты выводились для каждого шага. Это обеспечит большую наглядность и понимания анализа.

В общем случае анализ будет проводиться до тех пор, пока не произойдет одно из четырех событий:

- все переменные введены или отброшены;
- достигнуто максимально установленное число шагов;
- не осталось переменных вне модели, имеющих уровень значимости *F*, большее чем значение *F*-включить;
- следующая переменная имеет значение толерантности меньше заданного (0,01).

Результат после первого шага представлен на рисунке 2.3.2:

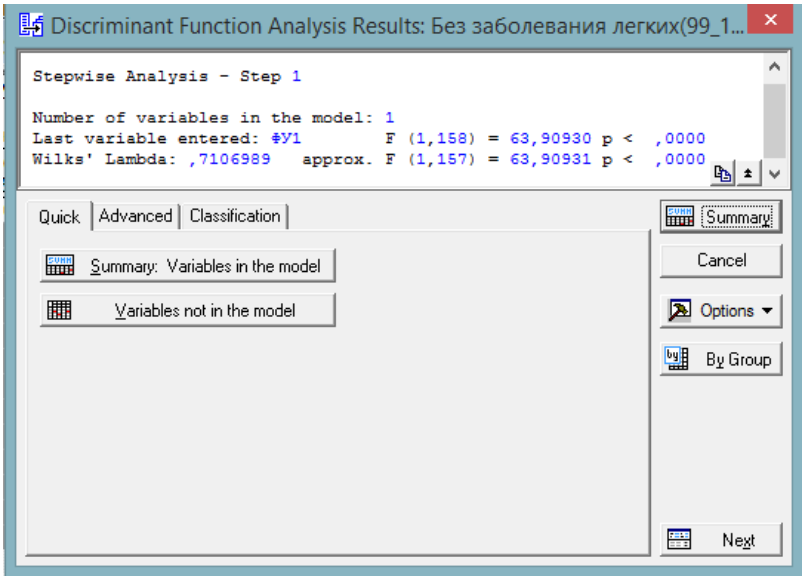


Рис 2.3.2. Первый шаг пошагового дискриминантного анализа с включением

Видно, что на первом шаге в модель была введена переменная ФУ1, а статистика Лямбда-Уилкса начала уменьшаться (на этом шаге она равна 0,71), следовательно дискриминация увеличивается. На каждом шаге также можно просмотреть переменные вне модели вместе с их показателями, в частности значение толерантности.

Аналогично проведем следующие шаги анализа. На рисунке 2.3.3 можно увидеть результаты после финального шага:

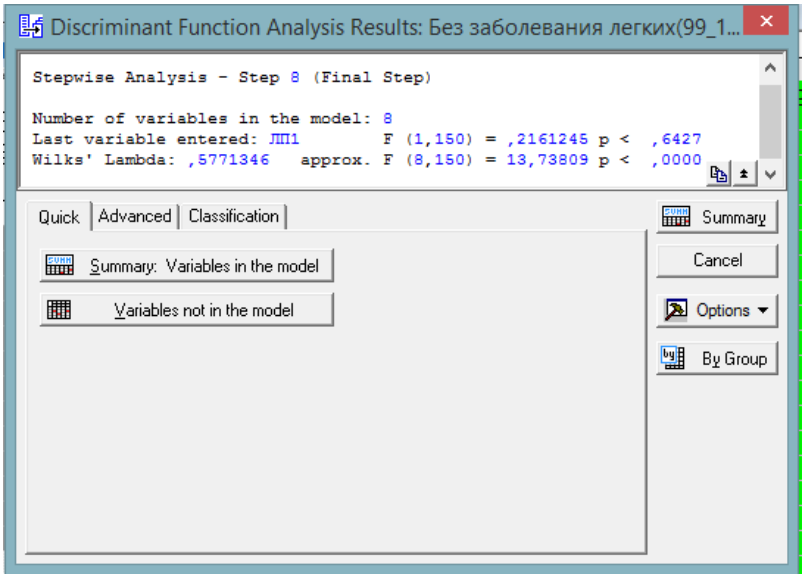


Рис.3.1.3. Результаты после заключительного шага

Всего получилось 8 шагов. Статистика Лямбда-Уилкса в конечном итоге достигла отметки 0,578.

Посмотрим на переменные, которые оказались в модели (таблица 2.3.1):

Discriminant Function Analysis Summary (Без заболевания легких(99_1).sta)
Step 8, N of vars in model: 8; Grouping: ослож (2 grps)
Wilks' Lambda: ,57713 approx. F (8,150)=13,738 p< ,0000

| N=159 | Wilks' Lambda | Partial Lambda | F-remove (1,150) | p-level | Toler. | 1-Toler. (R-Sqr.) |
|---|---|---|---|---|---|---|
| **ФУ1** | 0,584692 | 0,987075 | 1,96413 | 0,163138 | 0,494950 | 0,505050 |
| ИндНор1 | 0,636210 | 0,907145 | 15,35391 | 0,000135 | 0,752609 | 0,247391 |
| КСР1 | 0,605453 | 0,953228 | 7,36000 | 0,007449 | 0,060175 | 0,939825 |
| АДсист | 0,612540 | 0,942198 | 9,20214 | 0,002850 | 0,241678 | 0,758322 |
| АДдиаст | 0,598865 | 0,963715 | 5,64773 | 0,018740 | 0,249158 | 0,750842 |
| КСО1 | 0,588178 | 0,981225 | 2,87018 | 0,092310 | 0,068181 | 0,931819 |
| ЛПОНП | 0,584977 | 0,986594 | 2,03826 | 0,155463 | 0,937883 | 0,062116 |
| ЛП1 | 0,584176 | 0,987946 | 1,83016 | 0,178144 | 0,687416 | 0,312584 |

Табл. 2.3.1. Переменные, включенные в модель

Красным цветом выделены значимые переменные. Их оказалось 4: ИндНор1, КСР1, АДсист, АДдиаст. Присутствие остальных переменных, кроме перечисленных, не так важно для дискриминации, поэтому дальше используем только значимые (см. табл. 2.3.2):

Discriminant Function Analysis Summary (Без заболевания легких(99_1).sta)
Step 4, N of vars in model: 4; Grouping: ослож (2 grps)
Wilks' Lambda: ,62215 approx. F (4,154)=23,382 p< ,0000

| N=159 | Wilks' Lambda | Partial Lambda | F-remove (1,154) | p-level | Toler. | 1-Toler. (R-Sqr.) |
|---|---|---|---|---|---|---|
| **КСР1** | 0,752791 | 0,826460 | 32,33698 | 0,000000 | 0,839709 | 0,160291 |
| ИндНор1 | 0,691512 | 0,899697 | 17,16865 | 0,000056 | 0,886666 | 0,113334 |
| АДсист | 0,656742 | 0,947330 | 8,56222 | 0,003952 | 0,251292 | 0,748708 |
| АДдиаст | 0,640041 | 0,972050 | 4,42813 | 0,036975 | 0,264435 | 0,735565 |

Табл. 2.3.2. Повторение процедуры дискриминантного анализа для значимых переменных

Все переменные оказались значимы. Наиболее значима оказалась переменная КСР1.

Проводим анализ далее. Построим классифицирующие и дискриминантные функции, предварительно расставив априорные вероятности равными друг другу (обе по 0,5). Коэффициенты для классифицирующих функций содержатся в таблице 2.3.3:

| Variable | Classification Functions; grouping: ослож (Без заболевания легких(99_1).sta) G_1:1 p=,51572 | G_2:2 p=,48428 |
|---|---|---|
| КСР1 | 8,0369 | 9,4877 |
| ИндНор1 | 0,7633 | 1,0765 |
| АДсист | 0,2670 | 0,3126 |
| АДдиаст | 0,0886 | 0,0390 |
| Constant | -39,7573 | -50,1281 |

Табл. 2.3.3. Процедура классификации. Коэффициенты

Получаем две классифицирующие функции. Для первой группы (неосложненный инфаркт миокарда):

$h_1$ = 8,0369*(КСР1) + 0,7633*(ИндНор1) +0,2670*(АДсист) +0,0886*(АДдиаст) -39,7573.

Для второй группы (осложненный инфаркт миокарда):

$$h_2 = 9{,}4877 * (\text{КСР1}) + 1{,}0765 * (\text{ИндНор1}) + 0{,}3126 * (\text{АДсист}) + 0{,}0390 * (\text{АДдиаст}) - 50{,}1281.$$

Процент правильного прогноза представлен в таблице 2.3.4:

| Group | Classification Matrix (Без заболевания легких(99_1).sta) Rows: Observed classifications Columns: Predicted classifications Percent Correct | G_1:1 p=,51572 | G_2:2 p=,48428 |
|---|---|---|---|
| G_1:1 | 82,89474 | 126 | 26 |
| G_2:2 | 64,83517 | 64 | 118 |
| Total | 73,05389 | 190 | 144 |

Табл. 2.3.4.Процент правильного прогноза

С помощью построенных классифицирующих функций удалось правильно спрогнозировать 73% из имеющихся наблюдений, в том числе 82,89% в группе (Осложнения, 1) и 64,83% в группе (Осложнения, 2).

Теперь построим дискриминантную функцию (так как группы две, то ДФ будет одна) (табл. 2.3.5):

| Variable | Raw Coefficients (Без заболевания легких(99_1).sta) for Canonical Variables<br>Root 1 |
|---|---|
| КСР1 | 0,93627 |
| ИндНор1 | 0,20210 |
| АДсист | 0,02943 |
| АДдиаст | -0,03198 |
| Constant | -6,62780 |
| Eigenval | 0,60733 |
| Cum.Prop | 1,00000 |

Табл. 2.3.5. Коэффициенты дискриминантной функции

Тогда дискриминантная функция имеет вид:

$$f = 0{,}93627 * (\text{КСР1}) + 0{,}20210 * (\text{ИндНор1}) + 0{,}02943 * (\text{АДсист}) - 0{,}03198 * (\text{АДдиаст}) - 6{,}62780.$$

Посмотрим на среднее значение данной ДФ для двух групп (таблица 2.3.6):

| Group | Means of Canonical Variables (Без заболевания легких(99_1).sta)<br>Root 1 |
|---|---|
| G_1:1 | -0,750414 |
| G_2:2 | 0,799142 |

Табл. 2.3.6.Среднее значение дискриминантной функции для двух групп

Можно задать некоторый порог -0,7504 < $c$ < 0,799142 и считать, что если $R > c$, то наблюдение принадлежит второй группе, если $R < c$, то первой. Чем правее выберем $c$, тем лучше будет «угадываться» 1-ая группа, и тем хуже 2-ая, и наоборот. Если взять $c$ за середину отрезка, соединяющего центры двух групп, то результаты дискриминации совпадут с результатами, даваемыми классификационными функциями при априорных вероятностях 0,5 и 0,5.

По классифицирующим функциям и проценту правильного прогноза заметим, что процент «угадывания» в двух группах не сбалансирован (82% и 64%). Причем ошибка угадывания чаще происходит во второй группе (с

осложнениями), что может грозить тем, что пациенту с осложнением болезни выдаст диагноз «без осложнений». Данное статистическое исследование напрямую связано со здоровьем и жизнью пациентов, поэтому такая погрешность недопустима. В связи с этим, сбалансируем процент угадывания, путем расстановки априорных вероятностей как 0,4 и 0,6 соответственно в первой и второй группах. В таком случае получим следующие результаты (таблица 2.3.7 и 2.3.8):

Classification Matrix (Без заболевания легких(99_1).sta)
Rows: Observed classifications
Columns: Predicted classifications

| Group | Percent Correct | G_1:1 p=,40000 | G_2:2 p=,60000 |
|---|---|---|---|
| **G_1:1** | 70,39474 | 107 | 45 |
| G_2:2 | 71,42857 | 52 | 130 |
| Total | 70,95808 | 159 | 175 |

Табл.2.3.7. Сбалансированный процент прогнозирования

Classification Functions; grouping: ослож (Без заболевания легких(99_1).sta)

| Variable | G_1:1 p=,40000 | G_2:2 p=,60000 |
|---|---|---|
| **КСР1** | 8,0369 | 9,4877 |
| ИндНор1 | 0,7633 | 1,0765 |
| АДсист | 0,2670 | 0,3126 |
| АДдиаст | 0,0886 | 0,0390 |
| Constant | -40,0114 | -49,9138 |

Табл. 2.3.8. Обновленные коэффициенты классифицирующих функций

Общий процент угадывания немного снизился, но был достигнут баланс в группах.

Также в таблице 2.3.9 можно посмотреть классификацию случаев обучающей выборки. Неправильно угаданные варианты помечены звездочкой. В таблице 2.3.10 указаны квадраты расстояния Махаланобиса. Случай относится к группе, до которой расстояние Махаланобиса минимально

| Case | Classification of Cases (Без заболевания легких(99_1).sta)<br>Incorrect classifications are marked with *<br>Observed Classif. | 1<br>p=,40000 | 2<br>p=,60000 |
|---|---|---|---|
| *177,000000 | G_1:1 | G_2:2 | G_1:1 |
| 179,000000 | G_1:1 | G_1:1 | G_2:2 |
| 216,000000 | G_2:2 | G_2:2 | G_1:1 |
| 229,000000 | G_2:2 | G_2:2 | G_1:1 |
| 232,000000 | G_2:2 | G_2:2 | G_1:1 |
| 251,000000 | G_2:2 | G_2:2 | G_1:1 |
| 252,000000 | G_2:2 | G_2:2 | G_1:1 |
| 261,000000 | G_2:2 | G_2:2 | G_1:1 |
| 274,000000 | G_2:2 | G_2:2 | G_1:1 |
| 275,000000 | G_2:2 | G_2:2 | G_1:1 |
| *308,000000 | G_1:1 | G_2:2 | G_1:1 |
| *310,000000 | G_1:1 | G_2:2 | G_1:1 |
| 315,000000 | G_1:1 | G_1:1 | G_2:2 |
| 327,000000 | G_2:2 | G_2:2 | G_1:1 |
| 333,000000 | G_1:1 | G_1:1 | G_2:2 |
| 334,000000 | G_2:2 | G_2:2 | G_1:1 |
| *335,000000 | G_1:1 | G_2:2 | G_1:1 |
| 337,000000 | G_1:1 | G_1:1 | G_2:2 |
| *338,000000 | G_1:1 | G_2:2 | G_1:1 |
| 344,000000 | G_1:1 | G_1:1 | G_2:2 |
| *346,000000 | G_1:1 | G_2:2 | G_1:1 |
| 347,000000 | G_1:1 | G_1:1 | G_2:2 |
| *348,000000 | G_2:2 | G_1:1 | G_2:2 |
| 351,000000 | G_1:1 | G_1:1 | G_2:2 |
| 353,000000 | G_1:1 | G_1:1 | G_2:2 |
| *354,000000 | G_1:1 | G_2:2 | G_1:1 |
| 364,000000 | G_1:1 | G_1:1 | G_2:2 |
| 365,000000 | G_1:1 | G_1:1 | G_2:2 |

Табл. 2.3.9. Классификация случаев выборки

| Case | Squared Mahalanobis Distances from Group Centroids (Без заболевания легких(99_1).sta)<br>Incorrect classifications are marked with *<br>Observed Classif. | G_1:1<br>p=,40000 | G_2:2<br>p=,60000 |
|---|---|---|---|
| *410,000000 | G_1:1 | 2,47987 | 2,31698 |
| 413,000000 | G_2:2 | 5,63515 | 1,62068 |
| 415,000000 | G_2:2 | 1,82162 | 1,21625 |
| 428,000000 | G_2:2 | 1,84769 | 0,45716 |
| 429,000000 | G_1:1 | 2,60736 | 4,97708 |
| *431,000000 | G_1:1 | 2,03932 | 0,67857 |
| *438,000000 | G_1:1 | 1,50320 | 1,47641 |
| 446,000000 | G_2:2 | 5,83393 | 1,22787 |
| 450,000000 | G_1:1 | 3,93348 | 11,50159 |
| *451,000000 | G_1:1 | 3,35052 | 4,14880 |
| 452,000000 | G_2:2 | 2,24834 | 0,31913 |
| *454,000000 | G_1:1 | 3,05002 | 0,81800 |
| 457,000000 | G_1:1 | 2,74435 | 8,18593 |
| *461,000000 | G_1:1 | 4,58452 | 3,89771 |
| 463,000000 | G_1:1 | 1,97320 | 2,78864 |
| 470,000000 | G_1:1 | 3,77969 | 6,09581 |
| 474,000000 | G_1:1 | 2,37649 | 6,72811 |
| *475,000000 | G_1:1 | 2,67821 | 1,33677 |
| 476,000000 | G_1:1 | 1,81969 | 5,85689 |
| 479,000000 | G_1:1 | 7,21884 | 12,33256 |
| 481,000000 | G_2:2 | 5,39510 | 1,45941 |
| 482,000000 | G_1:1 | 1,11180 | 5,08456 |
| 483,000000 | G_2:2 | 3,73056 | 3,15635 |
| 485,000000 | G_1:1 | 7,00216 | 12,73851 |
| 487,000000 | G_2:2 | 4,20897 | 1,05294 |
| *499,000000 | G_2:2 | 2,83064 | 4,71192 |
| 502,000000 | G_1:1 | 1,33752 | 3,69360 |
| 506,000000 | G_2:2 | 10,50937 | 9,07988 |

Табл. 2.3.10. Квадраты расстояния Махаланобиса

Аналогично рассмотрим группу 99_23(пациенты с заболеванием легких). Изначально имеем 18 факторов. Значимыми оказались лишь три (таблица 2.3.11):

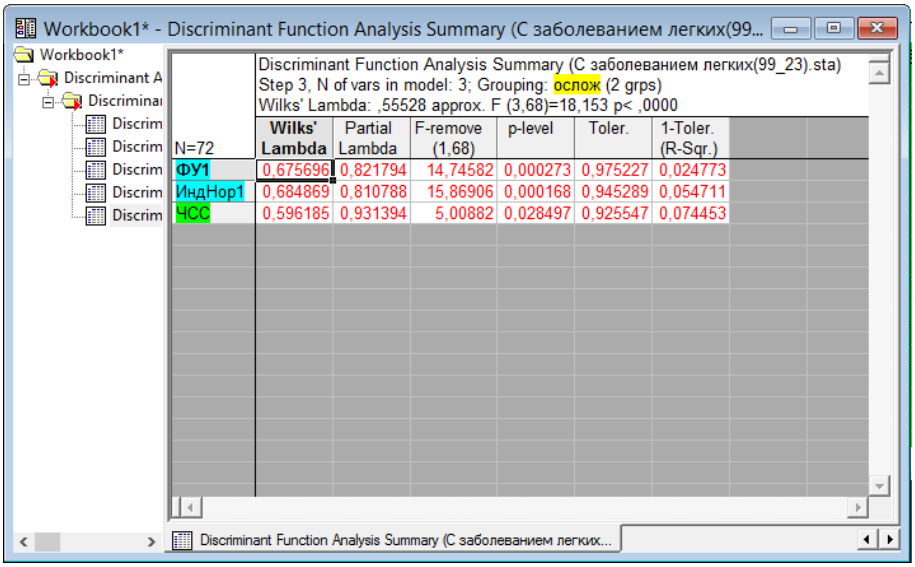


Табл. 2.3.11. Значимые факторы для группы с заболеванием легких

При априорных вероятностях 0,5 и 0,5 получаем следующий процент прогноза (таблица 2.3.12) и следующие коэффициенты классифицирующих функций (таблица 2.3.13), а также единственную дискриминантную функцию (таблица 2.3.14):

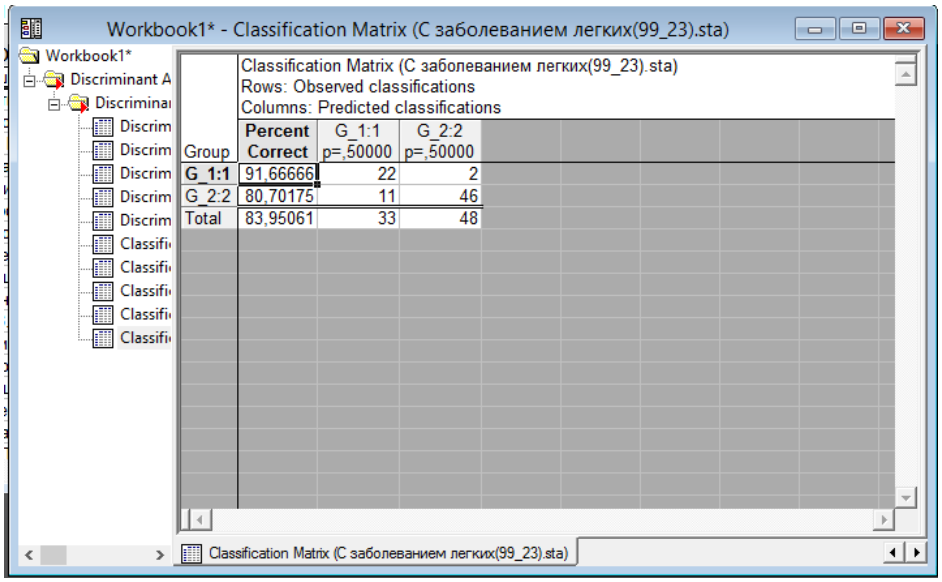


Табл. 2.3.12. Процент прогнозирования для второй группы

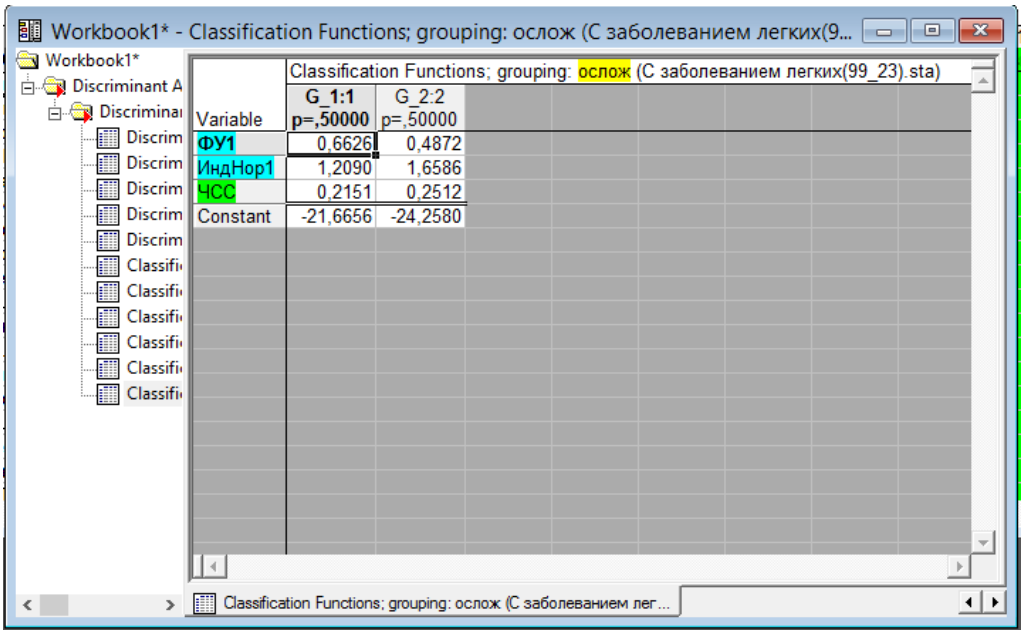


Табл. 2.3.13. Коэффициенты классифицирующих функций для второй группы

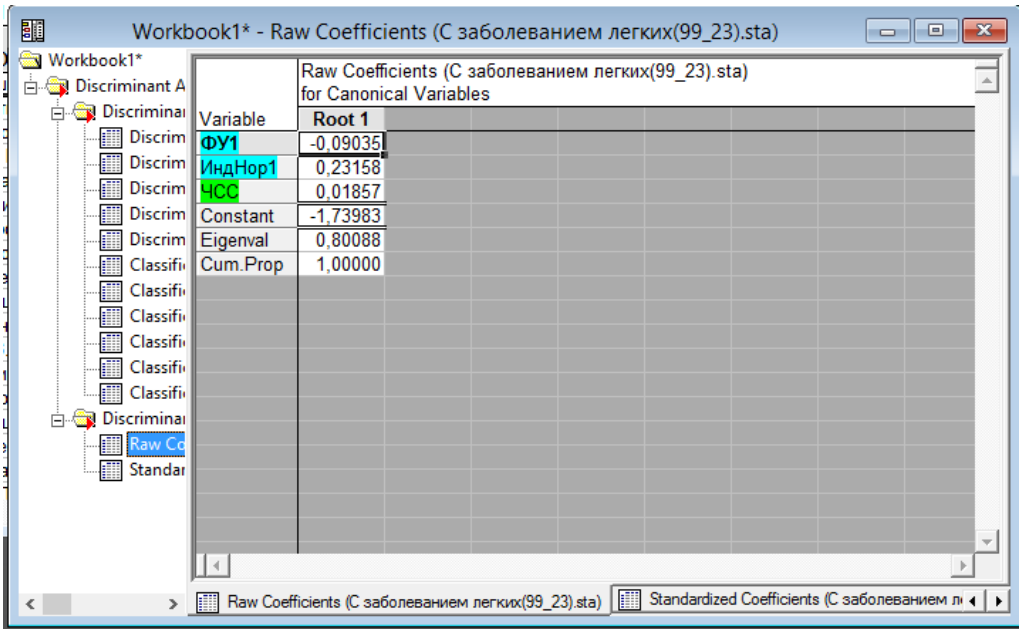


Табл. 2.3.14. Коэффициенты дискриминантной функции для второй группы

После этапов проведения дискриминантного анализа, построения классифицирующих функций для параметра № 20 «Осложнения», было реализовано приложение на языке *C#* с пользовательским интерфейсом для прогнозирования осложнений у поступивших пациентов. Описание данной программы и результаты ее работы можно найти в Приложении А.

Дискриминантный анализ для параметра «Исход» не был применен в связи с тем, что исходная база данных в данной переменной содержит резко различающиеся по количеству признаки, а именно количество выживших

пациентов резко превалирует над умершими. Кроме этого, число смертельных исходов в каждой из групп слишком мало, чтобы применять такой инструмент, как дискриминантный анализ. В связи с этим, выявленные после ANOVA переменные не прошли ограничения, указанные в параграфе 1.3 (конкретно ограничения 2 и 3).

# Выводы

В результате данной работы на основе базы пациентов с помощью дисперсионного анализа, непараметрических методов сравнения и дискриминантного анализа в среде *STATISTICA* были установлены зависимости между количественными факторами и интересующими нас переменными. Причем сделано это было в двух группах – у пациентов с заболеваниями легких и у пациентов без данного заболевания. Это позволило выяснить некоторые закономерности, объясняющие разницу в группах и связь между заболеванием легких и ИМ.

В обеих группах для переменной «Осложнения» был проведен дисперсионный и дискриминантный анализ в совокупности, в частности найдены особо влияющие факторы на наличие осложнений, построены дискриминантные и классифицирующие функции, позволяющие производить прогнозирование в дальнейшем.

В первой группе (без заболевания легких) наибольшее влияние оказывает фактор КСР1. Кроме этого, значимое влияние оказывают ИндНор1, АДсист, АДдиаст.

Во второй группе (с заболеванием легких) наибольшее влияние оказывает переменная ФУ1. Кроме нее, оказывают влияние такие факторы, как ИндНор1 и ЧСС.

В итоге выяснилось, что в этих двух группах общий значимый фактор лишь один – ИндНор1. Остальные же факторы абсолютно разные, причем в первой группе значимы оказались четыре переменные, а во второй лишь три. Кроме того, надо заметить, что во второй группе намного лучший результат в плане успешного процента квалификации (84% против 71%).

На основе данных результатов с целью удобства дальнейшего прогнозирования осложнений была написана программа с пользовательским интерфейсом, реализующая найденные классифицирующие функции.

Для переменной «Исход» в обеих группах был проведен лишь дисперсионный анализ и непараметрические методы сравнения. Дискриминантный анализ не был реализован в связи с недостатком количества данных в исходной базе пациентов. В данном случае можно опираться на результаты ANOVA, либо провести дискриминантный анализ с другой подобной базой данных, которая позволит это сделать.

Обобщая вышесказанное, можно сказать, что цели, описанные в постановке задачи, были достигнуты.

# Заключение

В данной работе была проведено исследование клинической картины такого серьезного заболевания, как инфаркт миокарда. Данное исследование позволило выявить у мужчин молодого и среднего возраста те количественные факторы, которые оказывают влияние на одни из самых важных характеристик данной болезни: смертность и наличие острых осложнений. Ведь именно люди этого возраста наиболее «беззащитны» перед этой болезнью, именно в этом возрастном сегменте возникает злокачественное течение болезни и высокий процент смертельных исходов. Кроме этого, было проведен анализ течения болезни с учетом заболевания легких пациента, что позволит лучше изучить связь инфаркта миокарда с легочными заболеваниями.

Проделанная работа должна помочь постановке правильного диагноза, вовремя предоставленного качественного лечения пациентам с инфарктом миокарда. Учитывая серьезность заболеваний сердца (в частности ИМ) и неутешительную статистику смертности в результате таких болезней, быстрое и правильное диагностирование и лечение просто необходимо.

Проведенная работа и полученные выводы могут найти продолжение в виде исследования также качественных и порядковых факторов, влияющих на течение заболевания; в виде сравнения полученных в данной работе результатов с результатами, полученными в других статистических пакетах; в медицинских системах автоматического принятия решений и т.д. Ведь чем больше и шире охватить доступные инструменты анализа, тем точнее и качественнее возможно произвести диагностику и лечение сердечных заболеваний, а значит увеличить вероятность сохранения здоровья и жизни пациентов.

# Список литературы


1. Оганов Р.Г., Масленникова Г.Я. Профилактика сердечно-сосудистых заболеваний реальный путь улучшения демографической ситуации в России // Кардиология. – 2007. – Т. 47,№1. – С.4-7.
2. Оганов Р.Г., Масленникова Г.Я. Смертность от сердечно-сосудистых и других неинфекционных заболеваний среди трудоспособного населения России // Кардиоваскуляр. терапия и профилактика. – 2002. – Т. 1, №3. – С. 4-8.
3. Яковлев В.А., Чепель А.И. Ишемическая болезнь сердца: учеб. пособие для слушателей I, VI факультетов и клинич. ординаторов: Ч. 1. СПб.: ВМедА., 2003. – 52 с.
4. Меньшикова И.Г., Лоскутова Н.В., Афонькин А.Н. и др. Факторы риска и особенности лечения инфаркта миокарда у лиц молодого возраста // Актуальные проблемы кардиологии в Приамурье: тез. докл. науч.-практ.конф. – Благовещенск, 1997. – С. 23-25.
5. Зяблов Ю.И., Округин С.А., Орлова С.Д. Острые коронарные катастрофы у лиц до 40 лет: результаты 10-летнего наблюдения в Томске (1988-1997) по программе ВОЗ «Регистр острого инфаркта миокарда» // Кардиология. – 1999. – Т. 39, №11. – С. 47-50.
6. Уускюла М.М., Ламп К.М. Ноозла С.А. Изучение причин заболеваемости острым инфарктом миокарда в молодом возрасте // Многофакторная профилактика ИБС: тез. докл. Всесоюз. симпоз. – Томск, 1989. – С. 150.
7. Bosetti C., Negri E., Tavani A. et al. Smoking and acute myocardial infarction among women and men: a case–control study in Italy // Prev. Med. – 1999. – Vol. 29, №5. – P.343–348.
8. Бойцов С.А. Методологические основы Российского многоцентрового эпидемиологического исследования острой ИБС (Резонанс) /С.А. Бойцов,

С.С. Якушин, Р.А. Лиферов и др. //Материалы III Национального 117 конгресса терапевтов «Новый курс: консолидация усилий по охране здоровья нации» (г. Москва, 5-7 октября 2008 г.).- 2008.- С.25-26.

9. Беленков Ю. Н., Привалова Е. В., Каплунова В. Ю., Хмелькова Е. В., Чекнева Н. С., Черноусов А. Ф., Хоробрых Т. В., Ветшев Ф. П. Роль экстракардиальных факторов в течении ишемической болезни сердца, нарушений ритма и проводимости сердца // Кардиология и сердечно-сосудистая хирургия. - 2009. – № 4. – С. 8-17.

10. Yusuf S., Hawken S., Ounpuu S. et al. On behalf of the INTERHEART Study Investigators. Effect of potentially moldable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study // Lancet. – 2004. – Vol. 364, №9438. – P. 937–952.

11. Чазов Е.И. Проблемы первичной и вторичной профилактики сердечно-сосудистых заболеваний // Терапевт. арх. – 2002. – Т. 74, №9. – С. 5-8.

12. Скрибник Э.Я. Редкие варианты инфаркта миокарда у больных молодого возраста // Клинич. медицина. – 1991. – Т. 69, №11. – С. 32-35.

13. Cole J. H., Miller J. I., Sperling L. S., Weintraub W.S. Long-term follow-up of coronary artery disease presenting in young adults // J. Am. Coll. Cardiol. – 2003. – Vol. 41, №4. – P.521-528.

14. Панова Т.Н., Копылова Н.А. Особенности лечения инфаркта миокарда в молодом возрасте // Вопросы диагностики и лечения внутренних и инфекционных болезней: (по материалам 77 науч.-практ. конф. сотрудников АГМА). – Астрахань, 2000. – Т. 1. – С. 26-30.

15. Гонохова Л.Г., Быканова Л.В., Кутенких Е.В. и др. Структура факторов риска сердечно-сосудистых заболеваний у мужчин трудоспособного возраста // Материалы 9-го Всерос. науч.-образоват. форума «Кардиология 2007». – М.: Б.и., 2007. – С. 57-59.

16. D'Agostino R. B., Grundy S., Sullivan L.M., Wilson P. Validation of the Framingham coronary heart disease predictions scores: results of multiple ethnic group investigations // JAMA. – 2001. – Vol. 286, №2. – P. 180-187.

17. Vaccarino V., Parsons I., Peterson E. D. Sex differences in mortality after acute myocardial infarction: changes from 1994 to 2006 // Arch. Intern. Med. – 2009. – Vol. 169, № 19. – P. 1767-1774.

18. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей - М.: Финансы и статистика, 1985. 488 с.

19. Дж.-О. Ким, Ч. У. Мюллер, У. Р. Клекка. «Факторный, дискриминантный и кластерный анализ», Издательство: Финансы и статистика, 1989. 216 с.

20. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. – СПб.: ВМедА, 2002. 266 с.

21. Гланц С. Медико-биологическая статистика – М.: Практика, 1998.459 с.

22. База примеров статистических решений StatSoft. URL: http://www.statsoft.ru/solutions/

23. Болч Б., Хуань К.Дж. Многомерные статистические методы для экономики – М.: Статистика, 1979. 317 с.

24. Ллойд Э., Ледерман У. Справочник по прикладной статистике. Том 1. М.: Финансы и статистика, 1989. 510 с.

25. Афифи А., Эйзенс С. Статистический анализ. Подход с использованием ЭВМ – М.: Мир, 1982. 488 с.

26. Кобзарь А.И. «Прикладная математическая статистика. Для инженеров и научных работников», М.: Физматлит, 2006. 816 с.

27. Shapiro S., Francia R. S. An approximate analysis of variance test normality // JASA. 1972. V. 67, №337. P. 215-216.

28. Levene H. Robust Tests for the Equality of Variance // Contributions to Probability and Statistics, ed. I. Olkin, Palo Alto, CA: Stanford University Press, 1960. P. 278 -292.

29. Brown M.B., Forsythe A.B. Robust Tests for Equality of Variances // Journal of the American Statistical Association, 69, 1974. 364 -367.

30. Закс Л. Статистическое оценивание. М.: Статистика, 1976. – 598 с.

31. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. 607 с.

# Приложение А
## Программа для прогнозирования осложнений на основе дискриминантного анализа

По результатам проведения дискриминантного анализа – нахождения наиболее влияющих факторов и построения классифицирующих функций для параметра № 20 «Осложнения» было реализовано приложение для прогнозирования осложнений у поступивших пациентов. В основу прогноза приложения положены классифицирующие функции, полученные в параграфе 2.3 для каждой из двух групп (с заболеванием легких и без него). Программа принимает на вход значения влияющих факторов, соответствующих группе пациента, а на выход выдает прогноз по развитию у данного пациента осложнений инфаркта миокарда.

Программа реализована в программной среде Visual Studio 2015 на языке *C#* и имеет пользовательский интерфейс, поэтому подойдет для использования врачом-специалистом.

Скриншоты вида и работы программы представлены на рисунках 1,2,3:

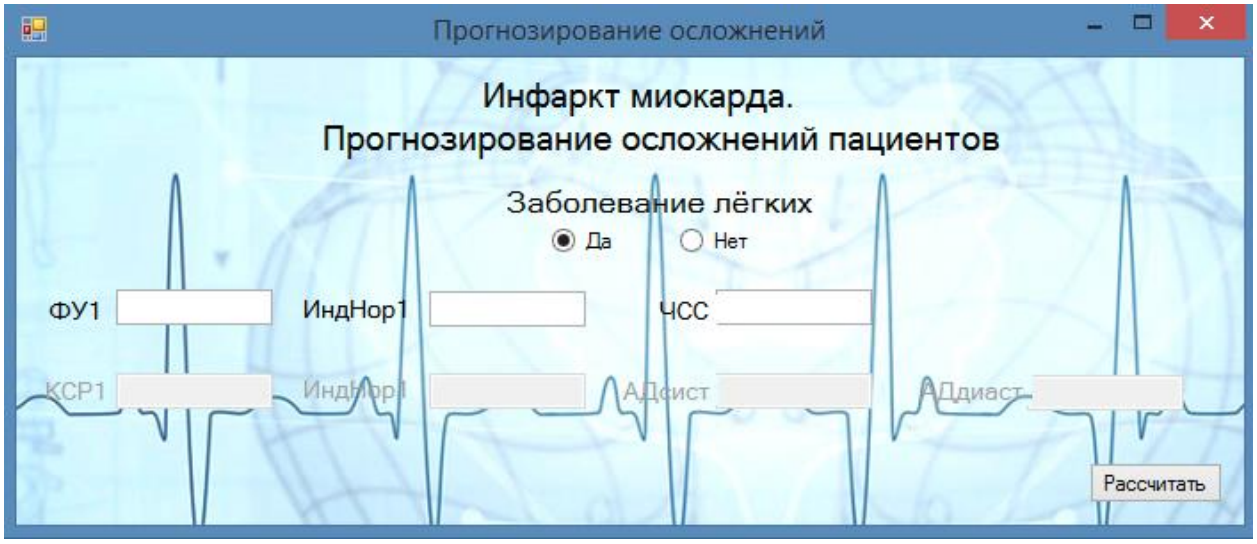


Рис.1. Общий вид программы

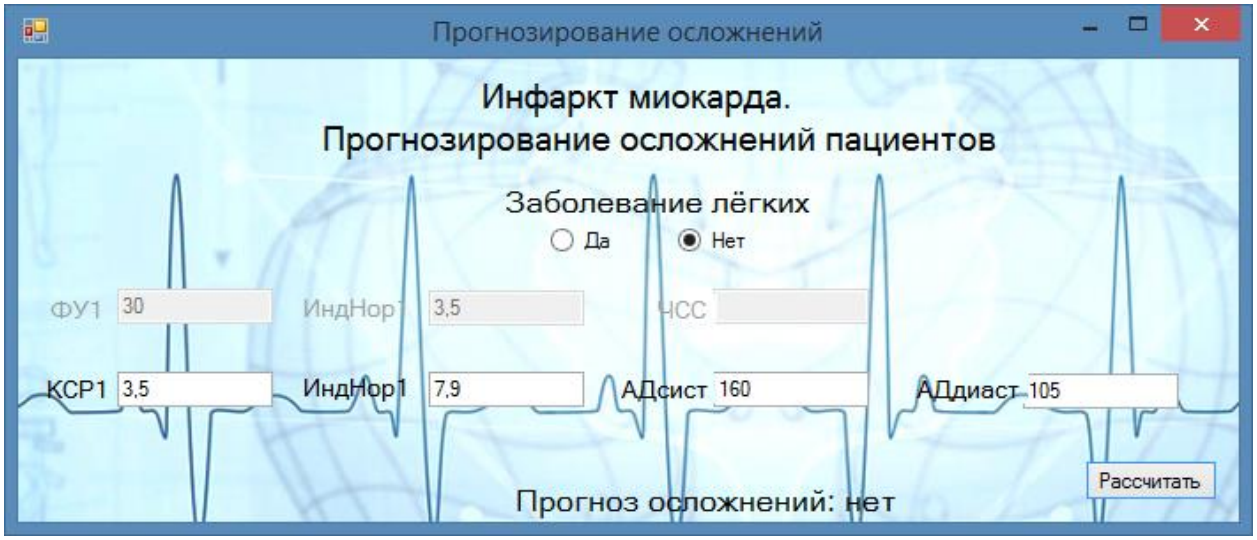


Рис.2. Пациент без заболевания легких. Прогноз: без осложнений


Рис.3. Пациент с заболеванием легких. Прогноз: осложненный ИМ