ФМБА РОССИИ
Федеральное медико-биологическое агентство

Медико-биологический университет
инноваций и непрерывного образования
ФМБЦ им. А.И. Бурназяна ФМБА России

Адрес: г. Москва, ул. Живописная, д. 46, стр. 8
Тел.: 8 (499) 190-96-92
Сайт: www.mbufmbc.ru

Бушманов А.Ю., Зубов А.С., Тихонова О.А., Касымова О.А., Кретов А.С., Галстян И.А.

ОБЩИЕ ТРЕБОВАНИЯ К ПРИМЕНЕНИЮ ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Учебно-методическое пособие для аспирантов и клинических ординаторов

Москва, 2025

Федеральное медико-биологическое агентство
Федеральное государственное бюджетное учреждение
«Государственный научный центр Российской Федерации —
Федеральный медицинский биофизический центр
имени А.И. Бурназяна»
МЕДИКО-БИОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ
ИННОВАЦИЙ И НЕПРЕРЫВНОГО ОБРАЗОВАНИЯ

**Бушманов А.Ю., Зубов А.С., Тихонова О.А.,
Касымова О.А., Кретов А.С., Галстян И.А.**

ОБЩИЕ ТРЕБОВАНИЯ К ПРИМЕНЕНИЮ ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

**Учебно-методическое пособие
для аспирантов и клинических ординаторов**

Москва 2025

УДК 51-7
ББК 22.16
С23

Бушманов А.Ю., Зубов А.С., Тихонова О.А., Касымова О.А., Кретов А.С., Галстян И.А. Общие требования к применению дискриминантного анализа для интеллектуальной поддержки принятия решений. Учебно-методическое пособие для аспирантов и клинических ординаторов. — М.: ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, 2025. — 36 с.

Авторы:

Бушманов А.Ю. — заместитель генерального директора по науке — начальник управления радиационной медицины ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, доктор медицинских наук, профессор, заведующий кафедрой медицины труда, гигиены и профпатологии МБУ ИНО ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России

Зубов А.С. — старший научный сотрудник лаборатории мультидисциплинарных клинических исследований, кандидат экономических наук, доцент

Тихонова О.А. — заведующая лаборатории мультидисциплинарных клинических исследований, кандидат медицинских наук

Касымова О.А. — заместитель главного врача по терапевтической и амбулаторно-поликлинической помощи ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, доцент кафедры терапии МБУ ИНО ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, к.м.н.

Кретов А.С. — руководитель центра профпатологии ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, старший преподаватель кафедры медицины труда, гигиены и профпатологии МБУ ИНО ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России

Галстян И.А. — заведующая лабораторией местных лучевых поражений и последствий лучевой болезни, профессора кафедры медицины труда, гигиены и профпатологии МБУ ИНО ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России

Рецензенты:

Шулаев А.В. — заведующий кафедрой общей гигиены ФГБОУ ВО Казанский ГМУ Минздрава России, д.м.н., профессор

Лавер Б.И. — заведующий отделом по координации деятельности медицинских организаций Центрального федерального округа ФМБА России, к.м.н., врач высшей квалификационной категории

Учебное издание предназначено для подготовки и самостоятельной работы аспирантов и ординаторов Медико-биологического университета инноваций и непрерывного образования ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России.

Содержание данного пособия направлено на внедрение в учебно-методическую работу единого подхода к оценке знаний обучающихся. Данное учебное пособие способствует стандартизации учебного процесса, обеспечивает доступность для обучающихся актуальной методической информации и инструментов для отработки практических навыков.

Издание освещает изучаемые на практических занятиях вопросы и содержит справочную информацию, вопросы для самоконтроля.

ISBN 978-5-93064-401-2

Содержание

ВВЕДЕНИЕ

Методические рекомендации устанавливают основные требования к алгоритму анализа первичных данных для классификации или группировки объектов по пороговому значению (константе дискриминации) с помощью модели дискриминантной функции. На реальных клинических данных представлен пример прогнозирования точности системы переменных-предикторов и обоснована целесообразность использования дискриминантного анализа в качестве статистического инструмента для мониторинга здоровья и экспертизы трудоспособности работников радиационно-опасных производств. Рассмотрены возможности использования дискриминантного анализа в профилактической медицине, необходимого для моделирования ситуаций при классификации групп лиц, работающих во вредных условиях труда. Описаны исчисления для интеллектуальной поддержки принятия решений [1] при комплексной оценке состояния здоровья работников атомной отрасли.

Предлагаемая дискриминантная математическая модель с расчетом линейных дискриминантных функций позволяет с большей вероятностью, чем при традиционном дифференциальном диагнозе, на основании отдельных результатов медицинского обследования и симптомов заболевания утверждать, что у пациента развиваются предпосылки к потере трудоспособности. Приведены примеры использования дискриминантной модели для дифференциальной диагностики у пациентов с противоречивыми данными различных методов обследования [2-6].

1. Область применения

1.1. Методические рекомендации распространяются на проблему продления трудового долголетия работников объектов использования атомной энергии и/или мониторинга здоровья, экспертизы трудоспособности работников радиационно-опасных производств

1.2. Документ устанавливает основные требования к алгоритму анализа первичных данных для классификации или группировки объектов по пороговому значению (константе дискриминации) с помощью модели дискриминантной функции.

1.3. Методические рекомендации предназначены для членов комиссий профцентров; врачей, принимающих решения о профилактическом лечении; в научно-исследовательской работе, при составлении учебного плана подготовки кадров высшей квалификации (аспирантура) по направлению «Профилактическая медицина».

2. Нормативные ссылки

В настоящих Методических рекомендациях учтены основные положения следующих нормативных документов:

Приказ Минздрава России от 28.07.2020 № 749н «Об утверждении требований к проведению медицинских осмотров и психофизиологических обследований работников объектов использования атомной энергии, порядка их проведения, перечня медицинских противопоказаний для выдачи разрешения на выполнение определенных видов деятельности в области использования атомной энергии и перечня должностей работников объектов использования атомной энергии, на которые распространяются данные противопоказания, а также формы медицинского заключения о наличии (отсутствии) медицинских противопоказаний для выдачи разрешения на выполнение определенных видов деятельности в области использования атомной энергии» (Зарегистрировано в Минюсте России 11.09.2020 № 59782);

Приказ Минздрава России от 28.01.2021 № 29н (ред. от 01.02.2022) «Об утверждении Порядка проведения обязательных предварительных и периодических медицинских осмотров работников, предусмотренных частью четвертой статьи 213 Трудового кодекса Российской Федерации, перечня медицинских

противопоказаний к осуществлению работ с вредными и (или) опасными производственными факторами, а также работам, при выполнении которых проводятся обязательные предварительные и периодические медицинские осмотры» (Зарегистрировано в Минюсте России 29.01.2021 № 62277);

Приказ Минздрава России от 05.05.2016 № 282н «Об утверждении порядка проведения экспертизы профессиональной пригодности и формы медицинского заключения о пригодности или непригодности к выполнению отдельных видов работ» (зарегистрирован Минюстом России от 02.06.2016, регистрационный номер № 42397);

Приказ Министерства здравоохранения Российской Федерации от 24.12.2018 № 911н «Об утверждении Требований к государственным информационным системам в сфере здравоохранения субъектов Российской Федерации, медицинским информационным системам медицинских организаций и информационным системам фармацевтических организаций» (Зарегистрирован 19.06.2019 № 54963);

Р ФМБА России 1-2023. Рекомендации. Порядок разработки, изложения, представления на согласование и утверждение нормативных и методических документов, разрабатываемых научными организациями по заказу ФМБА России, в Комиссию Федерального медико-биологического агентства по рассмотрению нормативных и методических документов, разработанных при выполнении научно-исследовательских и опытно-конструкторских работ, осуществлении научно-технической и инновационной деятельности.

Примечание: При пользовании настоящим документом целесообразно проверить действие ссылочных нормативных документов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет и по соответствующим ежемесячно издаваемым информационным указателям, опубликованным в текущем году

3. Термины и определения

В настоящем документе применяются следующие термины с соответствующими определениями:

анализ дискриминантной функции: Это статистический анализ, используемый для определения точности данной системы классификации или предикторных переменных;

валидация: Это процесс подтверждения того, что стандартизованная методика соответствует установленным требованиям в конкретных условиях, что она дает достоверные результаты;

выброс: Это точка данных, которая существенно отличается от других наблюдений;

двоичные переменные: Это переменные с двумя возможными значениями (например, да/нет, true/false);

дискриминантный анализ (от латинского discriminatio — различение): Это статистический метод, используемый в исследованиях, целью которого является классификация или прогнозирование категориальной зависимой переменной на основе одной или нескольких непрерывных или бинарных независимых переменных. Дискриминантный анализ позволяет классифицировать объекты по пороговому значению (константой дискриминации);

диоптрия: Это единица измерения оптической силы линз и других осесимметричных оптических систем, используется для характеристики остроты зрения;

индекс массы тела (ИМТ): Это величина, позволяющая оценить степень соответствия массы человека и его роста и тем самым косвенно судить о том, является ли масса недостаточной, нормальной или избыточной. Формула расчёта ИМТ: ИМТ = вес (кг): рост^2 (м);

категориальные переменные: Это переменные с различными категориями или группами (например, пол, тип лечения);

константа дискриминации: Это величина, которая определяет границу, разделяющую две рассматриваемые группы в рамках дискриминантного анализа;

лактатдегидрогеназа (ЛДГ): Это фермент, принимающий участие в реакциях гликолиза. Лактатдегидрогеназа катализирует превращение лактата в пируват (соль пировиноградной кислоты);

мониторинг (англ. monitoring — слежение, контроль): Это специальная форма наблюдения (слежения) за текущим изменением тех или иных процессов или объектов в пространстве и во времени, осуществляемая на постоянной основе;

номинальные переменные: Это категориальные переменные без какого-либо внутреннего порядка (например, виды фруктов: яблоко, банан, вишня);

показатель активности регуляторных систем (ПАРС, баллы) в психофизиологии: Это диагностический критерий для выявления нарушений адаптации и оценки адаптационных возможностей организма. ПАРС 1–2 балла — состояние оптимального напряжения регуляторных систем (норма). Увеличение баллов ПАРС свидетельствует о состоянии от умеренного (3-4)/выраженного (4-6)/ перенапряжения (6-8) регуляторных систем до избыточной активации (8-10) или состояние истощения (астенизации) регуляторных систем;

порядковые переменные: Это категориальные переменные со значимым порядком или ранжированием (например, уровни удовлетворённости: низкий, средний, высокий);

разведочные исследования: Это исследования, которые проводятся на начальном этапе, когда необходимо собрать максимально полные сведения о сущности изучаемой проблемы и ее структуре;

система переменных-предикторов: Это числовые переменные, которые могут принимать любое значение в пределах диапазона (например, возраст, температура).

4. Сокращения и обозначения

В настоящем документе применяются следующие сокращения:

CDA – canonical discriminant analysis (канонический дискриминантный анализ);

DA – discriminant analysis (дискриминантный анализ);

FDA – flexible discriminant analysis (гибкий дискриминантный анализ);

LDA – linear discriminant analysis (линейный дискриминантный анализ);

MDA – multiple discriminant analysis (множественный дискриминантный анализ);

QDA – quadratic discriminant analysis (квадратичный дискриминантный анализ);

RDA – regularized discriminant analysis (регуляризованный дискриминантный анализ);

ЛДГ – лактатдегидрогеназа;

ПАРС – показатель активности регуляторных систем.

5. Общие положения

5.1. Теоретические и методологические основы дискриминантного анализа

Дискриминантный анализ (Discriminant analysis, DA) применяется для определения точности данной системы классификации при прогнозировании принадлежности субъекта к определенной группе. DA включает разработку дискриминантных функций для каждого образца и выведение порогового значения, которое используется для классификации образцов по разным группам. Анализ дискриминантной функции — это статистический анализ, используемый для определения точности данной системы классификации или предикторных переменных.

DA рекомендуется применять, когда зависимая переменная является неметрической (категориальной), а независимые переменные являются метрическими (непрерывными или бинарными).

В данном документе приводятся основные положения (этапы) дискриминантного анализа с рассмотрением его применения на реальном клиническом примере в Приложении А.

При введении новой системы классификации анализ дискриминантной функции может использоваться для определения точности, с которой классификация способна дифференцировать конкретного субъекта по разным группам.

Необходимость в классификации наблюдений возникает, когда исследователь пытается классифицировать ряд лиц в две или более категорий или пытается решить, в какой категории следует оставить этих лиц в зависимости от количества измерений, доступных для каждого из этих лиц. Прямая идентификация этих лиц с их соответствующими категориями невозможна, и, следовательно, формулируется вопрос о построении подходящей «статистической функции принятия решения», предполагающей проведение классификации данных лиц.

5.2. Семантика дискриминантного анализа

Дискриминантный анализ — это параметрический метод определения того, какие веса количественных переменных или предикторов наилучшим образом различают две или более категорий зависимых переменных и делают это лучше, чем случайный выбор.

DA является наиболее популярным статистическим методом классификации отдельных лиц или наблюдений в непересекающиеся группы на основе оценок, полученных из подходящей «статистической функции принятия решений», построенной из одной или нескольких непрерывных предикторных переменных.

Например, если врач желает выявить пациентов с высоким, умеренным и низким риском развития сердечных осложнений, таких как инсульт, то он может выполнить дискриминантный анализ для классификации пациентов с различными группами риска инсульта. Такой метод позволяет врачу классифицировать пациентов на группы высокого-умеренного-низкого риска на основе личных характеристик (например, уровень липопротеинов высокой плотности [ЛПВП], уровень липопротеинов низкой плотности [ЛПНП], уровень холестерина, индекс массы тела) и/или образа жизни (например, количество часов упражнений или физической активности в неделю, статус курения, такой как количество сигарет в день).

DA выполняет ту же задачу, что и множественная линейная регрессия, прогнозируя результат на основе заданного набора предикторов. Однако множественная линейная регрессия ограничена случаями, когда зависимая переменная на оси Y является интервальной переменной, так что комбинация предикторов будет, через уравнение регрессии, производить оценочные средние числовые значения Y для популяции при заданных значениях взвешенных комбинаций X. Проблема возникает, когда зависимая переменная имеет категориальную природу, например, статус/стадии конкретного заболевания или статус мигранта/немигранта. Вместе с этим желательно также минимизировать вероятность неправильной классификации.

5.3. Цели дискриминантного анализа

При исследовании различий между группами или категориями необходимым шагом является определение атрибутов с наибольшим вкладом в максимальную разделимость между известными группами или категориями, чтобы классифицировать данное наблюдение в одну из групп. Для этой цели DA последовательно определяет линейную комбинацию атрибутов, известных как канонические дискриминантные функции (уравнения), которые вносят максимальный вклад в разделение групп. Предиктивный DA решает вопрос о том, как распределять новые

случаи по группам. Функция DA выдает оценки для отдельных лиц по предикторным переменным для прогнозирования категории, к которой принадлежит данное лицо. DA определяет наиболее экономичный способ различения групп. Статистические тесты значимости с использованием хи-квадрат позволяют исследователю увидеть, насколько хорошо функция разделяет группы. Также DA позволяет исследователю проверить, классифицируются ли случаи как предсказанные.

5.4. Ключевые предположения дискриминантного анализа

Для DA необходимы следующие предположения:

a) наблюдения представляют собой случайную выборку из разных популяций, характеризующихся разными распределениями вероятностей.
b) каждая предикторная переменная распределена нормально.
c) каждое из распределений для зависимых категорий в исходной классификации правильно классифицировано, и группы или категории должны быть определены до сбора данных.
d) должно быть не менее двух групп или категорий, причем каждый случай принадлежит только одной группе, чтобы группы были взаимоисключающими и в совокупности исчерпывающими (все случаи можно поместить в группу).
e) атрибуты, используемые для разделения групп, должны достаточно четко различать группы, чтобы перекрытие групп или категорий было явно несущественным или минимальным.
f) размеры групп зависимых переменных не должны сильно различаться и должны быть как минимум в пять раз больше числа независимых переменных.

В зависимости от числа категорий и метода построения дискриминантной функции существует несколько типов DA, таких как линейный, множественный и другие виды DA. В данном документе приводятся алгоритмы расчетов с применением линейного DA.

5.5. Типы дискриминантного анализа

Выделяют следующие типы дискриминантного анализа:

– линейный дискриминантный анализ (LDA);

– квадратичный дискриминантный анализ (QDA);
– регуляризованный дискриминантный анализ (RDA);
– гибкий дискриминантный анализ (FDA);
– множественный дискриминантный анализ (MDA);
– канонический дискриминантный анализ (CDA).

5.5.1. Линейный дискриминантный анализ (LDA)

Этот тип дискриминантного анализа используется, когда все предикторные переменные непрерывны и нормально распределены, а группы имеют равные ковариационные матрицы. LDA стремится найти линейную комбинацию предикторов, которая максимально разделяет группы.

5.5.2. Квадратичный дискриминантный анализ (QDA)

KDA похож на LDA, но он не предполагает, что группы имеют равные ковариационные матрицы. Это означает, что он может моделировать более сложные границы групп, но он также требует оценки большего количества параметров, чем LDA, и может быть более склонен к переобучению.

5.5.3. Регуляризованный дискриминантный анализ (RDA)

Регуляризованный дискриминантный анализ (RDA) — это расширение квадратичного дискриминантного анализа (QDA) и линейного дискриминантного анализа (LDA).

Этот вид анализа представляет собой комбинацию между LDA и QDA, который позволяет моделировать более сложные границы групп, чем LDA, но менее склонен к переобучению, чем QDA. Он делает это путем «сжатия» специфичных для группы ковариационных матриц до общей ковариационной матрицы со степенью сжатия, определяемой параметром настройки.

Цель RDA — улучшить точность классификации за счёт введения параметра регуляризации в матрицы ковариации. Этот параметр помогает стабилизировать оценки матриц ковариации, особенно в ситуациях, когда количество признаков близко или превышает количество наблюдений, или когда данные сильно коллинеарны.

RDA используется в ситуациях, когда в наборе данных большое количество признаков, присутствует мультиколлинеарность или когда в наборе данных есть выбросы, которые могут исказить матрицу ковариации.

5.5.4. Гибкий дискриминантный анализ (FDA)

Это расширение LDA, которое использует методы расширения базиса для моделирования нелинейных границ между группами. По сути, он применяет LDA в преобразованном пространстве предикторов.

Гибкий дискриминантный анализ (FDA) — это общая методология, направленная на предоставление инструментов для многогрупповой нелинейной классификации. Это модель классификации, основанная на комбинации непараметрических регрессионных моделей, например, MARS и линейного дискриминантного анализа.

Первым шагом FDA является непараметрическая регрессия, которая использует оптимальную оценку для преобразования переменной отклика, чтобы данные были в лучшей форме для линейного разделения. В FDA строятся множественные регрессионные модели, так называемые базисные функции, по всему диапазону значений предикторов. В этой процедуре диапазон значений предикторов разбивается на несколько групп/категорий.

5.5.5. Множественный дискриминантный анализ (MDA)

Он используется, когда требуется распределить данные на более двух групп, и вы хотите смоделировать вероятность членства в группе как функцию предикторов. Множественный дискриминантный анализ (MDA) — это многомерный метод уменьшения размерности. Он использовался для прогнозирования таких разнообразных сигналов, как нейронные следы памяти и корпоративные сбои.

5.5.6. Канонический дискриминантный анализ (CDA)

Этот тип дискриминантного анализа используется для выявления и измерения связей между набором переменных и между этим набором переменных и набором фиктивных переменных, которые представляют принадлежность к группам.

Канонический дискриминантный анализ — это метод, при котором классификационные функции вычисляются методом канонической корреляции.

Количество функций определяется количеством значимых корней канонической корреляции. Как правило, их много меньше, чем исходных признаков.

Корни задают новое пространство, в котором определяется «центр тяжести» каждой группировки — центроид. Объект, определённый в этом пространстве как точка, относится к той группировке, к центроиду, которой он расположен ближе.

Поскольку канонических корней, как правило, не более трёх, результат анализа допускает простую и очевидную графическую интерпретацию. Это большое преимущество канонического дискриминантного анализа перед линейным.

Канонический дискриминантный анализ — это метод редукции размерности, связанный с анализом главных компонентов и канонической корреляцией. При наличии переменной классификации и нескольких интервальных переменных канонический дискриминантный анализ выводит канонические переменные (линейные комбинации интервальных переменных), которые суммируют межклассовую вариацию примерно так же, как главные компоненты суммируют общую вариацию.

5.6. Области применения дискриминантного анализа

Дискриминантный анализ применяется для решения следующих задач:

Мультиклассовая классификация

DA часто используется, когда зависимая переменная является категориальной и имеет более двух категорий. В то время как другие методы, такие как логистическая регрессия, могут обрабатывать бинарные результаты, дискриминантный анализ особенно подходит для задач мультиклассовой классификации.

Прогностическое моделирование

DA применяется для прогнозировании членства в группе на основе набора предикторов. Например, в дерматологии методологию DA применяют для дифференциации клинико-иммунологических особенностей аллергических заболеваний кожи.

Понимание различий между группами

Дискриминантный анализ также можно использовать, когда интересно понять, какие переменные различают две или более естественно возникающих групп. Например, классификации больных инсультом по степени тяжести заболевания может проводиться методом DA.

Выполнение допущений

Дискриминантный анализ предполагает, что предикторы распределены нормально и что группы имеют равные ковариационные матрицы. Если ваши данные соответствуют этим допущениям, дискриминантный анализ может быть особенно эффективным методом.

Снижение размерности

Линейный дискриминантный анализ (LDA) также может использоваться для снижения размерности. То есть его можно использовать для сокращения количества переменных в наборе данных, сохраняя при этом как можно больше информации.

5.7. Преимущества и недостатки дискриминантного анализа

5.7.1. Преимущества дискриминантного анализа

Дискриминантный анализ имеет несколько преимуществ, которые делают его ценным инструментом в статистическом инструментарии исследователя:

Мультиклассовая классификация: дискриминантный анализ может обрабатывать ситуации, когда в зависимой переменной более двух классов, что является ограничением для некоторых других методов, таких как логистическая регрессия.

Понимание различий в группах: дискриминантный анализ не просто предсказывает принадлежность к группе; он также предоставляет информацию о том, какие переменные являются важными дискриминаторами между группами. Это делает его полезным инструментом для разведочных исследований, чтобы понять различия между группами.

Работа с большим количеством переменных

DA применяется для обработки большого количества переменных-предикторов. Он становится полезным, когда количество переменных очень велико, потенциально превышая количество наблюдений.

Снижение размерности: линейный дискриминантный анализ (LDA) можно использовать для снижения размерности — он может уменьшить количество переменных в наборе данных, сохраняя при этом как можно больше информации.

Априорные вероятности: дискриминантный анализ позволяет включать априорные вероятности, что означает, что исследователи могут включать априорные знания о пропорциях наблюдений в каждой группе.

Интерпретируемость модели: модель, созданная дискриминантным анализом, относительно интерпретируема по сравнению с некоторыми другими моделями машинного обучения, такими как нейронные сети. Веса признаков в модели могут дать представление об их относительной важности.

5.7.2. Недостатки дискриминантного анализа

Перечислим основные недостатки DA.

Предположение о нормальности: дискриминантный анализ предполагает, что предикторы распределены нормально. Если это предположение нарушается, это может повлиять на производительность модели.

Предположение о равных ковариационных матрицах: дискриминантный анализ, в частности линейный дискриминантный анализ (LDA), предполагает, что сравниваемые группы имеют равные ковариационные матрицы. Если это предположение не выполняется, это может привести к неточностям в классификации.

Мультиколлинеарность: Дискриминантный анализ может работать неэффективно, если среди предикторных переменных наблюдается высокая мультиколлинеарность. Такая ситуация может привести к нестабильным оценкам коэффициентов и трудностям в интерпретации результатов.

Выбросы: Выбросом называется точка данных, которая значительно отличается от других наблюдений. Дискриминантный анализ чувствителен к выбросам, которые могут оказывать большое влияние на функцию классификации.

Переобучение: Как и многие статистические методы, дискриминантный анализ может привести к переобучению, если модель слишком сложна. Переобучение происходит, когда модель очень хорошо соответствует обучающим данным, но плохо работает на новых, неизвестных данных.

Ограничено линейными отношениями: Линейный дискриминантный анализ (LDA) предполагает линейную связь между предикторными переменными и логарифмическими коэффициентами зависимой переменной. Это ограничивает его

полезность в сценариях, где отношения являются сложными или нелинейными. В таких случаях квадратичный дискриминантный анализ (QDA) или другие нелинейные методы могут быть более подходящими.

5.8. Методология дискриминантного анализа. Основные требования к алгоритму проведения дискриминантного анализа

Методология дискриминантного анализа заключается в поиске новых признаков, называемых дискриминантными функциями, на основе использования совокупности исходных показателей.

Методология дискриминантного анализа включает несколько основных этапов, которые могут варьироваться в зависимости от особенностей проводимого анализа. Выделим наиболее значимые:

- постановка задачи для проведения дискриминантного анализа. Определение зависимой переменной (изучаемое явление) и независимых переменных (данные на основании, которых производится классификация изучаемого явления)
- сбор данных и их проверка на предмет возможности математической и статистической обработки для проведения расчетов, при необходимости внесение дополнений и изменений имеющиеся данные
- определение нормальности распределения данных для анализа, при необходимости проведение нормализации данных
- проведение расчетов для определения критериев/коэффициентов дискриминации
- валидация — модели или методики путем проведения сравнения с экспериментальными результатами.

6. Алгоритм выполнения дискриминантного анализа

6.1. Первый этап. Постановка задачи для проведения дискриминантного анализа. Сбор первичных данных

1) четко определить проблему и цели анализа. Необходимо определить зависимую (объясняемую) переменную, которая представляет собой изучаемое явление, а также независимые переменные, т.е. те данные на основании, которых производится классификация изучаемого явления.

В DA в качестве зависимой переменной могут выступать пациенты (работники), которых необходимо классифицировать, разбить на группы и для этого определить критерии отнесения на основании имеющихся данных (анализы, тесты и т.д.).

2) собрать информацию о независимых переменных (предикторах). Зависимая переменная должна быть категориальной, а независимые переменные чаще всего бывают непрерывными и дискретными.

В статистике категориальная переменная (также называемая качественной переменной) — это переменная, которая может принимать одно из ограниченного и обычно фиксированного числа возможных значений, относя каждую отдельную или иную единицу наблюдения к определенной группе или номинальной категории (классу) на основе некоторого качественного свойства. Часто каждое из возможных значений категориальной переменной называется уровнем (например, степень ожога, группа инвалидности и т.д.).

Категориальные данные могут быть трех типов:

- бинарные или дихотомические переменные — переменные, которые имеют только два варианта: здоров и болен, выиграл и проиграл и т.д.
- номинальные переменные без определенного ранжирования: город, цвет, бренд и т д.
- порядковые переменные (имеют заданный порядок ранжирования), различные рейтинги, образование: начальное, среднее, высшее, оценка знаний: отлично, хорошо, удовлетворительно и т.д.

Сбор данных и их проверка на предмет возможности математической и статистической обработки для проведения расчетов, при необходимости внесение дополнений и изменений имеющиеся данные.

6.2. Второй этап. Предварительная обработка первичных данных

Предварительная обработка данных является неотъемлемым этапом любого анализа данных (будь то сравнение средних значений двух выборок или построение статистической модели), поскольку качество данных и полезная информация, которую можно извлечь из них, влияет на точность получаемых

результатов. Следует предварительно обработать данные, прежде чем приступить к полноценному анализу.

На данном этапе проводится работа с пропущенными значениями, выбросами, а также обеспечение того, чтобы данные соответствовали предположениям дискриминантного анализа. Эти предположения включают независимость наблюдений, нормальное распределение переменных-предикторов в каждой группе зависимой переменной и однородность дисперсий между группами.

Наиболее распространенной проблемой на данном этапе является работа с пропусками в данных или недостающими данными (missing data)

Предварительная обработка данных включает в себя работу с пропущенными значениями, выбросами. Если расчеты осуществляются в MS EXCEL, то необходимо привести все данные к числовому формату.

Пропуски в данных (missing data), часто вызывают ошибки при последующем анализе. Чаще всего, они возникают по причине неправильного ввода данных или при сокрытии информации.

Иногда пропуски в данных возникают по невнимательности респондентов при заполнении анкеты или невнимательности операторов базы данных. Также, достаточно распространены случаи, когда пропуски в данных являются неслучайными. Например, когда исследуют выборку пациентов мужчин и женщин, и гинекологические показатели указывают только для женщин, оставляя пропуски в базе данных для пациентов мужского пола.

Возможно два основных подхода к работе с пропусками данных:

a) заполнение пропусков среднеарифметическими или медианными значениями. Возможны в тех случаях, есть возможность рассчитать (например, частота сердечного ритма может быть рассчитана по предыдущему и последующему значению от пропущенной даты);
b) удаление пропусков (или строк), например, когда исследуется группа пациентов и у каждого пациента разовые анализы. Не рекомендуется заполнять пропуски нулевыми значениями, т.к. это может существенно исказить показатели, используемые в DA, например, среднее арифметическое по возрасту.

Дискриминантный анализ чувствителен к выбросам, которые могут оказывать большое влияние на функцию классификации. По отношению к выбросам можно применить те же методы, что и к пропускам в данных.

6.3. Третий этап. Проверка нормальности распределения

Нормальное распределение часто используются для приближенного описания многих случайных явлений, в которых на интересующие нас признаки оказывает воздействие большое количество независимых случайных факторов, среди которых нет резко выделяющихся. Максимально наглядно и определение нормального распределения — это график кривой Гаусса.

Кривая Гаусса (гауссиана) — это график нормального распределения, при котором большинство значений сосредоточено около среднего значения.

Среднее значение μ (мю) и стандартное отклонение σ (сигма) определяют форму кривой нормального распределения.

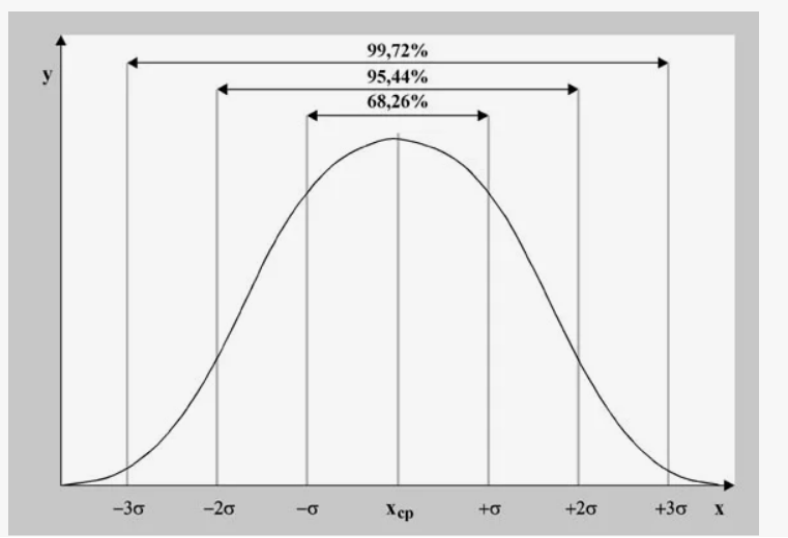


Рис. 1. Классический график двумерного нормального распределения (гауссиана)

По закону нормального распределения 68,26% значений находятся в пределах одной σ от μ, 95,44% — в пределах двух σ, а 99,72% — в пределах трех σ.

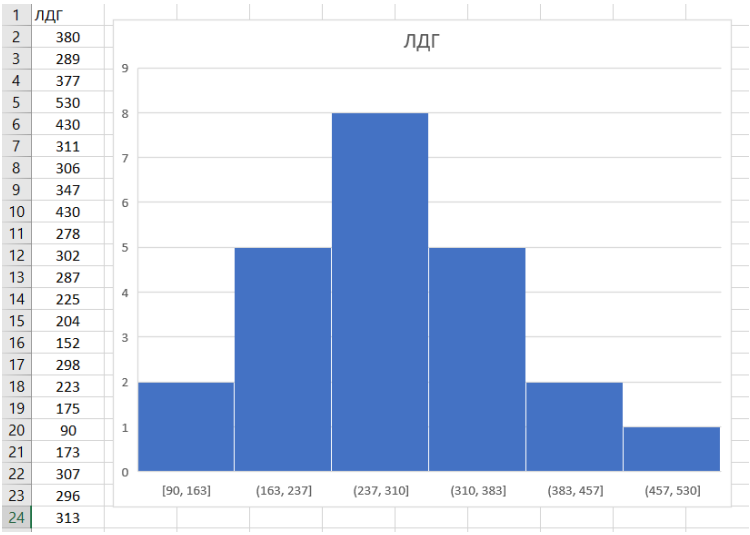
Зная среднее значение и стандартное отклонение распределения, можно устанавливать контрольные границы и принимать решения на их основе.

Существуют различные методы проверки нормальности распределения данных:

- графические;
- косвенные;
- расчетные.

Графические методы

Графический метод анализа предполагает построение соответствующего графика. Например, в MS Excel 2016 можно построить гистограмму частот без предварительной группировки и сортировки данных и провести визуальное сравнение с кривой Гаусса. Приведем пример значений показателей изоферментов ЛДГ (лактатдегидрогеназы) (рис.2).



	ЛДГ
1	
2	380
3	289
4	377
5	530
6	430
7	311
8	306
9	347
10	430
11	278
12	302
13	287
14	225
15	204
16	152
17	298
18	223
19	175
20	90
21	173
22	307
23	296
24	313

Рис. 2. График нормального распределения на примере показателя ЛДГ (расчеты выполнены в MS Excel)

Для построения графика нужно выбрать данные, для которых требуется построить гистограмму (выделить нужные ячейки). В главном меню выбрать пункт «Вставка» в группе «Диаграммы» нажать кнопку «Гистограмма», после чего в выпадающем списке еще раз выбрать «Гистограмма».

По горизонтальной оси гистограммы откладывается диапазон наблюдаемых значений величины, разбитый на определённое число (в данном случае 6) интервалов в порядке возрастания, а по вертикальной — вероятность или частота её попадания в каждый интервал. По графику видно соответствие кривой Гаусса.

Косвенные методы

Предполагают расчет двух показателей:

– коэффициент асимметрии;
– эксцесс распределения

Коэффициент асимметрии

Коэффициент асимметрии A (skewness) — характеризует скошенность распределения в сторону больших или меньших значений признака. Это мера отклонения распределения частоты от симметричного (нормального) распределения, то есть такого, у которого на одинаковом удалении от среднего значения по обе стороны выборки данных располагается одинаковое количество значений.

Формула расчета коэффициента асимметрии:

$$A_s = \frac{\sum_{i=1}^{n}(x_i - x_{\text{ср}})^3}{n\sigma^3} \qquad (2.1)$$

- Если $A_s = 0$, то распределение имеет симметричную форму;
- Если $A_s < 0$, то имеет место левосторонняя асимметрия (длинный хвост с левой стороны);

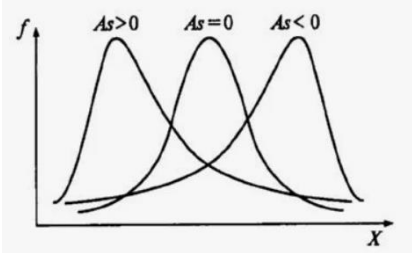


Рис. 3. Асимметрия (симметричность/скошенность) кривой распределения относительно среднего значения

- Если As > 0, то правосторонняя асимметрия (длинный хвост с правой стороны).

Для того чтобы оценить нормальность распределения с точки зрения асимметрии, необходимо сравнить полученный коэффициент асимметрии со средней квадратической ошибкой коэффициента асимметрии, которая рассчитывается по формуле:

$$\sigma_{As} = \sqrt{\frac{6\times(n-1)}{(n+1)\times(n+3)}} \qquad (2.2)$$

Если отношение $|A_s|$:σ_{As}>3, то асимметрия существенная. В противном случае асимметрию можно считать не существенной, обусловленной влиянием случайных факторов.

В программе MS Excel коэффициент асимметрия рассчитывается по каждому ряду путем введения функции СКОС().

Коэффициент эксцесса

Коэффициент эксцесса (Kurtosis) — в статистике мера остроты пика в распределении случайной величины.

Эксцесс является мерой остроты пика в распределении случайных величин. Для любого распределения величин можно рассчитать значение средней величины. В данном контексте коэффициент эксцесса показывает, находится ли большинство значений распределения в непосредственной близости к средней величине, либо же они распределены отдаленно от нее.

Формула расчета коэффициента эксцесса:

$$E_x = \frac{\sum_{i=1}^{n}\left(x_i - x_{\text{ср}}\right)^4}{n\sigma^4} - 3 \qquad (2.3)$$

Эксцесс характеризует распределения, в которых значения величин либо сосредоточены близко к средней величине, либо, наоборот, распределены далеко от нее.

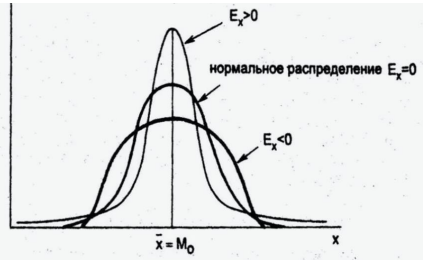


Рис. 4. Эксцесс (компактность или «размытость» распределения значений переменной вокруг среднего значения)

Положительный эксцесс (leptokurtic, Ex > 0) — острая вершина, когда пик выше, чем пик нормального распределения.

Отрицательный эксцесс (platykurtic, Ex < 0) — тупая вершина, когда пик ниже пика нормального распределения.

Для того чтобы оценить нормальность распределения с точки зрения асимметрии, необходимо сравнить полученный коэффициент эксцесса со средней квадратической ошибкой коэффициента асимметрии, которая рассчитывается по формуле:

$$\sigma_{Ex} = \sqrt{\frac{24n\times(n-2)\times(n-3)}{(n-1)^2\times(n+3)\times(n+5)}} \qquad (2.5)$$

Для нормального распределения величин значение эксцесса равно 0 (нулю). Если отношение $|E_x|$:$\sigma_{Ex} > 3$, то отклонение от нормального распределения можно считать существенным.

В программе MS Excel коэффициент эксцесса рассчитывается по каждому ряду путем введения функции ЭКСЦЕСС().

	A	B	C	D	E
1	**ЛДГ (Xi)**			**Расчет асимметрии (Xi - Xср)^3**	**Расчет эксцесса (Xi - Xср)^4**
2	380			674425,8168	59144211,85
3	289	**n (кол-во значений)**	**23**	-36,07923071	119,2183276
4	377	**Xср**	**292,3043**	607551,8527	51457000,4
5	530	**Станд. Отклон. (σ)**	**99,9**	13429619,67	3192162207
6	430			2610722,32	359485112,4
7	311			6534,642886	122169,4105
8	306			2568,905646	35182,8382
9	347			163628,2988	8949756,519
10	430			2610722,32	359485112,4
11	278			-2926,875072	41867,03907
12	302	**Асимметрия (S/(n*σ^3))**	**0,19**	911,4462891	8837,066195
13	287	**Эксцесс (S/(n*σ^4))**	**-0,12**	-149,243692	791,640453
14	225			-304880,2985	20519769,66
15	204			-688567,0906	60803467,87
16	152	***Ошибка (асимметрия)***	***0,46***	-2761934,584	387511430,5
17	298	***Ошибка (эксцесс)***	***0,83***	184,7695406	1052,383035
18	223			-332875,2021	23069698,79
19	175			-1614144,193	189346131,8
20	90			-8279719,987	1675023352
21	173			-1698121,705	202593302,5
22	307			3173,705268	46639,66873
23	296			50,47464453	186,5367298
24	313			8864,155174	183449,4723
25			**Сумма (S)**	**4435603,123**	**6589990849**

Рис. 5. Пример расчета коэффициентов эксцесса и асимметрии распределения в MS Excel

6.4. Четвертый этап. Нормализация данных[1]

Ключевая цель нормализации — приведение различных данных в самых разных единицах измерения и диапазонах значений к единому виду, который позволит сравнивать их между собой или использовать для расчёта схожести объектов.

Выделим три основных метода нормализации:

1) минимально-максимальная нормализация;
2) нормализация по формуле z-преобразования;
3) нормализация на основе среднего значения.

Метод минимально-максимальной нормализации позволяет привести значения данных к диапазону от 0 до 1 путём масштабирования значений в соответствии с минимальным и максимальным значениями.

Формула расчета:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.6)$$

где:

x — исходное значение,
x_{norm} — нормализуемое значение,
x_{max} — максимальное значение в наборе данных,
x_{min} — минимальное значение в наборе данных.

Z-преобразование — это метод, который нормализует данные на основе среднего значения и стандартного отклонения набора данных.

$$x_{norm} = \frac{x - x_{ср}}{\sigma} \quad (2.7)$$

где

x — исходное значение,
x_{norm} — нормализуемое значение,
$x_{ср}$ — среднее значение набора данных,
σ — стандартное отклонение набора данных.

Нормализация на основе среднего значения — это метод, который преобразует данные таким образом, что среднее значение набора данных становится равным 0.

$$x_{norm} = \frac{x - x_{ср}}{x_{max} - x_{min}} \quad (2.8)$$

[1] Практический пример нормализации приводится в Приложении А

x — исходное значение,
x_{norm} — нормализуемое значение,
x_{cp} — среднее значение набора данных,
x_{max} — максимальное значение в наборе данных,
x_{min} — минимальное значение в наборе данных.

6.5. Пятый этап. Проведение расчетов для определения коэффициентов дискриминации

Описание данного этапа предполагает разбиение совокупности на две группы (например, больные и здоровые). Данный этап является наиболее значимым при реализации метода на практике.

Рассмотрим реализацию расчета пошагово (подробный разбор конкретного примера см. в Приложении А):

a) расчет средних значений по каждому признаку (независимой переменной) для всех групп. И, таким образом, получаются векторы средних по матрицам разных групп (например, средний гемоглобин, средний уровень холестерина в крови, средний уровень сахара в крови по группе больных и группе здоровых пациентов);
b) для каждой группы рассчитываются квадратные ковариационные матрицы размерностью n×n (по количеству признаков (результатов тестов)). Для этого рассчитываются центрированные матрицы, где из каждого элемента каждого признака (исходного значения) вычитается среднее значение по данному признаку;
c) рассчитываются квадратные ковариационные матрицы;
d) вычисляется несмещенная оценка суммарной ковариационной матрицы для двух групп;
e) полученная матрица преобразуется в обратную матрицу;
f) формируется вектор оценок коэффициентов дискриминации, который рассчитывается как разница между векторами средних по матрицам обеих групп, умноженная на обратную матрицу, полученную на предыдущем шаге. Таким образом, дискриминантная функция имеет следующий общий вид:

$$F=V\times X_{ij} \qquad (2.9)$$

где X_{ij} — значение j-того признака (результат теста) для i-того объекта;
V — значения вектора оценок коэффициентов дискриминации;

g) рассчитываются оценки дискриминантной функции путем умножения первоначальных данных на вектор оценок коэффициентов дискриминации (V), по которым рассчитываются средние значения. Из полученных значений вычисляется константа дискриминации как среднее значение по двум средним значениям векторов оценок дискриминантной функции.

Таким образом, для классификации нового объекта (пациента) требуются данные, подлежащие дискриминации (результаты тестов), и с помощью вектора оценок дискриминации рассчитываются значения дискриминантных функций для новых объектов (пациентов). Полученный результат сравниваем с константой и проводится распределение (см. Приложение А)

6.6. Проверка и последующая работа с моделью

Проверку значимости дискриминантных функций осуществляют с помощью лямбда-теста Уилкса. Этот тест показывает значительно ли дискриминантные функции различают группы.

Классификация случаев

Дискриминантные функции используются для классификации случаев по группам, путем сопоставления конкретного случая с данными групп, с расчетом предельного дискриминантного балла, сравнение с которым позволяет произвести классификацию случая.

Интерпретация

На основе дискриминантной функции проводится интерпретация результатов. Веса или коэффициенты предикторных переменных в дискриминантной функции могут указывать, какие переменные наиболее важны для различения групп.

Прогнозирование

После того, как модель дискриминантного анализа построена и проверена, ее можно использовать для прогнозирования членства в группе для новых случаев.

ПРИЛОЖЕНИЕ А
(обязательное)
Применение дискриминантного анализа на примере «Базы данных больных, страдающих хронической бронхо-легочной патологией, вызванной воздействием промышленных аэрозолей»

В настоящем Приложении рассматриваются расчеты, которые проводятся при применении DA. Теоретические аспекты проведения расчетов рассмотрены в пп. 6.4 и 6.5.

Имеется «База данных больных, страдающих хронической бронхо-легочной патологией, вызванной воздействием промышленных аэрозолей».

Дата публикации и номер бюллетеня: 14.11.2024 Бюл. № 11)[1].

Зависимая переменная — статус заболевания человека (1 — да, 2 — нет);

Независимые переменные — острота зрения на худшем глазу (диоптрии) (X1), глюкоза (ммоль/л) (X2), общий холестерин (ммоль/л) (X3), показатель активности регуляторных систем (ПАРС, баллы) (X4), гемоглобин (X5).

Так как наша зависимая переменная имеет две категории (эталонных класса), то это пример двухгруппового линейного дискриминантного анализа. Здесь имеются как дискретные (ПАРС, гемоглобин), так и непрерывные переменные (диоптрии, глюкоза, холестерин).

В таблице содержатся данные о трудоспособных (A), нетрудоспособных (D) и новых (трудоспособность, которых не установлена) пациентах по результатам медицинских тестов (X1 … X5), именуемых признаками (независимые переменные) Изначально данные представлены в следующем виде (табл. 1).

[1] В рассматриваемом примере приведены условные значения.

Таблица А.1. Первичные медицинские данные пациентов

	Острота зрения на худшем глазу (диоптрии)	Глюкоза (ммоль/л)	Общий холестерин (ммоль/л)	ПАРС (баллы)	Гемоглобин	Группа
Пациент1	1,75	4,1	4,2	1	134	A
Пациент2	1	4,4	4,5	2	137	A
Пациент3	2	4,8	3,8	1	140	A
Пациент4	0,5	5,2	3,6	2	141	A
Пациент5	1	4,7	3,9	2	138	A
Пациент6	1,5	4,6	4,1	2	140	A
Пациент7	2	4,5	4	1	136	A
Пациент8	1,75	5,1	6,2	2	130	D
Пациент9	2,5	4,6	6	3	132	D
Пациент10	2	5,7	3,8	4	128	D
Пациент11	3	5	5,5	3	119	D
Пациент12	2,75	4,9	6	2	115	D
Пациент13	2,25	5,4	5,3	3	124	D
Пациент14	2,75	5,4	3,9	4	122	N
Пациент15	1,5	3,3	5,9	3	139	N

Пациенты 14 и 15 — это новые пациенты, по которым получены результаты тестов, но не сделан вывод об их трудоспособности, на начало проведения анализа они относятся к категории N («новые»). После проведения дискриминантного анализа мы сможем определить к какой группе (категории) они относятся.

На данном этапе необходимо провести нормализацию. Проведем минимально-максимальную нормализацию. Это метод нормализации, который позволяет привести значения данных к диапазону от 0 до 1 путём масштабирования значений в соответствии с минимальным и максимальным значениями (формула расчета 2.6).

Таким образом, нормализованные данные принимают следующий вид (табл. 2).

Следующим шагом является расчет средних значений по каждому признаку (независимой переменной) для двух групп по формуле:

$$\overline{x_{jk}} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ij} \qquad \text{(A.1)}$$

где

x_{ij} — значение j-того признака (результат тестов) для i-того объекта (пациента) в рассматриваемой группе (группа A «трудоспособные» и группа D «нетрудоспособные»);

Таблица 2. Нормализованные значения первичных медицинских данных пациентов

Пациенты	Острота зрения на худшем глазу (диоптрии)	Глюкоза (ммоль/л)	Общий холестерин (ммоль/л)	ПАРС (баллы)	Гемоглобин
A1	0,5000	0,3333	0,2308	0,0000	0,7308
A2	0,2000	0,4583	0,3462	0,3333	0,8462
A3	0,6000	0,6250	0,0769	0,0000	0,9615
A4	0,0000	0,7917	0,0000	0,3333	1,0000
A5	0,2000	0,5833	0,1154	0,3333	0,8846
A6	0,4000	0,5417	0,1923	0,3333	0,9615
A7	0,6000	0,5000	0,1538	0,0000	0,8077
D1	0,5000	0,7500	1,0000	0,3333	0,5769
D2	0,8000	0,5417	0,9231	0,6667	0,6538
D3	0,6000	1,0000	0,0769	1,0000	0,5000
D4	1,0000	0,7083	0,7308	0,6667	0,1538
D5	0,9000	0,6667	0,9231	0,3333	0,0000
D6	0,7000	0,8750	0,6538	0,6667	0,3462
N1	0,9000	0,8750	0,1154	1,0000	0,2692
N2	0,4000	0,0000	0,8846	0,6667	0,9231

$\overline{x_{jk}}$ — среднее значение j-того признака (результат теста) для данной группы k («трудоспособные» или «нетрудоспособные»);

n_k — общее количество объектов (пациентов) в группе k, где k=1 — группа A «трудоспособные», k=2 — группа D «нетрудоспособные»;

$i = 1,2, \ldots, n_k$ — количество объектов в группах. В данном кейсе в группе A — 7 объектов, $n_A = 7$; а в группе D — 6 объектов, $n_D = 6$;

$j = 1,2, \ldots, m$ — количество признаков (анализов) в рассматриваемом примере $m = 5$ (диоптрии (X1), глюкоза (X2), холестерин (X3), ПАРС (X4) и гемоглобин (X5)).

В MS Excel для этого можно воспользоваться функцией СРЗНАЧ().

Итоги расчетов выглядят следующим образом (табл. 3)

Таблица 3. Среднее значение по двум группам

	Острота зрения на худшем глазу (диоптрии)	Глюкоза (ммоль/л)	Общий холестерин (ммоль/л)	ПАРС (баллы)	Гемоглобин
срзнач A	0,3571	0,5476	0,1593	0,1905	0,8846
срзнач D	0,7500	0,7569	0,7179	0,6111	0,3718

И таким образом получены векторы средних по матрицам двух групп (трудоспособные и нетрудоспособные).

На следующем шаге для каждой группы рассчитывали квадратные ковариационные матрицы размерностью 5×5 (по количеству признаков (результатов тестов)). Для этого рассчитываются центрированные матрицы, где из каждого элемента каждого признака вычитается среднее значение по данному признаку. Формула расчета:

$$M_{kc} = \begin{pmatrix} x_{11} - \overline{x_1} & x_{12} - \overline{x_2} & \cdots & x_{1m} - \overline{x_m} \\ x_{21} - \overline{x_1} & x_{22} - \overline{x_2} & \cdots & x_{2m} - \overline{x_m} \\ \vdots & \vdots & & \vdots \\ x_{n1} - \overline{x_1} & x_{n1} - \overline{x_2} & \cdots & x_{nm} - \overline{x_m} \end{pmatrix}$$

где M_{kc} — центрированные матрицы для двух групп («трудоспособные» и «нетрудоспособные»);

x_{ij}– результаты тестов (признаки) для каждого пациента (объекта);

$\overline{x_j}$ — средние значения по каждому признаку (анализы).

Центрированная матрица по трудоспособным пациентам

$$\begin{bmatrix} 0{,}1429 & -0{,}3214 & 0{,}0714 & -0{,}1905 & -0{,}1538 \\ -0{,}1571 & -0{,}1339 & 0{,}1868 & 0{,}1429 & -0{,}0385 \\ 0{,}2429 & 0{,}1161 & -0{,}0824 & -0{,}1905 & 0{,}0769 \\ -0{,}3571 & 0{,}3661 & -0{,}1593 & 0{,}1429 & 0{,}1154 \\ -0{,}1571 & 0{,}0536 & -0{,}0440 & 0{,}1429 & 0{,}0000 \\ 0{,}0429 & -0{,}0089 & 0{,}0330 & 0{,}1429 & 0{,}0769 \\ 0{,}2429 & -0{,}0714 & -0{,}0055 & -0{,}1905 & -0{,}0769 \end{bmatrix}$$

Центрированная матрица по нетрудоспособным пациентам

$$\begin{bmatrix} -0{,}2500 & -0{,}0104 & 0{,}2821 & -0{,}2778 & 0{,}2051 \\ 0{,}0500 & -0{,}3229 & 0{,}2051 & 0{,}0556 & 0{,}2821 \\ -0{,}1500 & 0{,}3646 & -0{,}6410 & 0{,}3889 & 0{,}1282 \\ 0{,}2500 & -0{,}0729 & 0{,}0128 & 0{,}0556 & -0{,}2179 \\ 0{,}1500 & -0{,}1354 & 0{,}2051 & -0{,}2778 & -0{,}3718 \\ -0{,}0500 & 0{,}1771 & -0{,}0641 & 0{,}0556 & -0{,}0256 \end{bmatrix}$$

Далее рассчитываются квадратные ковариационные матрицы по формуле:

$$M_k = \frac{1}{n_k - 1}(M_{kc}^T \times M_{kc}) \qquad \text{(А.2)}$$

где M_k — ковариационная матрица для группы k;

n_k — общее количество объектов для группы k;

M_{kc} — центрированная матрица;

$M_{kc}{}^T$ — транспонированная центрированная матрица.

В итоге получаем ковариационные матрицы для двух групп.

Трудоспособные

$$\begin{bmatrix} 0{,}0529 & -0{,}01706 & 0{,}004121 & -0{,}03492 & -0{,}00897 \\ -0{,}01706 & 0{,}020503 & -0{,}01313 & 0{,}010251 & 0{,}012286 \\ 0{,}004121 & -0{,}01313 & 0{,}012539 & 0{,}000916 & -0{,}00666 \\ -0{,}03492 & 0{,}010251 & 0{,}000916 & 0{,}031746 & 0{,}008547 \end{bmatrix}$$

Нетрудоспособные

$$\begin{bmatrix} 0{,}0350 & -0{,}01542 & 0{,}014615 & -0{,}00333 & -0{,}03308 \\ -0{,}01542 & 0{,}025984 & -0{,}04573 & 0{,}022685 & 0{,}00203 \\ 0{,}014615 & -0{,}04573 & 0{,}115779 & -0{,}07521 & -0{,}00878 \\ -0{,}00333 & 0{,}022685 & -0{,}07521 & 0{,}062963 & 0{,}019658 \\ -0{,}03308 & 0{,}00203 & -0{,}00878 & 0{,}019658 & 0{,}064892 \end{bmatrix}$$

На следующем этапе вычисляется несмещенная оценка суммарной ковариационной матрицы для двух групп:

$$M_{sum} = \frac{1}{n_D + n_A - 2}(n_D \times M_D + n_A \times M_A) \qquad (A.3)$$

где M_{sum} — несмещенная оценка суммарной ковариационной матрицы;

$$\begin{bmatrix} 0{,}05273 & -0{,}01927 & 0{,}01059 & -0{,}02404 & -0{,}02375 \\ -0{,}01927 & 0{,}02722 & -0{,}03329 & 0{,}01890 & 0{,}00893 \\ 0{,}01059 & -0{,}03329 & 0{,}07113 & -0{,}04044 & -0{,}00902 \\ -0{,}02404 & 0{,}01890 & -0{,}04044 & 0{,}05455 & 0{,}01616 \\ -0{,}01804 & 0{,}00111 & -0{,}00479 & 0{,}01072 & 0{,}03540 \end{bmatrix}$$

Полученная матрица преобразуется в обратную матрицу $(M_{D,A}^{-1})$[1]. В программе MS Excel можно воспользоваться функцией МОБР().

$$\begin{bmatrix} 53{,}40 & 67{,}39 & 39{,}66 & 26{,}26 & 16{,}96 \\ 56{,}32 & 155{,}81 & 82{,}51 & 30{,}95 & 5{,}4 \\ 34{,}11 & 82{,}54 & 68{,}77 & 36{,}88 & 2{,}76 \\ 22{,}41 & 27{,}33 & 35{,}05 & 45{,}45 & -3{,}67 \\ 23{,}28 & 32{,}36 & 16{,}32 & 3{,}64 & 38{,}21 \end{bmatrix}$$

Для выполнения следующих действий предварительно рассчитаем разницу между векторами средних значений по матрицам двух групп (D — A). Значения полученного вектора следующие:

(0,3929 0,2093 0,5586 0,4206 -0,5128)

1 Преобразование с использованием методов линейной алгебры (метод Гаусса, метод Крамера и др.)

Далее находим вектор оценок коэффициентов дискриминации, который рассчитывается как разница между векторами средних по матрицам обеих групп («нетрудоспособные» и «трудоспособные») умноженная на обратную матрицу, полученную на предыдущем шаге.

$$V = M_{D,A}^{-1}(D - A) \qquad \text{(A.4)}$$

где V — вектор оценок коэффициентов дискриминации;
$M_{D,A}^{-1}$ — обратная матрица;
D — средние значения по признакам (результатам тестов) для группы I «нетрудоспособные»;
A — средние значения по признакам (результатам тестов) для группы II «трудоспособные».

Таким образом, мы получили оценки дискриминантной функции:

(59,59 111,08 83,19 55,11 6,97)

Дискриминантная функция для рассматриваемого примера имеет следующий общий вид:

$$F_{k,i} = v_1 \times x_{i1} + v_2 \times x_{i2} + v_3 \times x_{i3} + v_4 \times x_{i4} + v_5 \times x_{i5} \qquad \text{(A.5)}$$

где
k — группа (где k=1 — группа A «трудоспособные», k=2 — группа D «нетрудоспособные»)
$x_{i\,j}$ — исходное значение j-того признака (результат теста) для i-того объекта (пациента);
υ_1, υ_2, υ_3, υ_4, υ_5 — значения вектора оценок коэффициентов дискриминации V.

С полученными значениями формула выглядит следующим образом:

$$F_{k,i} = 59{,}59 \times x_{i1} + 111{,}08 \times x_{i2} + 83{,}19 \times x_{i3} + 55{,}11 \times x_{i14} + 6{,}97 \times x_{i5} \qquad \text{(A.6)}$$

Далее рассчитываются оценки дискриминантной функции путем умножения первоначальных данных на вектор оценок коэффициентов дискриминации. Из полученных значений вычисляется константа дискриминации как среднее значение по двум средним значениям векторов оценок дискриминантной функции.

$$V_D = D \times V \quad \text{(A.7)}$$

$$V_A = A \times V \quad \text{(A.8)}$$

$$\text{Const} = \frac{V_{D\text{ср}} + V_{A\text{ср}}}{2} \quad \text{(A.9)}$$

где D — первоначальная матрица по группе I «нетрудоспособные»;

H — первоначальная матрица по группе II «трудоспособные»;

V_D — оценки дискриминантной функции по группе I «нетрудоспособные»;

V_A — оценки дискриминантной функции по группе II «трудоспособные»;

V_{Dcp} — среднее значение всех элементов вектора V_D;

V_{Acp} — среднее значение всех элементов вектора V_A;

Const — константа дискриминации.

Для определения группы, к которой относится новый пациент, используются входные данные, подлежащие дискриминации (результаты тестов), и с помощью вектора оценок дискриминации рассчитываются значения дискриминантных функций для новых объектов (пациентов).

$$V_{Ni} = N_i \times V \quad \text{(A.10)}$$

Полученный результат сравниваем с константой и проводим распределение по следующему правилу:

если $V_{Ni} > \text{Const}$, то при $V_{Dcp} > V_{Acp}$ объект относится к группе I, а при $V_{Dcp} < V_{Acp}$ к группе II;

если $V_{Ni} < \text{Const}$, то при $V_{Dcp} > V_{Acp}$ объект относится к группе II, а при $V_{Dcp} < V_{Acp}$ к группе I.

В нашем примере оценки дискриминантной функции по группе трудоспособные (V_A)

(91,11 115,89 118,28 113,28 110,85 125,07 109,72)

Среднее значение — 112,03.

Оценки дискриминантной функции по группе нетрудоспособные (V_D)

(218,68 225,92 211,82 236,87 222,84 232,45)

Среднее значение V_{Dcp} — 224,77.

Константа дискриминации Const (по ф. А.9) — (112.03+224.77)/2 = 168,4

Далее определяется принадлежность к группе новых пациентов (табл. 1, Пациенты 14 и 15) для этого необходимо нормализованные значения тестов (см. таблица 2, значения по пациентам N1 и N2) умножить на соответствующие оценки дискриминантной функции и далее сравнить с коэффициентом дискриминации. Расчет по формулам А.5 и А.6:

V_{N1}=59,59×0,9+111,08×0,875+83,19×0,1154+55,11×1-6,97×0,2692

V_{N2}=59,59×111,08×0,4+0+83,19×0,8846+55,11×0,6667-6,97×0,9231

По итогу расчетов получаем V_{N1} = 217,41, V_{N2} = 140,6. Следовательно, первый пациент относится к группе «нетрудоспособные», второй — к группе «трудоспособные».

Таким образом, было проведено определение межгрупповой разницы для анализа работоспособности лиц, обследованных в рамках ежегодного профилактического осмотра. Описанный способ может быть использован в профилактической медицине при мониторинге состояния здоровья прикрепленного контингента.

БИБЛИОГРАФИЯ

1. Николаева Н.Г., Шадривова О.В., Борзова Ю.В., Григорьев С.Г., Ицкович И.Э., Климко Н.Н. Возможности дискриминантного анализа в дифференциальной диагностике хронического аспергиллеза и немикотических поражений легких. Вестник рентгенологии и радиологии. 2023; 104(1): 6–20. https://doi.org/10.20862/0042-4676-2023-104-1-6-20
2. Гринхалъх Т. Основы доказательной медицины. — М.: Гэотар-мед, 2004.
3. Сергиенко В.И., Бондарева КБ. Математическая статистика в клинических исследованиях. — М.: Гэотар-мед, 2001.
4. Мурик С. Э. Оценка функционального состояния организма человека. В 2 ч. Ч. 1. Теоретические основы: учеб. пособие / С. Э. Мурик. — Иркутск: Издво ИГУ, 2013. — 159 с. ISBN 978-5-9624-0934-4 ISBN 978-5-9624-0935-1 (ч. 1)
5. Лучинин А.С. Прогностические модели в медицине. Клиническая онкогематология. 2023;16(1):27–36. DOI: 10.21320/2500-2139-2023-16-1-27-36
6. База данных больных, страдающих хронической бронхо-легочной патологией, вызванной воздействием промышленных аэрозолей. Номер регистрации (свидетельства): 2024625156 Дата регистрации: 14.11.2024 Номер и дата поступления заявки: 2024624046 27.09.2024 Дата публикации и номер бюллетеня: 14.11.2024 Бюл. № 11

Формат 60x90/16, объем 2,25 усл. печ. л.
Бумага 80 г/м^2. Офсетная. Гарнитура Times New Roman.
Тираж 1000 Заказ № _____
Отпечатано в типографии ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России
123098, Москва, ул. Живописная, 46
Тел. +7 (499) 190-93-90
rcdm@mail.ru, lochin59@mail.ru
www.fmbafmbc.ru