Министерство здравоохранения Российской Федерации
РОССИЙСКИЙ КАРДИОЛОГИЧЕСКИЙ
НАУЧНО-ПРОИЗВОДСТВЕННЫЙ КОМПЛЕКС

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ МЕДИЦИНСКИХ ДАННЫХ

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ
для ординаторов, аспирантов медицинских учебных заведений,
научных работников

Составители: А.Г. Кочетов
О.В. Лянг
В.П. Масенко
И.В.Жиров
С.Н.Наконечников
С.Н.Терещенко

Москва 2012

УДК 519.6
ББК 65, 74
К756

Методы статистической обработки медицинских данных: Методические рекомендации для ординаторов и аспирантов медицинских учебных заведений, научных работников / сост.: А.Г. Кочетов, О.В. Лянг., В.П. Масенко, И.В.Жиров, С.Н.Наконечников, С.Н.Терещенко – М.: РКНПК, 2012. – 42 с.

Методические рекомендации в сжатой и доступной форме содержат информацию по описанию и сущности статистических методов обработки результатов медицинских исследований. Представлены базовые понятия и принципы статистической обработки. Использование предлагаемых рекомендаций позволит избежать ошибок в выборе статистических методов при использовании различных статистических программных пакетов и в изложении результатов статистического анализа.
Для ординаторов, аспирантов, научных работников.

УДК 519.6
ББК 65,74

Оглавление

АКТУАЛЬНОСТЬ СТАТИСТИКИ В МЕДИЦИНЕ

Статистика в медицине является одним из инструментов анализа экспериментальных данных и клинических наблюдений, а также языком, с помощью которого сообщаются полученные математические результаты. Однако, это не единственная задача статистики в медицине. Математический аппарат широко применяется в диагностических целях, решении классификационных задач и поиске новых закономерностей, для постановки новых научных гипотез. Использование статистических программ предполагает знание основных методов и этапов статистического анализа: их последовательности, необходимости и достаточности. В предлагаемом изложении основной упор сделан не на детальное представление формул, составляющих статистические методы, а на их сущность и правила применения.

Статистическая обработка медицинских исследований базируется на принципе того, что верное для случайной выборки верно и для генеральной совокупности (популяции), из которой эта выборка получена. Однако выбрать или набрать истинно случайную выборку из генеральной совокупности практически очень сложно. Поэтому следует стремиться к тому, чтобы выборка была репрезентативной по отношению к изучаемой популяции, т.е. достаточно адекватно отражающей все возможные аспекты изучаемого состояния или заболевания в популяции, чему способствует чёткое формулирование цели и строгое соблюдение критериев включения и исключения как в исследование, так и в статистический анализ.

ВИДЫ СТАТИСТИЧЕСКИХ ДАННЫХ В МЕДИЦИНЕ

Статистические данные могут быть представлены как количественными (числовыми непрерывными или дискретными), так и качественными (категориальными порядковыми или номинальными) переменными. Необходимо чётко указывать тип (вид) переменной при заполнении базы данных и точно придерживаться выбранного типа данных, так как от этого может зависеть дальнейшая обработка переменных во многих используемых в настоящее время

статистических программах. Например, нельзя одновременно вносить в столбец переменных и числовые и текстовые, даже аналогичные по смыслу, данные: если заполнение «да/нет» в виде 1 или 0, то не вносить буквенные аббревиатуры и наоборот.

Количественные (числовые) данные предполагают, что переменная принимает некоторое числовое значение. Из них выделяют **дискретные** данные, которые могут принимать строго определённые значения, в то время как **непрерывные** могут быть представлены любыми значениями. Уникальным примером количественных данных является представление возраста двумя типами: в виде непрерывной переменной – указывается точный возраст пациента, и в виде дискретной переменной – указывается только количество полных лет (50,3 года и 50 лет; 50,9 года и 51 год).

Категориальность является основой смыслового понимания качественных переменных. Категориальные данные применяются для описания состояния объекта путем присвоения ему номера, соответствующего категории, к которой этот объект принадлежит. Важным условием для применения категориальных данных является принадлежность одного объекта исследования только к одной возможной категории для одного критерия.

Качественные номинальные данные используются в том случае, если категории не упорядочены. Числа в данном случае являются лишь обозначением для состояния объекта и не упорядочивают это состояние. Например, по полу: 1 – мужской, 2 – женский.

Качественные порядковые (ранговые, ординарные) данные – данные, для которых категории могут быть упорядочены. Например, от плохого самочувствия к хорошему: 1 – хорошее, 2 – удовлетворительное, 3 – плохое. На практике часто используется перевод количественных данных в качественное категориальное упорядоченное представление, особенно при расчётах пороговых значений (cut-off) для последующих расчётов характеристик риска или прогностической значимости с использованием таблицы сопряжённости. Например, 1 – концентрация общего холестерина меньше или равна 5,2 ммоль/л (отношение рисков развития ИБС менее

1, прогностическая ценность положительного результата более 80%), 2 – концентрация общего холестерина более 5,2 ммоль/л (отношение рисков развития ИБС более 1, прогностическая ценность положительного результата более 80%).

ТИПЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

В практике обработки результатов проведённых исследований используются два типа статистического анализа данных — первичный (запланированный) и вторичный (незапланированный).

Первичный анализ данных — используется для изучения и описания закономерностей, существование которых предполагается исследователем, и которые являются собственно гипотезой исследования. В таком случае анализируются признаки, изучение которых учтено при планировании исследования, и проверяются заранее сформулированные гипотезы.

Вторичный анализ данных — используется для формирования перспектив проведённого исследования, поиска, разведки потенциальных закономерностей и гипотез. В таком случае выполняется «просеивание» незапланированных в конкретной работе данных, что часто бывает целесообразно уже на первом этапе знакомства с данными.

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Одной из основных составляющих любого анализа данных является описательная статистика (дескриптивная статистика). Её главной задачей является предоставление сжатой и концентрированной характеристики изучаемого явления в числовом и графическом виде.

Популяционное значение параметра (среднее значение, медиану, долю и т.д.) **получить невозможно** (исключение составляют случаи, когда исследование проводится на группе, которая включает всех членов популяции). **Однако** популяционное значение параметра **можно оценить** по выборке. Точность такой оценки зависит от метода измерения (ошибки измерения), объема и репрезентативности выборки (ошибка выборки) и биологической вариации.

Показатели описательной статистики можно разбить на несколько групп:

- показатели положения, описывающие положение экспериментальных данных на числовой оси. Примеры таких данных – максимальный и минимальный элементы выборки, среднее значение, медиана, мода и др.;

- показатели разброса, описывающие степень разброса данных относительно центральной тенденции. К ним относятся: выборочная дисперсия, разность между минимальным и максимальным элементами (размах, интервал выборки) и др.;

- показатели асимметрии: положение медианы относительно среднего и др.;

- графические представления результатов – гистограмма, частотная диаграмма и др.

Данные показатели используются для наглядного представления и анализа результатов всей исследовательской выборки, экспериментальной и контрольной группы.

При использовании описательной статистики важно учитывать тип данных и параметры распределения, характеризующиеся показателями асимметрии и гистограммой распределения. Наиболее часто употребляемыми критериями для проверки гипотезы о законе распределения являются критерий Пирсона, критерий χ^2 и критерий Колмогорова-Смирнова: при отличии распределения признака в изучаемой выборке от нормального распределения со статистической значимостью менее 0,05 ($p<0,05$) распределение признака в выборке признаётся ненормальным, и наоборот.

Основными типами распределений признаков являются: дискретные (для дискретных признаков – биномиальное, распределение Пуассона, распределение Бернулли) и непрерывные (для непрерывных признаков – нормальное (гауссово, или распределение Гаусса), логнормальное, постоянное, экспоненциальное, хи-квадрат χ^2). В соответствии с типом распределения применяется два принципа статистической обработки:

параметрический и непараметрический. Параметрический принцип включает все методы анализа нормально распределенных количественных признаков. Непараметрический принцип используется во всех остальных случаях – для анализа количественных признаков независимо от вида их распределения и для анализа качественных признаков.

Непараметрические методы считаются менее мощными по сравнению с параметрическими, т.е. иногда они не позволяют выявить статистические закономерности, которые могут быть выявлены с помощью параметрических методов. В то же время непараметрические методы более надежны в случаях, когда есть сомнения в том, что анализируемый признак имеет нормальное распределение. Для нормально распределенных признаков параметрические и непараметрические методы дают близкие результаты.

Указание в представлении данных меры центральной тенденции (среднее, медиана, мода) автоматически сообщает читателю о нормальности распределения признака. При нормальном распределении все три показателя более или менее совпадают, а при асимметричном распределении — нет.

Мода (Mo) — это наиболее частое значение в выборке, или среднее значение класса с наибольшей частотой. Мода как центральная тенденция используется чаще всего для того, чтобы дать общее представление о распределении. В некоторых случаях у распределения могут быть две моды, в таком случае это свидетельствует о бимодальном распределении, что указывает на наличие двух относительно самостоятельных групп.

Медиана (Me, Md) соответствует центральному значению в последовательном ряду всех полученных значений или среднему значению наиболее часто встречающихся значений выборки. Медиана вместе с квартилями используется для представления дискретных переменных или количественных непрерывных переменных с ненормальным распределением.

Среднее арифметическое (M) — это показатель центральной тенденции, полученный делением суммы всех значений данных на число этих данных. Среднее арифметическое используется для представления количественных переменных с

нормальным распределением. Среднее значение, как мера центральной тенденции в описательной статистике количественных данных, имеет одно из двух представлений. Первое в виде «M±S», или в зарубежной традиции M (S), где M – среднее, а S – стандартное отклонение (Standard Deviation, равное корню квадратному из дисперсии). Стандартное отклонение предназначено для описания выборок **с нормальным распределением** и не приспособлено для распределений, отличных от нормального. При нормальном распределении в диапазон M±S укладывается порядка 70% всех значений признака.

Второе представление результатов - в виде *« $M \pm m$ »*, где m - стандартная ошибка среднего (Standard Error of Mean), определяемая следующим образом: $m = s/\sqrt{n}$. Однако, подобная форма представления данных в медицине является малоинформативной. Использование стандартной ошибки среднего используется в физике, где при измерении параметров одинаковых объектов вариабельность результатов определяется только случайными погрешностями, и при увеличении количества измерений можно получить значение среднего, более близкое к истинному, с меньшей стандартной ошибкой среднего. В медицине объектами наблюдения выступают сложные системы, значительно различающиеся по своим свойствам, что определяет практическое отсутствие истинного значения параметра. В действительности, в биологии (соответственно, и в медицине) определяется не точное значение, а диапазон, в который укладывается большинство значений исследуемого признака, т.е. **ширина распределения.** Поэтому оптимальным описанием ширины распределения в медицинских исследованиях в настоящее время принимается представление **95% доверительного интервала** с указанием нижней (5%) и верхней (95%) границы.

Доверительный интервал представляет собой диапазон значений, который с определённой исследователем вероятностью (чаще всего в медицине это α=0,05 или 95%) включает в себя настоящее популяционное значение. Например: при размере выборки исследования из 30 пациентов с

ИБС средний возраст составил 56,3 года (СО 4,26 лет) или 56,3 года (95% ДИ от 54,7 до 57,9 лет).

Наиболее адекватная **непараметрическая** характеристика ширины – это квантили. Квантили представляют собой частоту попадания значений переменной в определённые интервалы. Чаще всего используется разделение на 10 (по 10%) или на 4 интервала (25%, 50%, 75%). При разделении на четыре квантиля (именуемых квартилями) для предоставления оценки центральной тенденции, ширины и асимметрии распределения результатов достаточно трёх чисел: нижний квартиль (25%), 50% квартиль, который соответствует **медиане**, и верхний квартиль (75%). Подобный метод предоставления данных является одним из наиболее компактных и удобных. Например: при размере выборки исследования из 30 пациентов с ИБС возраст по медиане составил 56,3 года (интерквартильная широта от 55,2 до 57,8 года).

Для качественных данных единственной корректной характеристикой будет являться **число объектов** с данным конкретным значением критерия. Представляются подобные данные в виде гистограммы или количества объектов с данным конкретным значением критерия относительно общего количества объектов. **Проценты**, как относительное долевое выражение числа объектов от общего числа объектов равного 100, указываются **при объёме выборки более 20**. Причём, при объёме выборки от 20 до 99 необходимо указывать **целое число процентов**, при объёме выборки более **100** – не более чем **с одним знаком после запятой** (например: количество выживших пациентов 5 из 10 прооперированных; количество выживших пациентов 16 (53%) из 30 прооперированных; количество выживших пациентов 59 (57,8%) из 102 прооперированных). В последнее время получает широкое применение использование 95% доверительного интервала в представлении процентов, долей и, обязательно, в связанных с ними отношениях частот при анализе таблицы сопряжённости в качестве оценки вероятности событий, получившей название отношения рисков, особенно при популяционных исследованиях и мета-анализе. Наиболее

удобным и простым в таком случае для расчёта 95% доверительного интервала представляется метод Уилсона, который может использоваться при любом объёме выборки.

ИНДУКТИВНАЯ СТАТИСТИКА

Задачей индуктивной статистики является проверка статистических гипотез о законе распределения, а основной областью применения – использование в медико-биологических исследованиях для сравнения двух разных выборок на предмет принадлежности к общей генеральной совокупности. Принадлежность двух выборок к одной генеральной совокупности свидетельствует об отсутствии различия между ними.

Для этого формулируются статистические гипотезы:

- H_0 гипотеза об отсутствии различий (*нулевая гипотеза*);

- H_1 гипотеза о значимости различий (*альтернативная гипотеза*).

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

NB!!! - *p*-значение является количественной характеристикой только лишь статистической, **НО** не клинической значимости. При наличии статистической значимости необходимо принять решение о клинической важности выявленных различий. Особенно это касается вторичного анализа данных, незапланированного. При первичном запланированном анализе данных обычно проверяется статистическая значимость клинически важных различий.

Теория вероятностей в основе своей оперирует понятием допустимой

ошибки, и ошибка является обязательным компонентом статистического анализа, влияющая на *p*-значение. Допустимый уровень ошибок, от которого зависит *p*-значение, выбирается исследователем. В медико-биологических исследованиях принято использовать два вида ошибок: ошибка первого рода, которой соответствует понятие уровня статистической значимости α (альфа), и ошибка второго рода β (бета), которой соответствует понятие статистической мощности 1-β.

Ошибка первого рода (**уровень значимости α**) – допустимость ошибочного признания различий, то есть альтернативной гипотезы. В медико-биологических исследованиях в качестве критического порога значимости традиционно выбирается уровень 0,05, что допускает наличие ошибки первого рода 5 раз в 100 сравнениях. При $p \leq \alpha$ различия принимаются статистически значимыми. И чем меньше *p*-значение, тем меньше подобных ошибок: например, при p=0,01 считается, что ошибка первого рода возможна 1 раз в 100 сравнениях, при p=0,001 – 1 раз в 1000 сравнениях. Однако в разведочных/пилотных исследованиях допускается уровень значимости α=0,1 для выявления намечающихся различий и/или взаимосвязей с целью дальнейшего планирования на их основе новых исследований с достаточной значимостью.

Ошибка второго рода β (**статистическая мощность 1-β**) – допустимость ошибочного отказа от наличия различий или, что то же самое, ошибочного признания отсутствия различий, соответственно ошибочного признания нулевой гипотезы, обусловленное недостаточным количеством данных. Ошибка второго рода выражается в виде статистической мощности равной 1- β. Мощность необходима для определения достаточности объёма выборки, особенно при доказательстве отсутствия статистических значимых различий в биоэквивалентных исследованиях. При адекватной статистической мощности отсутствие статистических значимых различий действительно признаётся таковым. При неадекватной мощности нельзя утверждать об эквивалентности (схожести) групп. В медико-биологических

исследованиях в качестве критического порога принимается значение ошибки второго рода β=0,1 или β=0,2, что в виде статистической мощности, выраженной в процентах, равно 90% или 80%, чаще всего – 80%: вероятность того, что из 100 в 80 случаях действительно существующее различие будет выявлено и в 20 случаях – упущено.

Необходимым условием формирования гипотезы является **предположение о смещении признака** между изучаемыми группами: **одностороннее или двустороннее.** Вычисляемое для односторонних тестов значение статистической значимости (p) примерно в 2 раза меньше, чем для двусторонних тестов, что позволяет при обосновании одностороннего тестирования чаще выявлять клинически важные статистические закономерности. **Односторонние тесты учитывают** исходное (априорное) предположение о том, что в одной из групп распределение признака смещено в определенную сторону (в сторону увеличения либо уменьшения) по отношению к другой. Однако для того чтобы воспользоваться таким тестом, необходимо обосновать свое предположение. **Двусторонние тесты используются** в отсутствие исходного (априорного) предположения о том, что в одной из групп распределение признака смещено в определенную сторону (в сторону уменьшения или увеличения) по отношению к другой. Экспертным медицинским сообществом рекомендуется чаще использовать двусторонние тесты.

Выборки могут быть **независимыми**, если идёт сравнение контрольной и опытной группы, или **зависимыми**, если обе выборки представлены одними и теми же пациентами до и после вмешательства.

Для расчёта *p*-значения используют решающие правила – *статистические критерии*. То есть, на основании информации о результатах наблюдений (характеристиках членов экспериментальной и контрольной групп) вычисляется число, называемое *эмпирическим значением* критерия. Это число сравнивается с известным (заданным таблично) эталонным числом, называемым *критическим значением* критерия. Математическим результатом

такого сравнения является *p*-значение.

Главная задача исследователя при использовании индуктивной статистики заключается в формулировке статистических гипотез и выборе правильного статистического критерия для проверки этих гипотез (схема 1).

Схема 1. Методология индуктивной статистической обработки исследования.

Выбор критерия зависит от поставленной задачи, типа данных и количества измерений. Так, для количественных данных **при распределениях, близких к нормальным**, используют параметрические методы, основанные на таких показателях, как среднее значение и стандартное отклонение. Для сравнения двух **независимых выборок** используется **непарный t-критерий**, для двух **зависимых** выборок используется **парный t-критерий.**

При обработке **малых выборок** (**менее 16** объектов, при котором t-распределение начинает существенно отличаться от нормального) для сравнения **неколичественных данных** используют **непараметрические методы** — U-тест Манна-Уитни для двух независимых выборок, критерий Вилкоксона для сравнения

двух зависимых выборок, критерий χ^2 (хи-квадрат) для проверки статистической гипотезы о наличии связи между двумя качественными признаками.

Таблица сопряжённости

Таблица сопряжённости – это форма представления данных об объектах исследования на основе группировки двух или более признаков по принципу их сочетаемости (рис. 1).

Если первая переменная может принимать m значений, а вторая переменная n значений, то результирующая таблица сопряжения признаков будет представлять собой матрицу размером m*n, в каждую ячейку которой заносятся частоты встречаемых комбинаций признаков.

Группа	Изучаемый эффект (критерий риска, признак)		Размер группы (выборки)
	Да	Нет	
Группа 1 (экспериментальная)	36	48	84
по строке	42,9%	57,1%	100,0%
по столбцу	70,6%	43,6%	52,2%
Группа 2 (контрольная)	15	62	77
по строке	19,5%	80,5%	100,0%
по столбцу	29,4%	56,4%	47,8%
Сумма	51	110	161
по строке	31,7%	68,3%	100,0%
по столбцу	100,0%	100,0%	100,0%

Рисунок 1. Общий вид четырехпольной таблицы сопряжённости.

При составлении таблицы сопряженности следует помнить, что ее строки и колонки являются взаимоисключающими, то есть информация, касающаяся конкретного пациента, может находиться только в клетке таблицы на пересечении одной строки и одного столбца. Для применения данного критерия необходимо, чтобы в каждой клетке рассматриваемой таблицы ожидаемая частота была не ниже 5.

Для обработки данных представленных в таблицах сопряжённости признаков используют критерий χ^2. Этот критерий отвечает на вопрос о том, с одинаковой ли частотой встречаются разные значения признака в эмпирическом и теоретическом распределениях (проверка гипотезы о законе распределения) или в двух и более эмпирических распределениях (сравнение двух качественных признаков).

Для использования непараметрического метода χ^2 не требуется вычисление среднего значения или стандартного отклонения. Его преимущество состоит в том, что необходимо знать лишь зависимость распределения частот результатов от двух переменных; это позволяет выяснить, связаны они друг с другом или, наоборот, независимы. Поэтому указанный статистический метод чаще всего используется для обработки качественных данных.

Метод χ^2 для расчёта p-значения, состоит в том, что оценивают, насколько сходны между собой распределения эмпирических и теоретических частот. Если разница между ними невелика, то можно полагать, что отклонения эмпирических частот от теоретических обусловлены случайностью. Если же, напротив, эти распределения будут достаточно разными, можно будет считать, что различия между ними значимы и существует связь между действием независимой переменной и распределением эмпирических частот. Как раз для подсчёта χ^2 и строится таблица сопряжённости.

Критерий χ^2 в случае таблицы 2x2 (то есть при 1 степени свободы) дает несколько завышенные значения. Это вызвано тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен. На практике это приводит к тому, что при малом числе наблюдений нулевая гипотеза будет отвергаться слишком часто.

Аппроксимация статистики χ^2 для таблиц 2x2 с малыми числом наблюдений в ячейках может быть улучшена уменьшением абсолютного значения разностей между ожидаемыми и наблюдаемыми частотами на величину 0,5 перед возведением в квадрат – это так называемая поправка Йетса. Поправка Йетса, делающая оценку более умеренной, обычно

применяется в тех случаях, когда таблицы содержат только малые частоты, например, когда некоторые ожидаемые частоты становятся меньше 10.

Четырёхпольная таблица сопряжённости, представленная выше на рис. 1, помимо расчёта *p*-значения, используется также для расчёта характеристик риска по признакам, предположительно являющихся факторами риска или характеризующих благоприятное состояние: значений риска, шанса, отношения рисков и шансов, повышения или снижения абсолютного и относительного рисков, индекса потенциальной пользы или вреда. При изучении качественных или пороговых значений количественных лабораторных показателей эта же таблица используется для расчётов чувствительности, специфичности, прогностической значимости положительного и отрицательного результатов, отношения правдоподобий.

Точный критерий Фишера

Критерий χ^2 годится для анализа таблиц сопряженности 2x2, если ожидаемые значения в любой из ее клеток не меньше 5. Когда число наблюдений невелико, это условие не выполняется и критерий χ^2 неприменим. В этом случае используют *точный критерий Фишера*. Он основан на переборе всех возможных вариантов заполнения таблицы сопряженности при данной численности групп, поэтому, чем она меньше, тем проще его применять.

ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

На данном этапе анализа данных происходит исследование зависимостей между переменными. С этой целью применяются корреляционный анализ (для установления факта наличия или отсутствия зависимости между переменными, выраженной в виде числового значения), а также регрессионный анализ (для нахождения количественной зависимости между переменными, выраженной в виде уравнения и/или графика), и часто в последнее время - факторный анализ, рассмотренный в главе «Снижение размерности (схема 2).

Схема 2. Статистические методы поиска зависимостей между переменными.

Корреляционный анализ

Корреляция – взаимосвязь между двумя или более переменными (в последнем случае корреляция называется множественной или совокупной). Цель корреляционного анализа – установление наличия или отсутствия этой взаимосвязи. В случае, когда имеются две переменных, значения которых измерены в шкале отношений, используется коэффициент линейной корреляции Пирсона **r**, который принимает значения от -1 до +1 (нулевое его значение свидетельствует об отсутствии корреляции).

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

Для данных, измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена, так как он является непараметрическим и улавливает тенденцию – изменения переменных в одном направлении, который обозначается $\mathbf{r_s}$ и определяется сравнением *рангов* – номеров значений сравниваемых переменных в их упорядочении. Коэффициент корреляции Спирмена является менее чувствительным, чем коэффициент корреляции Пирсона.

Важно отметить, что близкое к плюс единице или к минус единице значение коэффициента корреляции говорит о силе взаимосвязи переменных прямой или обратной, но ничего не говорит о причинно-следственных отношениях между ними.

Регрессионный анализ.

В отличие от корреляционного анализа, регрессионный анализ — не только говорит о наличии зависимости между независимой переменной и одной или несколькими зависимыми переменными, но и позволяет определить эту зависимость количественно. Независимые переменные называют регрессорами или предикторами, а зависимые переменные — критериальными. Опять же, терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Существует несколько видов линейного и нелинейного регрессионного анализа, позволяющие обнаружить математическую зависимость между несколькими переменными, однако все эти методы являются параметрическими, что делает невозможным их применение для обработки качественных данных. Непараметрическим аналогом множественной регрессии является логистическая регрессия с двумя градациями зависимого признака (бинарная логистическая регрессия) и более (мультиноминальная логистическая регрессия).

Бинарная логистическая регрессия

С помощью метода бинарной логистической регрессии можно исследовать зависимость дихотомических (бинарных, имеющих только 2 категориальных значения) переменных от независимых переменных, имеющих любой вид шкалы. Как правило, в случае с дихотомическими переменными речь идёт о некотором событии, которое может произойти или не произойти; бинарная логистическая регрессия в таком случае рассчитывает вероятность наступления события в зависимости от значений независимых переменных с выводом коэффициентов регрессии для каждой такой переменной и её статистической значимости.

Вероятность наступления бинарного события рассчитывается по формуле:

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

где $z = b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n + a$,

X_1 — значения независимых переменных, b_1 — коэффициенты, расчёт которых является задачей бинарной логистической регрессии, a — константа полученного регрессионного уравнения.

Если рассчитанная вероятность имеет значение меньшее 0,5, то можно предположить, что событие не наступит; в противном случае предполагается наступление события.

Коэффициенты в полученной регрессии не следует интерпретировать как эффект от изменения X. Для правильной трактовки следует найти производную логистической функции по параметру X и вычислить предельный эффект (marginal effect) при конкретном значении переменной X (обычно вычисляется в среднем значении).

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

Оценка адекватности модели бинарной логистической регрессии

Точность результатов расчёта логистической регрессии целиком и полностью зависит от выборки, на основании которой рассчитывались коэффициенты в уравнении логистической регрессии. Таким образом, построенная модель требует проверки её адекватности.

Простейшим способом оценки адекватности модели является проверка этой модели на исходных данных и сравнение полученных данных с предварительно используемыми исходами («исходными исходами»). Оценка выражается как процент наблюдений с исходами, верно предсказанными с помощью модели регрессии.

Проверка значимости отличия коэффициентов от нуля проводится при помощи статистики Вальда, использующей распределение хи-квадрат и представляющей собой квадрат отношения соответствующего коэффициента к его стандартной ошибке.

Качество приближения регрессионной модели к гипотетически реальной оценивается при помощи функции подобия. Мерой правдоподобия служит отрицательное удвоенное значение логарифма этой функции (-2LL). В качестве начального значения для -2LL применяется значение, которое получается для регрессионной модели, содержащей только константы. Эта величина имеет распределение Хи-квадрат Пирсона и показывает уровень согласованности модели регрессии со всеми независимыми переменными.

Мультиномиальная логистическая регрессия

Этот метод является вариантом логистической регрессии, при которой зависимая переменная имеет больше двух категорий. В то время как, при бинарной логистической регрессии независимая переменная может иметь непрерывную шкалу, то мультиномиальная логистическая регрессия пригодна только для категориальных независимых переменных, причём имеет значение, относятся ли они к шкале наименований или к порядковой шкале.

Для построения мультиномиальной логистической регрессии формируется n недублированных логитов для n+1 возможных значений независимой переменной, причём одна категория используется как эталонная, ее коэффициенты принимаются равными 0:

$$g1 = ln\frac{p1}{pn} = b_{10}+b_{11}+\ldots+b_{1(n-2)}$$

$$g2 = ln\frac{p2}{pn} = b_{20}+b_{21}+\ldots+b_{2(n-2)}$$

$$gn = 0$$

Нахождение коэффициентов b_{10}, b_{11}, b_{20} и b_{21} (называемых параметрическими оценками) является основной задачей мультиномиальной логистической регрессии. Первая цифра индекса указывает на номер логита, а

вторая на порядковый номер коэффициента в данном логите, причём цифра 0 на второй позиции индекса означает константу, за которой далее следует ровно столько коэффициентов, сколько независимых переменных (факторов) взято в рассмотрение. Коэффициентам последней (эталонной) категории присваивается значение 0.

Получив значения для недублирующихся логитов, можно рассчитать значения дублирующихся логитов, используя правила вычисления логарифма.

$$ln\frac{p1}{p2} = ln\frac{p1}{pn} - ln\frac{p2}{pn}$$

Следует отметить, что прямое определение вероятности для каждой категории, значительно информативней, чем соотношение этих вероятностей между собой. Для каждой i-ой категории зависимых переменных эта вероятность может быть вычислена по следующей формуле:

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

В случае наличия лишь одной независимой переменной проведение расчёта с применением столь громоздкого метода является достаточно бессмысленным — все соотношения могут быть выяснены проще, при помощи таблиц сопряженности.

Регрессия Кокса

Регрессия Кокса, или модель пропорциональных рисков, — графическое построение и математическое представление в виде коэффициентов регрессионного уравнения, экспонент коэффициентов (отношения шансов) риска наступления события как функции, зависящей от времени, и оценка влияния каждой из независимых переменных на этот риск.

Риск наступления события — функция от времени — измеряет правдоподобие наступления события в самом ближайшем будущем для тех, кто еще находится в группе риска. Риск наступления события равен

предельному значению условной вероятности наступления события во временном промежутке *[t, t + dt]* для объектов, еще оставшихся в группе риска на момент времени *t*, деленному на длину временного интервала *dt*.

Метод Кокса основан на следующих положениях:

Логарифмическая линейность. Все объясняющие переменные влияют линейно на логарифм функции риска наступления события;

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

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

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

$\ln h_i = \ln h_0(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$, где

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

$X_1, \ldots, X_p$ — независимые переменные, регрессоры;

$\beta_1, \ldots, \beta_p$ — соответствующие коэффициенты.

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

Коэффициенты $\beta_1, \ldots, \beta_p$ отражают влияние каждой из независимых переменных (регрессоров) на функцию риска: при увеличении X_j на единицу и фиксированных значениях остальных регрессоров, риск наступления события возрастает в $exp(\beta_j)$ раз.

Визуально функция риска представлена в современных программах статистической обработки данных кривыми Каплана-Мейера по выживаемости и по кумулятивному риску и соответствующими таблицами

регрессионного логистического анализа.

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

СНИЖЕНИЕ РАЗМЕРНОСТИ

В биологических, в том числе и медицинских исследованиях всегда имеется большое общее количество исследуемых связей. Однако многие из них будут статистически не значимыми, и при отсутствии клинической значимости их не надо принимать во внимание и интерпретировать. Сколько же будет ценных с точки зрения поставленных целей исследования взаимосвязей – заранее предположить невозможно. Как правило, статистически и клинически значимыми оказывается 5-25% всех возможных связей.

Кроме того, многие показатели могут иметь между собой различно выраженные зависимости. Используя их в качестве исходного подмножества признаков можно поставить и решить задачу конструирования на их основе более сложных, интегрированных признаков с помощью факторного анализа. При этом число таких комплексных признаков (индексов) будет значительно меньше, нежели число исходных признаков. В результате этой процедуры мы получаем новые признаки, которые компактно несут в себе гораздо больший объем информации, нежели каждый из исходных признаков в отдельности. В итоге появляется возможность отфильтровать случайную составляющую и получить более надёжную информацию о структуре как самих исходных признаков, так и о структуре исследуемых групп пациентов.

Факторный анализ

Факторный анализ - это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводят к меньшему количеству независимых влияющих величин, называемых факторами (факторными комплексами, компонентами). При этом в один фактор

объединяются переменные, сильно коррелирующие между собой. Переменные из разных факторов слабо коррелируют между собой. Таким образом, целью факторного анализа является нахождение таких комплексных факторов, которые как можно более полно объясняют наблюдаемые связи между переменными, имеющимися в наличии. С помощью факторного анализа возможно выявление скрытых (латентных) переменных факторов, отвечающих за наличие линейных статистических связей (корреляций) между наблюдаемыми переменными.

Сильная ассоциация между двумя разными переменными свидетельствует об избыточности двух пунктов исследования. Зависимость между переменными можно обнаружить с помощью диаграммы рассеяния. Полученная путем аппроксимации (подгонки) линия регрессии дает графическое представление зависимости. Если определить новую переменную на основе линии регрессии, изображенной на этой диаграмме, то такая переменная будет включить в себя наиболее существенные черты обеих переменных. Новый фактор в действительности является линейной комбинацией двух исходных переменных.

Выделение главных компонентов в факторном анализе проводится по диаграмме рассеяния изучаемых переменных. Процедура выделения главных компонентов подобна вращению, доводящему до максимума дисперсию (так называемый варимакс) исходного пространства переменных. Например, на диаграмме рассеяния вы можете рассматривать линию регрессии как ось X, повернув ее так, что она совпадает с прямой регрессии. Этот тип вращения называется вращением, максимизирующим дисперсию, так как критерий (цель) вращения заключается в максимизации дисперсии (изменчивости) «новой» переменной (фактора, факторного комплекса) и минимизации разброса вокруг нее.

После того, как найдена линия, для которой дисперсия максимальна, вокруг нее остается некоторый разброс данных. И процедура повторяется. В анализе главных компонентов именно так и делается: после того, как первый

фактор выделен, то есть, после того, как первая линия проведена, определяется следующая линия, максимизирующая остаточную вариацию (разброс данных вокруг первой прямой), и т.д. Таким образом, факторы последовательно выделяются один за другим. Так как каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, то факторы оказываются независимыми друг от друга. Другими словами, некоррелированными или ортогональными. При повторных итерациях выделяются факторы все с меньшей и меньшей дисперсией.

Решение об остановке процедуры выделения факторов принимается на основании двух наиболее распространённых рекомендаций: критерия Кайзера и критерия каменистой осыпи.

Критерий Кайзера. Если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser, 1960), и является, вероятно, наиболее широко используемым.

Критерий каменистой осыпи. Критерий каменистой осыпи является графическим методом, впервые предложенным Кэттелем (Cattell, 1966). Собственные значения (факторную нагрузку) можно изобразить в виде простого графика (рис. 2). Компонентополагающим числом на графике является место, где убывание собственных значений слева направо представляет собою крутой склон, расстояние между точками примерно равно единице и собственное значение компонента больше единицы. Незначимые компоненты находятся далее на максимально замедленной части кривой, расстояние между точками меньше единицы и собственное значение компонента меньше единицы, так называемый «щебень».

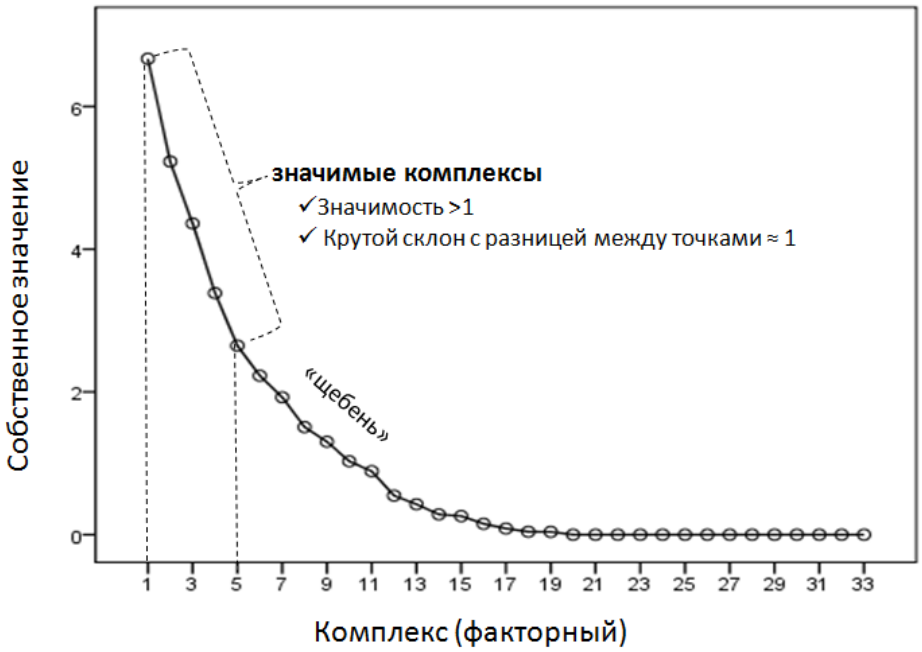


Рисунок 2. Точечная диаграмма значимости факторных моделей.

Первый критерий (критерий Кайзера) часто сохраняет слишком много факторов, в то время как второй критерий (критерий каменистой осыпи) может сохранить слишком мало факторов; однако оба критерия вполне хороши при условиях, когда имеется относительно небольшое число факторных комплексов (моделей) и много переменных. На практике возникает важный дополнительный вопрос: какое количество компонентов может быть содержательно интерпретировано. Поэтому обычно исследуется несколько решений с большим или меньшим числом факторных комплексов, и затем выбирается одно наиболее «осмысленное» и клинически значимое.

Оценка изучаемой переменной в новом факторном комплексе (компоненте) представляется в матрице компонентов векторной нагрузкой переменной, которая по сути своей является корреляционным коэффициентом взаимосвязи переменной и нового факторного комплекса (компонента). Какую матрицу компонентов, с повёрнутым решением или нет, представлять в результатах статистической обработки, зависит от выбора исследователя, учитывающего и описывающего клиническую целесообразность зависимостей переменных в компоненте и количества компонентов.

Пример факторного анализа по статистическому программному пакету SPSS:

1) Выбор количества компонентов по полной объяснённой дисперсии и точечной диаграмме (табл. 1, рис. 3)

Таблица 1. Полная объяснённая дисперсия компонентов.

Компонент	Начальные собственные значения			Суммы квадратов нагрузок извлечения			Суммы квадратов нагрузок вращения		
	Итого	% Дисперсии	Кумулятивный %	Итого	% Дисперсии	Кумулятивный %	Итого	% Дисперсии	Кумулятивный %
1	1,386	27,724	27,724	1,386	27,724	27,724	1,347	26,935	26,935
2	1,035	20,690	48,414	1,035	20,690	48,414	1,074	21,479	48,414
3	0,964	19,271	67,685						
4	0,896	17,913	85,598						
5	0,720	14,402	100,000						

Метод выделения: Анализ главных компонентов.

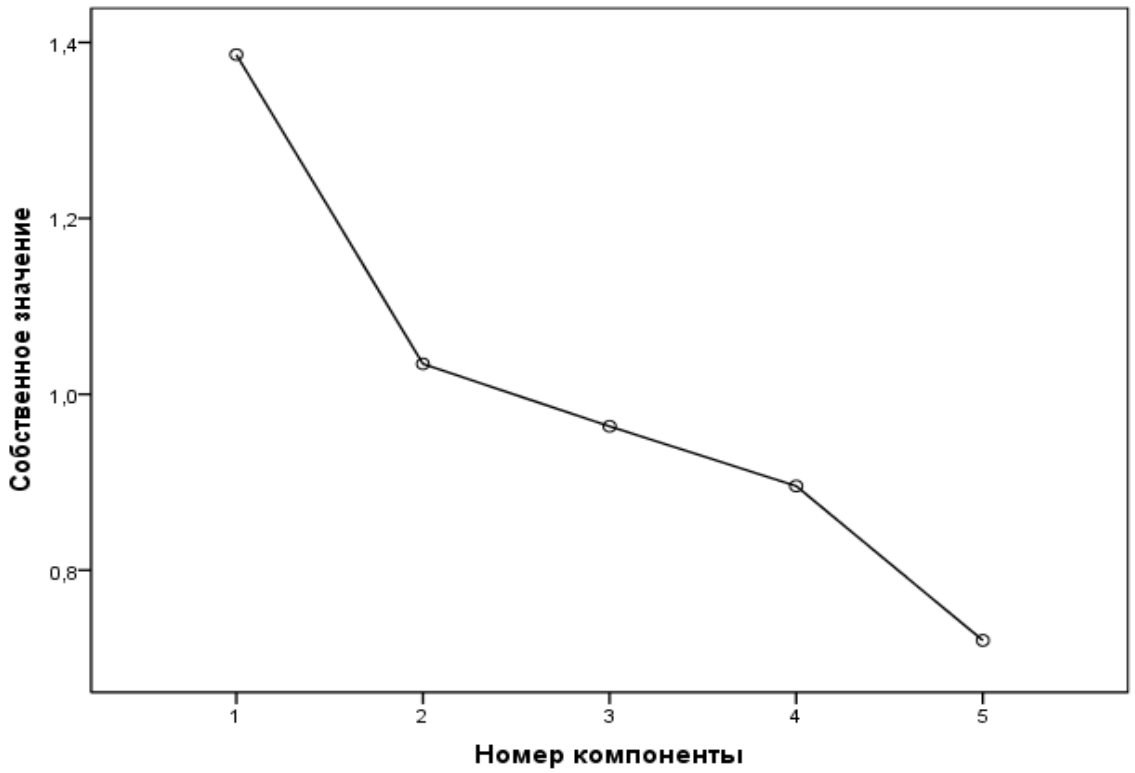


Рисунок 3. График нормализованного простого стресса.

2) Визуальная оценка диаграммы рассеяния (рис. 4).

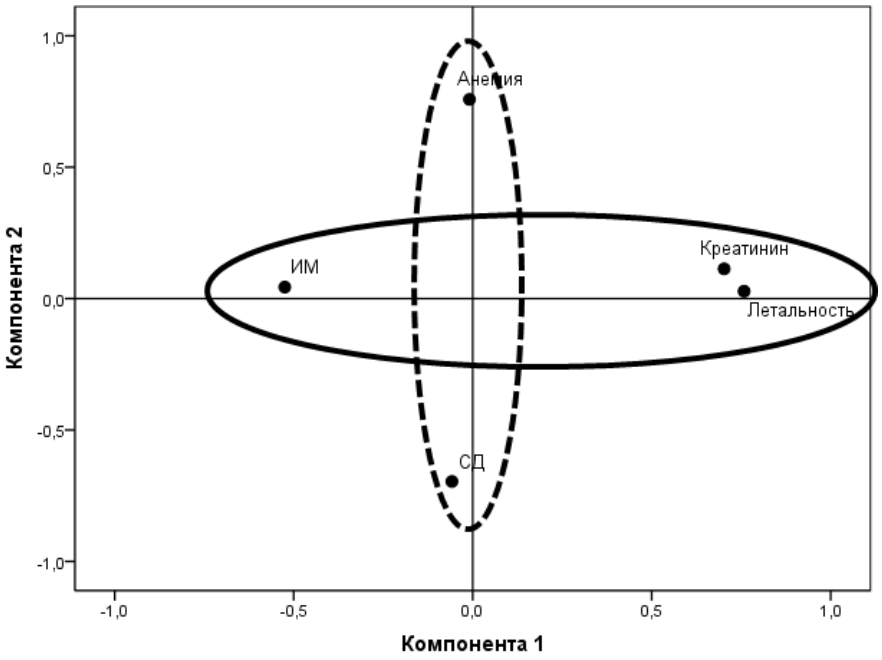


Рисунок 4. График компонентов в повёрнутом пространстве.

3) Выбор матрицы компонентов (табл. 2 и табл. 3).

Таблица 2. Матрица компонентов

	Компонент	
	1	2
Летальный исход	0,724	
Креатинин при поступлении	0,699	
Инфаркт миокарда	0,480	
Анемия при поступлении		-0,717
Сахарный диабет		-0,636

Метод выделения: Анализ методом главных компонентов.
а. Извлеченных компонентов: 2

Матрица неповёрнутых компонентов в первом наиболее значимом компоненте с факторной нагрузкой 1,386 показывает сильную прямую ассоциацию наличия летального исхода (векторная нагрузка переменной +0,724) с концентрацией креатинина (+0,699) и среднюю – с наличием инфаркта миокарда. Второй, менее значимый компонент (факторная нагрузка

1,035), сильно ассоциирует отсутствие анемии (-0,717) с отсутствием сахарного диабета (-0,636).

Таблица 3. Матрица повёрнутых компонентов

	Компонент	
	1	2
Летальный исход	0,758	
Креатинин при поступлении	0,702	
Инфаркт миокарда	0,525	
Анемия при поступлении		0,758
Сахарный диабет		0,696

Метод выделения: Анализ методом главных компонентов.
а. Извлеченных компонентов: 2

В матрице повёрнутых компонентов в первом компоненте сохранились и усилились векторные нагрузки переменных. Во втором компоненте с большей силой ассоциированы уже наличие анемии и сахарного диабета. С точки зрения здравого смысла и клинической интерпретации полученных результатов факторного анализа наиболее целесообразным является представление в результатах статистической обработки данных матрицы повёрнутых компонентов.

КЛАССИФИКАЦИЯ И ПРОГНОЗ

Важной задачей статистического анализа данных является классификация. Принято выделять три подобласти теории классификации: группировка, дискриминация (дискриминантный анализ) и кластеризация (кластерный анализ).

Группировка

Часто используемым подходом к классификации объектов исследования может служить запланированная группировка. При данном подходе разделение на группы производится условно самим исследователем. На этом принципе основаны когортные исследования.

Когортное исследование — это обсервационное исследование, в котором выделенную группу людей (когорту) наблюдают в течение некоторого времени. Исходы у испытуемых в разных подгруппах данной когорты, тех, кто подвергался

или не подвергался воздействию определённых факторов риска, сравниваются. В проспективном когортном исследовании когорты составляют в настоящем и наблюдают их в будущем. В ретроспективном (или историческом) когортном исследовании когорту подбирают по архивным записям и прослеживают их исходы с того момента по настоящее время.

Когортные исследования могут быть либо фиксированными, либо динамическими. В фиксированной когорте число пациентов остаётся неизменным, если пациенты покидают фиксированную когорту, то их не заменяют. В динамических когортах возможно, как исключение пациентов из когорты, так и включение новых пациентов в когорту.

Главным недостатком когортных исследований является их высокая стоимость, вызванная как необходимостью формирования больших выборок, так и длительностью наблюдения за больными. Кроме того лонгитюдность (продолжительность наблюдения, изучения в динамике) исследований приводит к тому, что к окончанию исследования увеличение потери больных ведет к смещению результатов. Кроме того трудно сохранить последовательность измерений и результатов в течение значительного времени, а исход заболевания, его вероятность или этиология как таковая могут измениться.

Дискриминантный анализ

В *дискриминантом анализе* классы (группы) предполагаются уже заданными, а задача заключается в том, чтобы вновь появляющийся объект отнести к одному из этих классов на основании значения некой переменной. Основная идея дискриминантного анализа заключается в том, чтобы определить, отличаются ли разные совокупности по среднему какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе. Таким образом, априорная классификация (предсказание для новых объектов) строится на основании значения весов классификации, построенных с помощью функции классификации на основе

апостериорной (на основе имеющихся данных) классификации.

Подобный подход к классификации содержит в себе несколько значительных ограничений. Во-первых, дискриминантный анализ требует, чтобы классы были заданы заранее, что в ряде случаев может быть затруднительно, и во-вторых, результаты априорной классификации (для новых наблюдений) всегда менее точны, чем результаты апостериорной классификации. К тому же, использование среднего в дискриминантном анализе предполагает нормальное распределение анализируемых выборок.

И группировка, и дискриминантный анализ имеют один общий недостаток: оба этих метода привносят структуру классов извне, вместо определения реальной структуры. С точки зрения вторичного анализа данный подход является вредоносным, поскольку прямо противоречит основной задаче вторичного анализа – незапланированный поиск новых структур и взаимосвязей. В отличие от этих двух методов, кластерный анализ предназначен для построения структуры классов на основании данных выборки и идеально подходит для задачи построения классификации в рамках вторичного анализа данных.

Кластерный анализ

Кластерный анализ является методом поиска закономерностей группирования, как объектов исследования, так и признаков в отдельные локальные подмножества (кластеры).

Задача *кластерного анализа* заключается в выделении по эмпирическим данным резко различающихся групп (кластеров) объектов, которые схожи между собой внутри каждой из групп. С помощью кластерного анализа можно производить группировку объектов исследования в кластеры, группировку признаков в кластеры (редукция числа переменных), одновременную группировку объектов исследования и признаков.

Группировка объектов исследования в кластеры применяется в тех случаях, когда предполагается, что имеющаяся выборка гетерогенна, но

причина гетерогенности при этом неизвестна. Результатом применения процедуры кластеризации может быть формирование нескольких подгрупп (кластеров) объектов исследования, в каждой из которых содержатся сходные наблюдения. Дальнейший анализ подгрупп может выявить некоторые объективные признаки, по которым эти подгруппы различаются.

Группировка признаков в кластеры применяется на достаточно однородной (в отношении наблюдений, или объектов исследования) выборке с целью поиска неизвестных закономерностей связи признаков (или групп признаков). Результатом может быть формирование нескольких групп признаков, в каждой из которых содержатся признаки, обнаружившие статистически значимые взаимосвязи.

Исследования, использующие кластерный анализ, характеризуют следующие пять основных шагов:

1) отбор выборки для кластеризации;

2) определение множества признаков, по которым будут оцениваться объекты в выборке, и способа их стандартизации;

3) вычисление значений той или иной меры сходства между объектами;

4) применение метода кластерного анализа для создания групп сходных объектов;

5) проверка достоверности результатов кластерного решения.

Проведение кластерного анализа и интерпретация его результатов достаточно сложны. Существует около 100 разных алгоритмов кластеризации, однако наиболее часто используемые: иерархический кластерный анализ и кластеризация методом k-средних.

В иерархических методах каждое наблюдение образовывает сначала свой отдельный кластер, состоящий из одного объекта. На первом шаге два соседних кластера объединяются в один; этот процесс может продолжаться до тех пор, пока не останутся только два кластера. Расстояние между кластерами является средним значением всех расстояний между всеми возможными парами точек из обоих кластеров.

Для определения количества кластеров, которое следовало бы считать оптимальным, решающее значение имеет расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учётом предусмотренного преобразования значений. Для метрических данных это квадрат евклидового расстояния, определенный с использованием стандартизованных значений. На этапе, когда мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить, так как в противном случае будут объединены уже кластеры, находящиеся на относительно большом расстоянии друг от друга.

Оптимальным считается число кластеров равное разности количества наблюдений и количества шагов, после которого коэффициент увеличивается скачкообразно.

Иерархические методы объединения, хотя и точны, но трудоёмки: на каждом шаге необходимо выстраивать дистанционную матрицу для всех текущих кластеров. Расчётное время растёт пропорционально третьей степени количества наблюдений, что при наличии нескольких тысяч наблюдений может серьёзно замедлить работу статистической программы.

Поэтому при наличии большого количества наблюдений применяют другие методы. Их недостаток заключается в том, что необходимо заранее задавать количество кластеров, а не так, как в иерархическом анализе, получить это в качестве результата. Эту проблему можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров. Если количество кластеров указать предварительно, то появляется следующая проблема: определение начальных значений центров кластеров. Их также можно взять из предварительно проведённого иерархического анализа, в котором для каждого наблюдения рассчитывают средние значения переменных, использовавшихся при анализе.

Если нет желания проходить весь этот длинный путь, то можно воспользоваться другим методом. Если количество кластеров k, которое необходимо получить в результате объединения, задано заранее, то первые k наблюдений, содержащихся в файле, используются как первые кластеры. На последующих шагах кластерный центр заменяется наблюдением, если наименьшее расстояние от него до кластерного центра больше расстояния между двумя ближайшими кластерами. По этому правилу заменяется тот кластерный центр, который находится ближе всего к данному наблюдению. В результате получается новый набор исходных кластерных центров. Для завершения шага процедуры рассчитывается новое положение кластерных центров, а наблюдения перераспределяются между кластерами с изменёнными центрами. Этот итерационный процесс продолжается до тех пор, пока кластерные центры не перестанут изменять свое положение или пока не будет достигнуто максимальное число итераций.

Главной информацией, получаемой по окончанию кластерного анализа, является принадлежность конкретного наблюдения к тому или иному кластеру. Кроме того, в результате кластерного анализа выделяются такие показатели, как средние по кластерам, дендрограммы, дисперсионный анализ, а в ряде программ приводится и такая важная информация, как среднее расстояние до центра кластера (для каждого из кластеров), максимальное и минимальное расстояние и, соответственно, наиболее удаленное и наиболее близкое к центру кластера наблюдение (типичный, эталонный представитель данного кластера), а также доля дисперсии расстояния, объясняемая кластерным разбиением (коэффициент детерминации R^2) и т.д.

Разные кластерные методы дают различные решения для одних и тех же данных. Это обычное явление в большинстве прикладных исследований. Цель кластерного анализа заключается в поиске существующих группировок данных. В то же время структуризация данных может приводить к порождению артефактов.

В качестве одного из возможных способов проверки устойчивости результатов кластерного анализа может быть использован метод сравнения результатов полученных для различных алгоритмов кластеризации. Другое средство проверки устойчивости кластерного решения может заключаться в том, чтобы исходную выборку случайным образом разделить на две примерно равные части, провести кластеризацию обеих частей и затем сравнить полученные результаты. Более сложный путь предполагает последовательное исключение вначале первого объекта и кластеризацию оставшихся (N - 1) объектов. Далее последовательно проводя эту процедуру с исключением второго, третьего и т.д. объектов анализируется структура всех N полученных кластеров. Другой алгоритм проверки устойчивости предполагает многократное размножение, дублирование исходной выборки из N объектов, объединение всех дублированных выборок в одну большую выборку (псевдогенеральную совокупность) и случайное извлечение из нее новой выборки из N объектов. После чего проводится кластеризация этой выборки, далее извлекается новая случайная выборка и вновь проводится кластеризация и т.д. Это тоже достаточно трудоемкий путь.

Не меньше проблем возникает и при оценке качества кластеризации, так как не существует универсального критерия оптимизации кластерного решения. Наилучшим способом утвердиться в том, что найденное кластерное решение является на данном этапе исследования оптимальным, является только согласованность этого решения с выводами, полученными с помощью других методов многомерной статистики, либо проверка предсказывающих моментов полученного решения уже на других объектах исследования.

АНАЛИЗ ВРЕМЕНИ ДО НАСТУПЛЕНИЯ СОБЫТИЯ

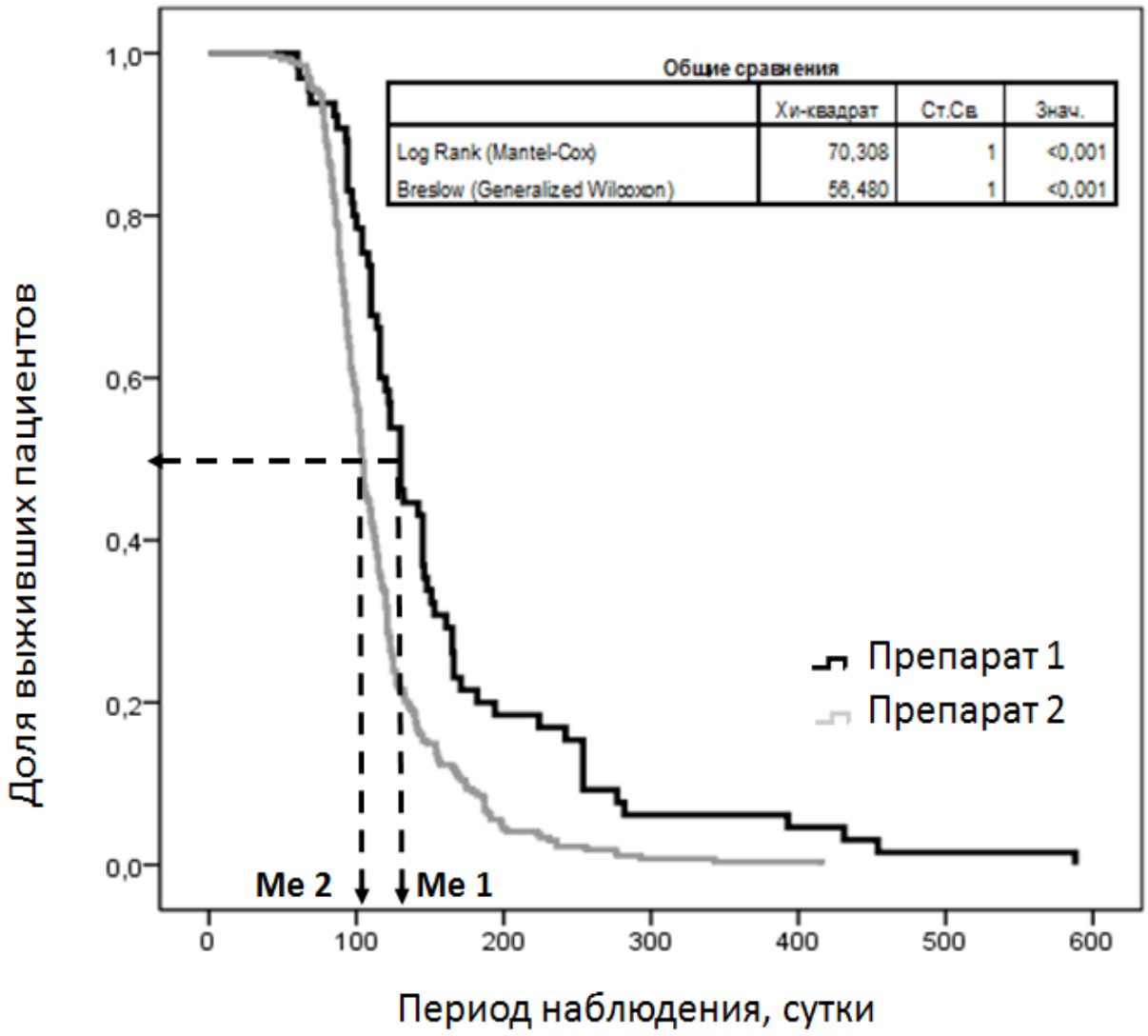
Анализ времени до наступления события в медицине чаще всего называется анализом выживаемости, так как в медицинских исследованиях принято обязательно оценивать вероятность выживания во времени, а событием (конечной точкой), по которому она определяется, является смерть.

На самом деле этим событием может быть не только смерть, но и любое другое. Функция же времени тоже может быть функцией любой другой количественной непрерывной переменной.

Ожидаемое событие при указанном анализе может не наступить. Такие случаи, когда событие ещё не наступило или о нём неизвестно, называются цензурированными. На цензурированных событиях, при конечной точке «смерть» это будет событие «жив», и основан весь анализ времени до наступления события, поэтому он получил название анализа выживаемости, а кривая её отражающая – кривой выживаемости. Другое название представляемого анализа – метод Каплана-Мейера, а кривой – кривая Каплана-Мейера.

По оси Y в кривой Каплана-Мейера откладывается процент или доля выживших (для других переменных это будет процент или доля пациентов с отсутствием события или отсутствием проецируемой точки по оси X), по оси X – время наблюдения (или для других переменных – числовое количественное непрерывное значение). Примеры представлены на рис. 5 и 6: в первом случае изображены кривые выживаемости в двух группах лечения, во втором – кривые скорости клубочковой фильтрации (СКФ) по Каплану-Мейеру у выживших и умерших больных, использованные для подтверждения и нахождения пороговых значений СКФ в прогнозе летального исхода.

Важным в анализе является представление не только кривой, но и оценок частоты наступления события (медианы и/или средней и 95% доверительного интервала) с указанием статистической значимости различий оценок наступления событий в анализируемых выборках (расхождения кривых на момент оценки) по критериям Бреслау (другое название - генерализованный критерий Вилкоксона), учитывающий ранние различия в вероятностях события, и Лог Ранка (другое название – Мантела-Кокса), учитывающий поздние различия в вероятностях события.



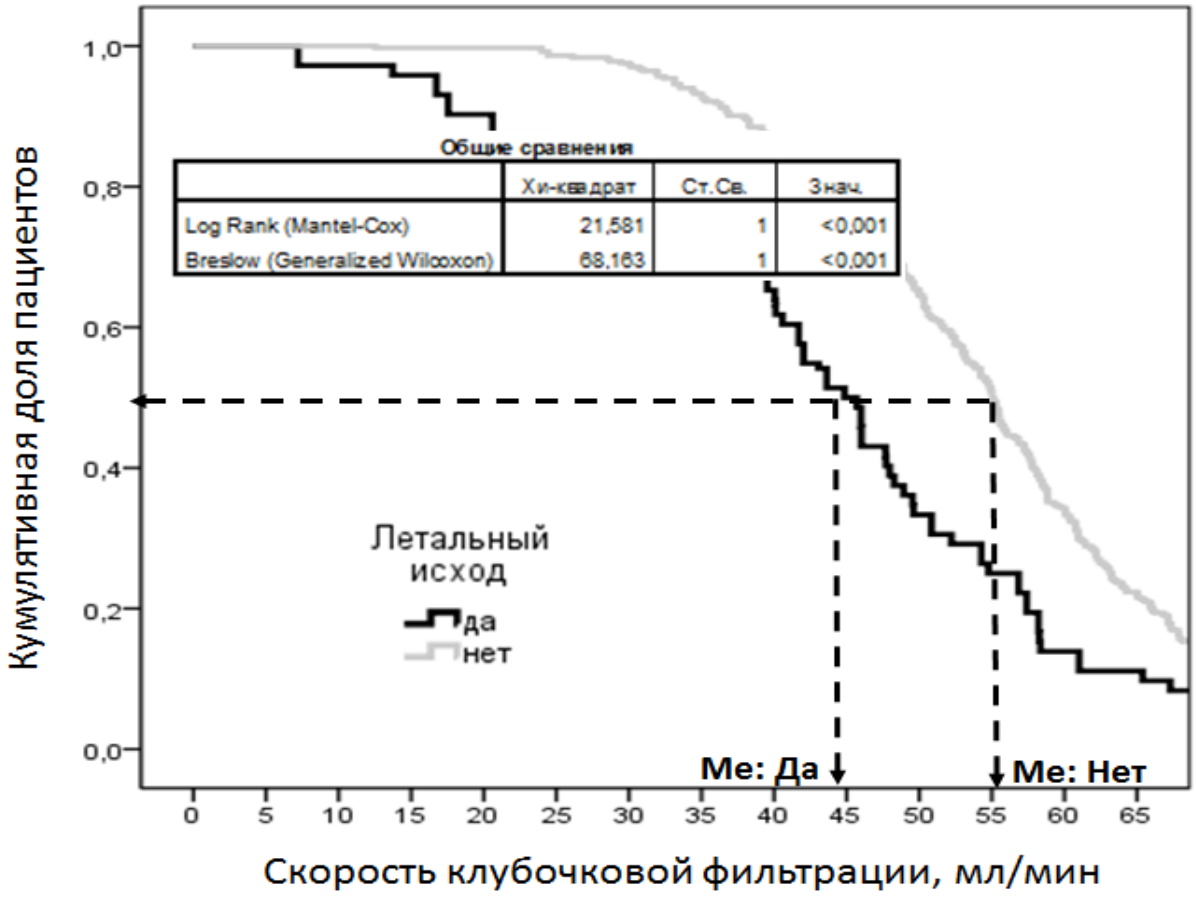
Число живых, процентили

Препарат	25%		50% (Me)		75%	
	Оценка	Ст. ошибка	Оценка	Ст. ошибка	Оценка	Ст. ошибка
1	151	6,43	125	2,43	108	2,97
2	121	0,92	107	0,93	94	0,54
Всего	125	1,36	109	0,84	95	0,63

Период наблюдения, сутки

Рисунок 5. Кривая выживаемости (Каплана-Мейера).

Из рис. 5 видно, что кривая выживаемости (Каплана-Мейера) представляет собою ступенчатую функцию, которая показывает оценки числа пациентов, выраженного в долях (или процентах), остающихся в живых на различных временных интервалах с начала исследования.



Число пациентов с представленной в таблице СКФ, процентили

Летальный исход	25,0%		50,0%		75,0%	
	Оценка	Ст. ошибка	Оценка	Ст. ошибка	Оценка	Ст. ошибка
да	54,78	2,13	44,84	1,23	34,56	2,35
нет	63,24	0,77	55,03	0,49	45,87	0,78
Всего	62,14	0,70	54,17	0,60	43,99	0,63

Рисунок 6. Концентрационные кривые СКФ по Каплану-Мейеру у выживших и умерших больных.

Необходимо помнить, что анализируемое событие (или переменная отклика) является на самом деле временем или концентрацией (и т.п.) **ДО** наступления события, а не самим событием.

Начало времени исследования у всех пациентов при анализе выживаемости должно быть одинаковым (например, при 6-визитовом исследовании - визит 1, но не с визита 3). Иначе возникает упреждающее смещение, которое ложно показывает статистически значимые различия, обусловленные тем, что одним больным диагноз поставили раньше, чем другим.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ.

1. http://www.biometrica.tomsk.ru/urolitiaz.htm. Леонов, В.П. Применение разведочного анализа для оценки исходных данных (на примере наблюдений по уролитиазу) / В.П. Леонов // Электронный журнал Биометрика. – 2005.
2. http://www.biometrica.tomsk.ru/problem.htm. Леонов, В.П. Общие проблемы применения статистики в биомедицине или что разумнее: ДДПП или ДППД? / В.П. Леонов // Электронный журнал Биометрика. – 2005.
3. Бурдяк, А.Я. Применение метода «анализ наступления события (event history analysis)» с помощью пакета SPSS / А.Я. Бурдяк // Spero. – 2007. - №6. – С.189-202.
4. Гланц, С. Медико-биологическая статистика / С. Гланц. – М. : Практика, 1999. – 334 с.
5. Гринхальх, Т. Основы доказательной медицины / Т. Гринхальх. – М. : ГЭОТАР-МЕД, 2004. – 240с.
6. Иванов, О.И. Обработка результатов медико-биологических исследований на микрокалькуляторах по программам / О.И. Иванов, О.Н. Погорелюк. - М.: Медицина. - 1990. – 218 с.
7. Ким, Дж.О. Факторный, дискриминантный и кластерный анализ / Дж.О. Ким, Ч.У. Мьюллер, У.Р. Клекка. – М. : Финансы и статистика, 1989. – 215с.
8. Количественные методы в исторических исследованиях / Под ред. И.Д. Ковальченко. – М.,1984. – 384с.
9. Ланг, Т.А. Описание статистики в медицине. Руководство для авторов, редакторов и рецензентов / Т.А.Ланг, М.Сесик. - М. : Практическая медицина. – 2011. – 477с.
10. Мандель, И.Д. Кластерный анализ / И.Д. Мандель. - М. : Финансы и статистика, 1988. – 176с.

11. Наследов, А. SPSS. Компьютерный анализ данных в психологии и социальных науках / А. Наследов. – СПб.: Питер, 2005. – 416с.
12. Новиков, Д.А. Статистические методы в медико-биологическом эксперименте (типовые случаи) / Д.А. Новиков, В.В. Новочадов. – Волгоград: ВолГМУ, 2005. – 84 с.
13. Петри, А. Наглядная медицинская статистика / А. Петри, К. Сэбин. – Москва : ГЭОТАР-МЕД, 2010. – 169 с.
14. Платонов, А.Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компъютерные методы / А.Е. Платонов. – М. : Издательство РАМН, 2001. – 52 с.
15. Пэтри, А. Наглядная статистика в медицине / А. Пэтри, К. Сэбин. – М. : ГЭОТАР-МЕД, 2003. – 144 с.
16. Реброва, О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О.Ю. Реброва. – МедиаСфера, 2002. – 70с.
17. СПСС (SPSS): искусство обработки информации / Под редакцией А. Бююль, П. Цёфель. – Москва, Санкт-Петербург, Киев: ТИД «DiaSoft», 2005. – 602 с.
18. Юнкеров, В.И. Математико-статистическая обработка данных медицинских исследований / В.И. Юнкеров, С.Е. Григорьев. – СПб.: ВМедА, 2002. – 266 с.
19. Altman, D.G. Statistical guidelines for contributors to medical journals / D.G. Altman, S.M. Gore, M.J. Gardner, S.J. Pocock // British Medical Journal. – 1983. - №286. – P. 1489-1493.
20. Brown, L.D. Interval estimation for a binomial proportion / L.D. Brown, T.T. Cai, A. Dasgupta // Statistical science. – 2001. – №2. – P. 101–133.
21. Garcia-Perez, M.A. On the confidence interval for the binomial parameter / M.A. Garcia-Perez // Quality and quantity. – 2005. – N 39. – P. 467–481.
22. Spriestersbach, A. Descriptive Statistics The Specification of Statistical Measures and Their Presentation in Tables and Graphs. Part 7 of a Series on

Evaluation of Scientific Publications / A. Spriestersbach, B. Röhrig, J.-B. du Prel et al. // Dtsch Arztebl Int. – 2009. – 106. – P. 578–583.

23. Tandy, R.D. Technical Note: The Initial Stages of Statistical Data Analysis / R.D. Tandy // Journal of Athletic Training. – 1998. – P.69-71.
24. Tzoulaki, I. Risk of cardiovascular disease and all cause mortality among patients with type 2 diabetes prescribed oral antidiabetes drugs: retrospective cohort study using UK general practice research database / I. Tzoulaki, M. Molokhia, V. Curcin et al. // British Medical Journal. – 2009. - №339. - b4731.
25. Wilson E.B. Probable inference, the law of succession, and statistical inference / E. B. Wilson // Journal of American Statistical Association. – 1927. – №22. – P. 209-212.