Румянцев П.О., Саенко В.А., Румянцева У.В., Чекин С.Ю.*

ГУ-Медицинский радиологический научный центр РАМН, г. Обнинск (директор - академик РАМН А.Ф. Цыб)
** - автор, принимавший участие в написание второй части обзора.*

Статистические методы анализа в клинической практике

Статистический анализ является интегральной частью клинического исследования. Цель настоящей работы - помочь клиницистам разобраться в сути различных методов статистической обработки медицинских данных, не углубляясь в детали математических расчетов. Рассматриваются наиболее востребованные и популярные виды анализа, применяемые в клинической и экспериментальной медицине. В первой части обзора внимание уделено описательной статистике и методам одномерного анализа, вторая часть посвящена анализу выживаемости и многомерной статистике. Ключевые слова: *методы статистического анализа, медицина, описательная статистика, алгоритм, распределение, параметрическая и непараметрическая статистика, достоверность, статистическая мощность, линейная регрессия, диагностическая информативность. логистическая регрессия, отношение рисков, анализ выживаемости, таблицы дожития, метод Каплан-Мейера, лог-ранк, модель Кокса, моделирование, многомерная статистика.*

Statistical analysis is an integral part of clinical studies. The aim of this work is to assist clinicians in getting insights into various methods of medical data processing. Here we describe the most relevant types of statistical analyses widely used in clinical and experimental medicine. First part of the paper focuses on descriptive statistics and univariate analysis, and the second reviews survival analysis and multivariate methods. *Key words: statistical analysis, medicine, descriptive statistics, algorithm, distribution, parametric and nonparametric statistics, significance, statistical power, linear regression, diagnostic test efficacy, logistic regression, hazard ratio, survival analysis, mortality tables, Kaplan-Meir method, log-rank, Cox model, modeling, multivariate statistic.*

Оглавление

ВВЕДЕНИЕ

На протяжении всей своей истории медицина искала пути повышения эффективности результатов диагностики и лечения. Начиная с интуитивный обобщений, методом проб и ошибок, через осмысление разрозненного эмпирического опыта, она вступила в эпоху доказательности. В настоящее время каждый вывод, предлагаемый специалистам и общественности, основывается на убедительных аргументах, а данные, из которых этот вывод вытекает, должны быть получены в ходе четко спланированного исследования, использующего адекватные методы статистического анализа.

Любое исследование начинается с определения его цели. Таковой, например, может быть изучение эффективности фармакологического препарата или новой процедуры в лечении заболевания. В протоколе будущего исследования четко указываются все данные, которые должны быть собраны в ходе его выполнения, методика получения каждого результата, а также, подчеркнем, заранее определяются методы статистической обработки. Производится предварительная оценка необходимой мощности исследования, также основывающаяся на статистических методах. Только при соблюдении такой методологии протокола результаты исследования могут считаться доказательными.

Ввиду того, что объемы данных и размеры групп (выборок) могут сильно варьировать, а данные быть весьма разнообразными, возникает необходимость использования методов статистического анализа, адекватных задаче. Расчет статистических показателей, которые позволяют оценить достоверность различия, корреляцию и взаимное влияние анализируемых факторов происходит по определенной технологии с использованием математических функций и создания моделей. Назначение статистического анализа состоит в объективизации суждений о результатах исследования и обеспечении

доказательствами правомочности сформулированных выводов.

Сегодня нет недостатка в статистических программных пакетах (SPSS, Statistica, S-Plus, MedCalc, StatDirect и др.), а также в персональных компьютерах, производительность которых вполне достаточна для сложных математических вычислений. Необходимо отметить, что практически все статистические пакеты разработаны за рубежом и имеют оригинальный интерфейс на английском языке. Большинство научных публикаций в мире также выходит на английском языке. Все это предопределяет необходимость знания специальных иностранных терминов и определений. Чтобы успешно использовать имеющиеся программно-технические ресурсы, клиницисту нужно также понимать основы и логику применения статистического анализа. Без этого даже наличие доступных программно-технических средств автоматически не приводит к доказательности. Скорее наоборот, для неискушенного исследователя они представляют соблазнительную возможность попытаться быстро проанализировать свои данные с целью обнаружить статистическую значимость собственных результатов. Нередко это достигается путем загрузки имеющихся данных в статистическую программу, после чего практически наугад выбирается статистический тест, который возвращает желаемый, предпочтительно максимально высокий, показатель «статистической значимости». Очевидно, подобный подход никак не отвечает принципу доказательности.

Несмотря на упомянутую доступность компьютерной техники и программного обеспечения с приемлемо дружественным интерфейсом, комплексная статистическая обработка представляет собой сложную задачу. Во многих, если не в большинстве, случаев для глубокого анализа клинических данных необходимым является участие специалиста с профессиональной подготовкой в математической статистике. Подобное сотрудничество является характерным примером того, что современный уровень развития науки все больше нуждается в интенсивном взаимодействии специалистов различных областей знания.

Целью данного обзора является попытка донести до клиницистов в упрощенной и доступной для понимания форме логику и методологию современной аналитической статистики, применяемой в мировой медицине. Хотелось бы надеяться, что это поможет врачам взвешенно осуществлять планирование (дизайн) исследования, корректно анализировать полученные данные и верно интерпретировать результаты анализа. В этой работе мы намеренно не углубляемся в математические расчеты и, не претендуя на всеохватность, рассматриваем базисные концепции наиболее востребованных в медицине методов статистического анализа.

Часть 1. Одномерный статистический анализ

1.1. Формирование статистической гипотезы

Статистическая обработка данных является инструментом для обоснования выводов, касающихся интересующей нас популяции (группы лиц, объединенных каким-либо признаком) на основе анализа репрезентативной (представительной) выборки из этой популяции. К примеру, для изучения эффективности какой-либо операции невозможно собрать данные на всех пациентов, когда-либо ей подвергавшихся. Вместо этого подбирается и анализируется репрезентативная выборка. Если выборка обладает достаточной статистической мощностью и анализ выполнен корректно, то полученные выводы могут быть экстраполированы на весь контингент больных, которым данная операция выполнялась. При этом, однако, любой статистический анализ допускает, что обнаруженные (или не обнаруженные) закономерности до известной степени могут оказаться случайными.

Переходя от общей постановки проблемы и дизайна исследования к расчетам, необходимо прежде всего сформулировать статистическую **гипотезу**. Она служит своеобразным связующим звеном между данными и возможностью применения статистических методов анализа, формулируя вероятностный закон разброса данных.

Выдвинутая статистическая гипотеза даёт описание *ожидаемых* результатов исследования, с которыми сравниваются *наблюдаемые*. Если гипотеза верна - наблюдаемое отличается от ожидаемого лишь случайным образом, а именно в соответствии с вероятностным законом этой гипотезы. *Нулевая гипотеза* (обозначается H_o) предполагает отсутствие различий (корреляции, связи) между сравниваемыми выборками. В качестве контрольной выборки чаще всего выступает общепринятый стандарт (метод, подход). Если же нулевая гипотеза отвергается, то принимается *альтернативная гипотеза* (H_a) о наличии различия между группами.

Отличие наблюдаемого от ожидаемого измеряется *вероятностной мерой*. Если отличия между наблюдаемым и ожидаемым настолько велики, что вероятность того, что они являются случайными мала - можно отвергнуть выдвинутую гипотезу как неверную. Обычно она отвергается, если вероятностная мера оказалась меньше или равна заранее установленному *уровню значимости* (см. раздел 1.5.).

Во многих случаях исследователь интуитивно ставит перед собой задачу доказать, что «новый метод лучше старого», т.е. подтвердить альтернативную гипотезу. Это достаточно распространенное заблуждение относительно порядка применения статистических методов.

1.2. Типы данных, их независимость и распределение

Для правильного выбора статистического теста необходимо учитывать характер данных, включаемых в анализ: типы переменных, возможные зависимости между ними и формы их распределений.

Первая попытка классификации переменных в статистике, сохранившая своё значение до настоящего времени, была предпринята в 1946 г. Стэнли Смитом Стивенсом (Stanley Smith Stevens). Схема классификации была основана на типах операций, допустимых для данной переменной. Например, для переменных, обозначающих пол или религию допустимы только сравнения типа равно – не равно, а сравнения типа больше – меньше или арифметические операции не допустимы; как следствие, для этих переменных может быть определена такая статистика, как *мода* (наиболее вероятное значение), и не может быть определено *математическое ожидание* (среднее значение).

В порядке возрастания числа допустимых операций Стивенс ввёл следующие уровни классификации переменных: *номинальный* (nominal), *порядковый* (ordinal) и *непрерывный* (continuous), причём последний делился на подуровни *интервальный* (interval) и *относительный* (ratio).

Дискуссия о «правильной» классификации переменных в статистике продолжается до сих пор. На сегодняшний день общепринятого согласия в этом вопросе не достигнуто, и некоторые статистические компьютерные программы требуют определения типа переменных (например, *PSPP).* Пользователь должен тщательно следить по документации за схемой классификации, использующейся в компьютерной программе, чтобы гарантировать корректный выбор вычисляемых статистик и тестов.

Для простоты мы примем за основу три типа переменных: *непрерывные*, *дискретные* и *категориальные* (номинальные). **Непрерывные переменные** (continuous variables) могут принимать любые численные значения, которые естественным образом упорядоченны на числовой оси (например, рост, вес, АД, РОЭ). **Дискретные переменные** (discrete variables) могут принимать счётное множество упорядоченных значений, которые могут просто обозначать целочисленные данные или ранжировать данные по степени проявления на упорядоченной ранговой шкале (клиническая стадия опухоли, тяжесть состояния пациента). **Категориальные переменные** (categorial variables) являются неупорядоченными и используются для качественной классификации (пол, цвет глаз, место жительства); в частности, они могут быть *бинарными* (дихотомическими) и иметь категорические значения: 1/0, да/нет, имеется/отсутствует.

Форма **плотности распределения** (distribution density) - для непрерывных

переменных, или форма *весовой функции* (probability mass function) - для дискретных переменных, может выражаться эмпирической *гистограммой*, показывая с какой частотой значения переменной попадают в определенные интервалы или принимают определённые значения.

Нормальное (или гауссово) **распределение** имеет колоколообразную форму абсолютно симметричную относительно оси, проходящей через среднее значение (рис. 1) и математически описывается формулой, включающей два параметра, *среднее* и *стандартное отклонение* (см. раздел 1.3.).

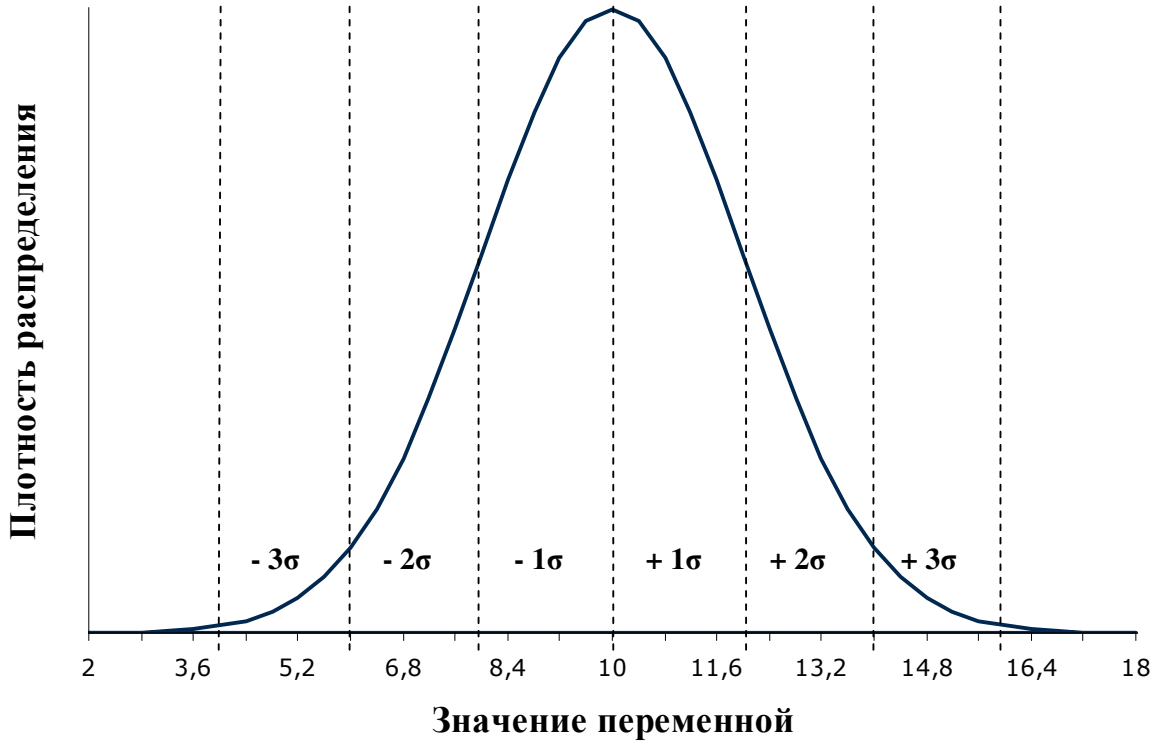

Рис. 1. Плотность нормального распределения

Оценка соответствия распределения данных гауссовому выполняется в статистических программах с помощью критериев нормальности (например, Колмогорова-Смирнова). Визуальная проверка с помощью гистограммы также весьма наглядна. В тех случаях, когда данные не распределены нормально, но подчиняются другому распределению (что может быть определено с помощью статистических программ), приведение к нормальности может быть сделано путем математических операций, например, логарифмирования, извлечения квадратного корня или обращения.

Независимость (англ. independence) данных предполагает, что значения переменных в одной выборке не связаны со значениями переменных в другой, с которой производится сравнение. Примером независимых выборок могут быть показатели

артериального давления (АД) в группе мужчин по сравнению с группой женщин: АД у мужчин не зависит от аналогичного показателя у женщин. Примером зависимых выборок является показатели АД, измеренные у пациентов в 9 часов утра и измеренные у них же в 5 часов вечера. Результаты этих измерений для каждого человека и в целом между выборками скорее всего будут коррелировать, поэтому они считаются парными и оцениваются как зависимые.

1.3. Описательная статистика

Для составления представления о выборке в целом существует ряд показателей, объединяемых понятием «описательная статистика». Каждому исследователю известен такой показатель как **среднее** (mean), который вычисляется путем деления суммы значений переменной на количество значений и характеризует «центральное положение» количественной переменной. Показатель среднего сильно зависит от разброса данных (т.е. наличия экстремально больших и малых значений) и размера выборки. Из-за того, что значения суммируются и делятся на количество случаев (наблюдений), очень высокие или низкие значения переменных (выбросы, англ. outlier) в малых выборках могут существенно влиять на значение среднего. По мере того, как выборка количественно увеличивается в размере, влияние экстремальных значений на среднее снижается.

Медиана (median) – значение, которое занимает среднее положение среди точек данных, разбивая выборку на две равные части. Половина значений переменной лежит по одну сторону значения медианы, и половина – по другую. Очевидно, что выбросы, т.е. экстремальные значения переменной оказывают на медиану гораздо меньшее воздействие, чем на среднее (сами значения, но не их количество). В связи с этим медиану часто используют для описания, например, среднего роста или веса тела в группах.

Стандартное отклонение (standard deviation, SD) отражает изменчивость (разброс, вариацию) значений переменной и оценивает степень их отличия от среднего. Оно рассчитывается на основании вычисленного показателя рассеяния данных, называемого **дисперсией** (variance), путем извлечения из него квадратного корня, в связи с чем в отечественной литературе его также называют «среднеквадратичным отклонением» и обозначают греческим символом σ (сигма). Стандартное отклонение может меняться непредсказуемо, т.е. расти или уменьшаться с увеличением размера выборки, однако обычно не слишком сильно. Наверняка многие исследователи слышали о так называемом «правиле трех сигма». Оно гласит, что практически все наблюдения укладываются в интервал «среднее ± 3σ». Действительно, в интервал «± 3σ» попадают 99,7% наблюдений, ± 2σ включает 95,4%, а ± 1σ – всего 68,3% всех наблюдений. Это правило подходит для

различных распределений, включая нормальное.

Стандартная ошибка (среднего) (англ. standard error, SE, иногда standard error mean, SEM) является оценкой возможного отличия между значением среднего в анализируемой выборке, и истинным средним для всей популяции (которое на самом деле не может быть определено без анализа бесконечно большого числа наблюдений). Стандартная ошибка рассчитывается путем деления стандартного отклонения на квадратный корень из числа наблюдений в выборке и, следовательно, ее значение уменьшается с ростом размера выборки. Это уменьшение является естественным, поскольку чем больше имеется наблюдений, тем больше вероятность, что рассчитанное среднее приближается к истинному.

Доверительный интервал (англ. confidence interval, CI) – диапазон значений, область, в которой с определенным уровнем надежности (или доверия) содержится истинное значение параметра (например, среднего). 90%-ный доверительный интервал означает, что истинное значение величины попадет в рассчитанный интервал с вероятностью 90%. В биомедицинских исследованиях доверительный интервал среднего обычно устанавливается на уровне 95% и определяется как ±1,96 стандартной ошибки (коэффициент 1,96 вытекает из предположения о нормальности распределения значения переменной при условии, что выборка достаточно велика). Для примера, если значение среднего систолического давления в исследованной группе составляет 125 мм рт.ст., а стандартная ошибка 5 мм рт.ст., то при 95% доверительном интервале границы диапазона значений среднего будут 115,2 и 134,8 мм рт.ст. (что составляет ± 9,8 (5 х 1,96) мм рт.ст. в обе стороны от значения среднего). Совмещая значение среднего и доверительный интервал, можно констатировать, что определенное значение систолического АД в группе составляет 125 мм рт.ст., и при этом мы на 95% уверены, что истинное значение находится в интервале между 115,2 и 134,8 мм рт.ст. (в англоязычной литературе описывается как 125.0 [115.2 – 134.8], mean [95%CI]).

У исследователей часто возникает вопрос какие описательные статистические характеристики изучаемой выборки нужно указывать в тексте: среднее или медиану ± стандартное отклонение или стандартную ошибку? Это зависит от того, разброс чего – исходной случайной величины, или оценки её среднего значения (медианы) – изучает исследователь. Если непрерывные переменные распределены нормально (или близко к таковому) и разброс данных обусловлен естественными причинами (люди разного роста, веса и т.п.), то принято указывать среднее ± стандартное отклонение. Если же рассеяние связано с неточностью измерения (например, техническое ограничение или погрешность прибора), то рекомендуется приводить среднее ± (95%) доверительный интервал или стандартная ошибка. Во всяком случае, необходимо указать какие именно характеристики

приведены. Когда непрерывные данные не подчиняются нормальному распределению, для их описания обычно используется медиана и (95%) доверительный интервал. На графиках при этом рекомендуется указать весь интервал значений и обозначить границы 25%, 50% (собственно медиану) и 75% квартилей. Для описания дискретных данных, которые по определению принимают лишь ограниченное число значений и не подчиняются нормальному распределению, используется представление в виде пропорций (процента, доли) или таблиц сопряжения.

1.4. Размер выборки и статистическая мощность

На стадии планирования исследования очень важно определить, какое минимальное число наблюдений необходимо включить в изучаемую группу чтобы результаты тестирования гипотезы оказались правомочными. Для ответа на этот вопрос необходимо понимать что такое **статистическая мощность** и разбираться в сути **ошибок 1 и 2 типа**.

При проверке гипотезы принимается во внимание возможность ошибок измерений, что может стать причиной ложного результата. В зависимости от характера возможного ложного результата, подразделяют ошибки 1 и 2 типа. **Ошибка 1 типа** (обозначается **α**) определяется как вероятность обнаружить различие, тогда как в действительности оно отсутствует («ложноположительный результат»). Другими словами, это вероятность неправомерно отбросить нулевую гипотезу (H_o) в пользу альтернативной (H_a). **Ошибка 2 типа** (обозначается **β**) – это вероятность сделать вывод об отсутствии различия, в то время как фактически оно имеется («ложноотрицательный результат»), т.е. неправомерно принять H_o. В биомедицинских исследованиях предельно допустимый предел ошибки 1 типа обычно устанавливается на уровне 5%, а ошибка 2 типа – не более 20% ($\alpha = 0{,}05$; $\beta \leq 0{,}2$). Ошибка 1 типа рассматривается как более критическая, потому что менее всего хотелось бы неправомерно отвергнуть общепринятую (нулевую) гипотезу. На практике это отражает разумную консервативность, поскольку рекомендация нового метода лечения как более эффективного - в то время как он таковым не является - может нанести больше вреда (например, здоровью пациента, экономический и моральный ущерб), чем отказ от его внедрения (по крайней мере хуже не будет).

Таблица 1. Типы ошибок и статистическая мощность исследования

Результаты проверки гипотезы	Истинный, но неизвестный характер взаимодействия	
	Гипотеза H_o не верна	Гипотеза H_o верна
Отвергнуть гипотезу H_o	Корректное решение (достаточная статистическая мощность)	Ошибка 1 типа (α)
Принять гипотезу H_o	Ошибка 2 типа (β)	Корректное решение

Понимая природу ошибок 1 и 2 типа, можно переходить к оценке мощности исследования. **Статистическая мощность** (statistical power) вычисляется как 1 - β и означает вероятность сделать заключение о наличии различия, в то время как оно имеется на самом деле (т.е. получить «истинно положительный результат»). Таблица 1 показывает взаимосвязь между ошибками 1 и 2 типа и статистической мощностью.

Статистическая мощность напрямую зависит от размера выборки (поскольку связана со стандартной ошибкой, которая в свою очередь уменьшается с увеличением размера выборки), а также от степени различия, которое ожидается обнаружить. Выявление бóльших различий требует меньшего числа наблюдений и, наоборот, для определения небольших различий потребуется более многочисленная выборка. Если планируемая численность выборки не обеспечивает приемлемый уровень статистической мощности (≥ 80%) чтобы убедительно отвергнуть нулевою гипотезу или согласиться с ней, результаты исследования не будут доказательными. Например, если исследователь хочет определить различие в среднем весе тела между двумя группами (получавшими и не получавшими препарат, снижающий аппетит) и хочет доказать разницу в 1 кг при стандартном отклонении 10 кг в контрольной и изучаемой группах, то при α=0,05 и мощности 80% необходимо иметь не менее 1570 людей в каждой группе. Однако, если необходимо оценить различие в 5 кг, достаточно включить в группы по 64 человека.

Расчет размера выборки для желаемого уровня статистической мощности исследования не является сложной процедурой и производится с помощью ряда статистических программных пакетов (например, Statmate). При их использовании нужно обратить внимание на правильную постановку задачи при оценке абсолютных (как в приведенном выше примере), или относительных (например, снижение частоты рецидива в 1,5 раза) изменений.

1.5. Статистическая достоверность

При сравнении групп мы изначально исходим из того, что они не отличаются (это - H_o). Если вероятность того, что выявленные различия являются случайным результатом весьма мала, тогда правомочным будет отвергнуть нулевую гипотезу и заключить, что различие действительно имеется (верна H_a). Показатель **достоверности** различий обозначается **p** (probability, в англоязычной литературе встречается обозначение P или *P*). *Величиной p* (или «пи-величина», англ. «*P*-value») для конкретной выборки называют вероятность получения по крайне мере таких же или ещё больших отличий наблюдаемого от ожидаемого, чем в данной конкретной выборке, при условии, что выдвинутая гипотеза верна. Величина *p* меняется от выборки к выборке, т.е. является случайной величиной на

множестве выборок (причём, с равномерным распределением на интервале 0 - 1).

С помощью статистических расчетов вычисляется значение p, которое затем сравнивается с заранее выбранным *уровнем значимости*, часто обозначаемому греческой буквой α (альфа) (не путать с ошибкой 1-го типа). Обычно в биомедицинских исследованиях уровень значимости устанавливается на уровне $\alpha \leq 0{,}05$ ($\leq 5\%$) . Если выбран уровень значимости $\alpha=0{,}05$, то все выборки, которые для выдвинутой гипотезы возвращают величину $p \leq 0{,}05$, отвергают эту гипотезу, а выборки с величиной $p>0{,}05$ не дают оснований для того, чтобы её отвергнуть. Величину уровня значимости следует понимать в том смысле, что мы задаём, что не более чем в 5% попыток сравнения (какого-либо параметра в разных группах) обнаруженная разница может быть обусловлена чистой случайностью, а не тем, что разница действительно существует. Иными словами, мы задаём вероятность ложного отказа от нулевой (стандартной) гипотезы H_o в пользу альтернативной (изучаемой) гипотезы H_a. В итоге, повторимся, если статистический анализ показывает, что $p \leq 0{,}05$ - правомочным будет заключение о том, что выявленное различие неслучайно и, следовательно, оно является достоверным.

Для демонстрации достоверности различия часто используется наглядный **метод доверительных интервалов**. Напомним, что доверительный интервал устанавливается на уровне ±1,96 стандартной ошибки, в который попадает 95% данных при условии их нормального или близкого к нему распределения. Если доверительный интервал интересующего нас параметра в изучаемой группе «накрывает» значение среднего в группе сравнения, то априори следует вывод о том, что наблюдаемое различие является статистически недостоверным. Если среднее значение параметра в контрольной группе лежит вне доверительного интервала изучаемой группы, то скорее всего различие является достоверным. Среди исследователей бытует представление, что для того, чтобы быть уверенным в наличии разницы по какому-либо параметру между сравниваемыми группами нужно, чтобы «усы ошибок» (границы доверительных интервалов) не пересекались. В определенном смысле это верно: не пересечение «усов» служит гарантией достоверности различия. Однако даже если доверительные интервалы перекрываются, достоверность различий вполне может сохраняться – по крайней мере до тех пор, пока один из «усов» сравниваемых групп не достиг значения среднего другой группы.

1.6. Выбор одномерного статистического теста

Выбор статистического теста является чрезвычайно важной задачей. От его правильности будет зависеть качество анализа и, в конечном итоге, надежность выводов. Выбор теста является задачей нетривиальной, но разбираясь в статистических

характеристиках данных и используя пошаговый алгоритм, исследователь в состоянии осуществить его корректно. Успешное продвижение по алгоритму выбора подходящего статистического метода анализа предполагает знание ответов на следующие вопросы: а) тип данных (непрерывные или дискретные), б) данные зависимые или независимые, в) распределение параметрическое (нормальное) или непараметрическое (отличное от нормального) и г) количество сравниваемых групп.

Заметим, что в зависимости от количества сравниваемых параметров (переменных) различают одномерную (univariate) и многомерную (multivariate) статистику. Одномерная статистика применяется при анализе двух и более групп с целью сравнения лишь одной переменной. Многомерная статистика используется для анализа двух и более групп, но с учетом одновременного изменения двух или более переменных. В данной части работы приведены методы одномерной статистики, многомерная статистика рассматривается во второй части.

Еще на стадии планирования анализа полученных результатов нужно определить, какая статистика будет использоваться, одномерная или многомерная. При этом, даже если планируется использование многомерных методов, сперва все равно необходимо использовать описательную статистику и провести одномерный анализ. Это позволяет лучше ориентироваться в наборе данных и сформировать первичное представление о соотношениях различных переменных в сравниваемых группах.

На Рис. 2 показана блок-схема выбора методов одномерного статистического анализа, а ниже кратко обсуждаются области применения основных из них.

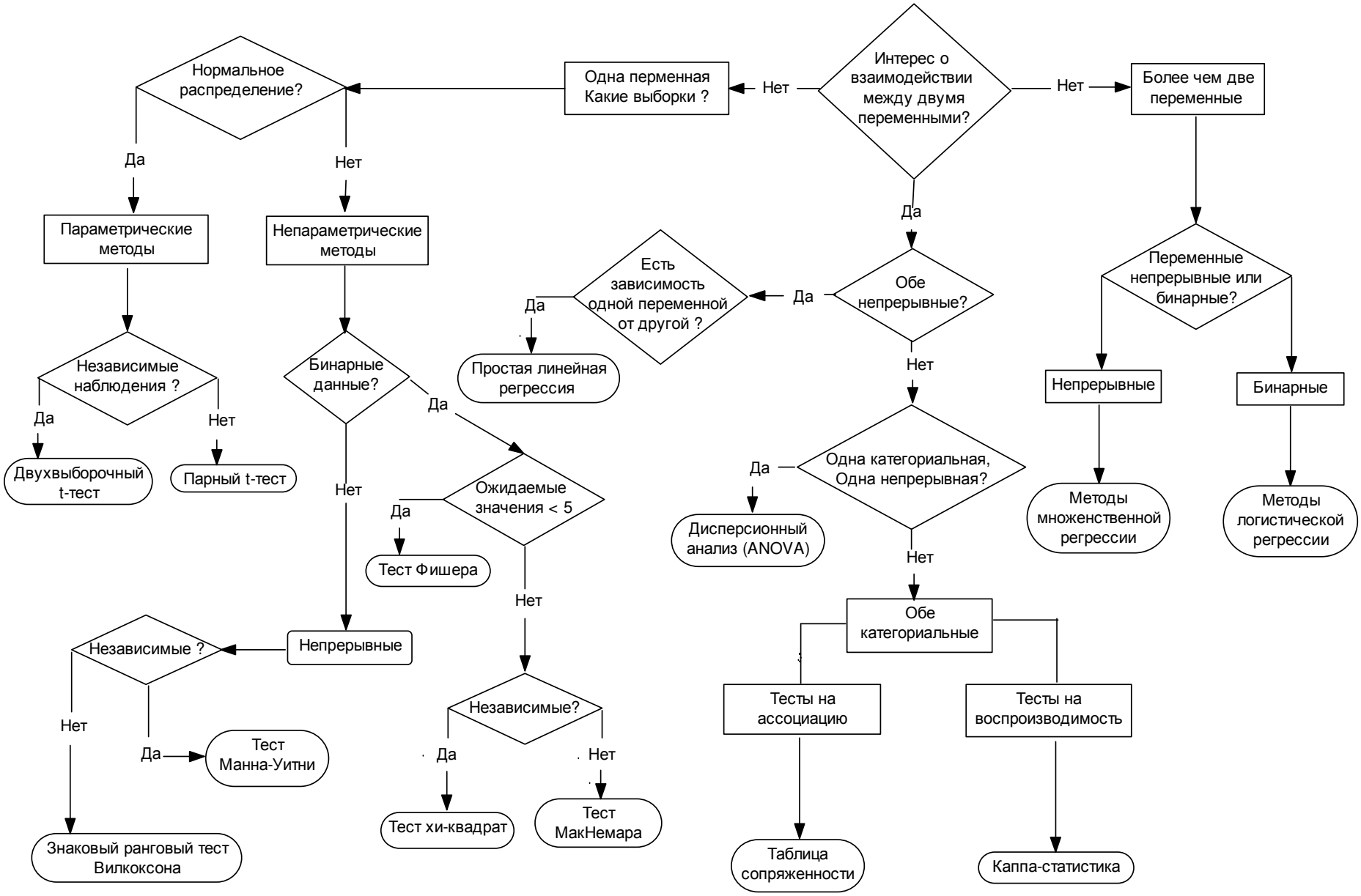

Рис 2. Алгоритм выбора метода одномерного статистического анализа.

1.6.1. Параметрическая статистика

Параметрическая статистика используется для анализа непрерывных (численных) переменных, значения которых распределены нормально. Наиболее часто используется так называемый **непарный *t*-тест** (распространенное название - «тест Стьюдента»; *t*-test), с помощью которого возможно провести проверку гипотезы (H_o) об отсутствии различия средних значений переменной в двух независимых выборках, исходя из предположения об одинаковости стандартного отклонения в них.

Если данные являются зависимыми (например, получены в процессе повторных наблюдений за одним и тем же пациентом (repeated measurements) или используются показатели пациентов, подобранных в пары (по возрасту или полу), рекомендуется **парный** (paired) ***t*-тест**.

Распространённой ошибкой является применение t-тестов к показателям состояния пациентов (пациента) до и после проведения двух разных методов лечения (H_0 – методы не различаются или лечение не действует) без проверки равенства стандартных отклонений показателей. При неуверенности в одинаковых дисперсиях (стандартных отклонениях) выборок используется модифицированный ***t*-тест Уэлча** (Welch's t-test), но он применим только к независимым выборкам (непарный тест).

Различают t-тесты односторонние и двусторонние. Термин **двусторонний** (двунаправленный, англ. two-tailed) означает, что поиск различий будет производиться в обе стороны: для увеличения показателей и для их уменьшения. В биомедицинских исследованиях рекомендуется применять двухсторонние тесты, так как чаще всего неизвестно, будет ли знак отличия положительным или отрицательным.

Для сравнения независимой переменной в более чем двух выборках может выполняться **дисперсионный анализ** (**AN**alysis **O**f **Va**riance, ANOVA). К примеру, его можно применить для выявления разницы среднего систолического АД в различных возрастных группах. Для зависимых данных, оцениваемым в более чем двух группах используется **дисперсионный анализ с повторным измерением** (Repeated-Measures ANOVA, RM-ANOVA).

1.6.2. Непараметрическая статистика

Непараметрические методы анализа применяются как к непрерывным, так и к дискретным данным.

1.6.2.1. Непрерывные переменные

***U* тест Манна-Уитни** (Mann-Whitney *U*), также известный как тест **Вилкоксона ранговых сумм** (Wilcoxon Rank Sum) или тест **Манна-Уитни-Вилкоксона (MWW)**

проверяет, являются ли две сравниваемые группы выборками из одного и того же распределения, используя в качестве статистики (U) медиану всевозможных разностей между элементами одной и второй выборки. По этой причине на результат практически не влияют редкие экстремальные значения. Для ранговых шкал, когда t-тест не применим, MWW-тест остаётся логичным выбором. Проблемы с интерпретацией теста, как и в случае t-тестов, возникают, когда распределения для двух выборок отличаются по форме, например, имеют сильно отличающиеся дисперсии.

Для иллюстрации важности адекватного выбора статистического теста предположим, что исследователь сравнивает вес тела в двух независимых группах пациентов. В первой группе, помимо людей с «нормальным» весом, имеется два полных человека; средний вес в группе составил 100,3 кг, а медиана – 75,1 кг. Во второй группе, напротив, есть несколько худощавых людей; средний вес группы – 60,8 кг, медиана - 72,5 кг. Известно, что в обеих группах распределение отклоняется от нормального, т.е. выборки не проходят тест на нормальность распределения данных. При сравнении средних показателей (100,3 кг и 60,8 кг) может создаться впечатление, что группы существенно отличаются и вполне возможно, что t-статистика выявит достоверность различий. Однако сравнение средних было бы оправдано в том случае, если распределение переменной веса тела в обеих группах являлось нормальным. Но оно таковым не является, поэтому следует использовать непараметрическую статистику. Тест MWW обнаружит очень схожие медианы (75,1 и 72,5 кг) в группах сравнения и, скорее всего, будет сделан вывод об отсутствии отличия между группами.

При сравнении переменной более чем в двух *независимых* группах, непараметрическим аналогом дисперсионного анализа является **тест Крускала-Уоллиса** (Kruskal-Wallis), в котором данные заменены их рангами и сравниваются медианы выборок. Нормальность распределений не требуется, но они должны быть похожей формы и иметь сравнимые по величине дисперсии.

Если данные не распределены нормально, являются непрерывными и *зависимыми* (парными), может быть рекомендован тест **знаковых рангов Вилкоксона** (Wilcoxon signed-rank). Принцип метода заключается в вычислении разницы между парными данными с последовательным ранжированием по положительному или отрицательному значению разницы и определением критического (порогового) значения для опровержения нулевой гипотезы.

1.6.2.2. Дискретные переменные

Для *независимых* категориальных, в частности, бинарных данных обычно

используются методы таблиц сопряжения (англ. contingency tables). Сравнительный анализ проводится чаще всего с помощью **точного теста Фишера** (англ. Fisher's exact test) или **хи-квадрат (χ^2) теста** (англ. chi-square test; или «хи-квадрат Пирсона», англ. Pearson's chi-square).

Хи-квадрат тест может быть применен к таблицам практически любой размерности. В некоторых статистических программах реализовано продолжение точного теста Фишера для таблиц сопряжения размерностью большей, чем 2 X 2 (точный тест Фишера изначально разработан для таблиц сопряжения размерностью 2 X 2), однако многие исследователи традиционно предпочитают статистику хи-квадрат, что в принципе правомерно. Отметим, что последняя не может использоваться, если ожидаемое (но не наблюдаемое) значение признака в какой-либо ячейке таблицы менее 5.

Точный тест Фишера и хи-квадрат тест основываются на принципиально разной идеологии расчета. Точный тест Фишера использует перебор вариантов заполнения таблицы сопряженности (перестановочный тест), в то время как хи-квадрат нацелен на сравнение наблюдаемой и ожидаемой частоты появления признака. Их общее назначение состоит в проверке значимости связи между двумя категориальными переменными, но при разных выборочных схемах (например, при разных дизайнах исследования).

Какой тест более предпочтителен для расчетов? Для таблиц сопряжения размерностью 2 X 2 предпочтителен точный тест Фишера, поскольку он дает более точную оценку, чем хи-квадрат тест. Однако применение хи-квадрат теста как для таблиц 2 X 2, так и для таблиц большей размерности, также является правомерным. Выбор остается за исследователем, необходимо всегда указывать какой из методов использовался.

В большинстве случаев оценки значимости различия (т.е. значения p), полученные с помощью этих двух разных тестов для одной и той же таблицы сопряжения, не совпадают. Вместе с тем и точный тест Фишера, и хи-квадрат тест, как правило, непротиворечиво выдают значение p, которое будет либо больше, либо меньше установленного порогового уровня значимости, например на уровне 0,05.

Таблица 2. Таблица сопряжения непарных дискретных данных

Воздействие фактора (применение препарата)	Эффект имеется (наличие побочного эффекта)	Эффект отсутствует (нет побочного эффекта)	Итого
Да (пациенты)	А (45)	Б (75)	А+Б (120)
Нет (контрольная группа)	В (55)	Г (85)	В+Г (140)
Итого	А+В (100)	Б+Г (160)	Ч (260)

Пример данных, организованных в таблицу сопряжения размерностью 2 X 2, приведен в таблице 3. В ней рассматривается абстрактная ситуация возникновения

побочного эффекта (например, тахикардии) после применения какого-либо препарата.

Расчеты, проведенные с помощью точного теста Фишера и хи-квадрат теста в рассматриваемом случае возвращают значения p равные 0,80 и 0,87, соответственно. Это говорит о том, что связь побочного эффекта с применением данного препарата недостоверна.

Из таблицы сопряжения также можно рассчитать еще один важный статистический показатель. Он называется **отношение шансов** (англ. odds ratio, **OR**) и вычисляется как (А*Г)/ (Б*В). Отношение шансов используется, чтобы оценить насколько велики шансы положительных и отрицательных исходов (например, развитие нежелательного побочного эффекта после применения препарата, как показано в примере выше). Если OR=1 (или очень близко к 1), то это означает, что шансы события в обеих группах практически совпадают.

Для данных, приведенных в таблице 3, отношение шансов составляет 0,93, а 95% доверительный интервал – от 0,56 до 1,53. В англоязычной литературе показатель часто записывается в виде: 0,93 [0.56-1.53] (т.е. OR [95% CI]). Из значения отношения шансов (0,93), которое меньше 1, можно составить представление о том, что побочный эффект в группе, принимавшей препарат, наблюдался несколько реже, чем в контрольной группе (соответственно 60% и 65%). Однако поскольку доверительный интервал включает значение 1, различие недостоверно.

Если категориальные данные являются *зависимыми*, используется **тест МакНемара** (McNemar test), который представляет собой модификацию хи-квадрат теста для парных или соотнесенных данных. Примером уместного использования теста МакНемара было бы сравнение доли пациентов, ответивших на лечение по какому-то показателю, когда сравнение проводится до и после лечения у одних и тех же людей. Тест МакНемара часто используется в исследованиях типа «случай-контроль» (case-control study), в которых каждому случаю противопоставляется конкретный контроль. Для расчетов с помощью теста МакНемара составляется таблица сопряжения, подобная Таблице 3, однако в каждой ячейке указывается не количество лиц, соответствующих какому-либо исходу, а количество пар (до/после лечения, случай/контроль).

1.6.2.3. Преимущества и недостатки непараметрических методов

К преимуществам непараметрических методов можно отнести следующие:

- могут быть использованы, когда характеристики популяции, из которой делается выборка, частично неизвестны;
- бóльшая мощность (робастность);
- относительная несложность вычислений (в большинстве случаев);
- менее жесткие начальные допущения;

Недостатками являются:

- меньшая эффективность, чем у параметрических методов;
- меньшая специфичность;
- потенциальная трудоемкость при применении к большим массивам данных.

1.7. Корреляционный и регрессионный анализ

На практике часто возникают задачи, когда нужно проверить взаимосвязь между какими-либо *непрерывными данными*, например, между АД и весом тела. В этих случаях используются корреляционный и регрессионный анализ. Корреляционный анализ определяет характер взаимосвязи переменных (прямой или обратный), а регрессионный - форму зависимости (насколько сильно изменяется переменная в ответ на изменение другой).

1.7.1. Корреляционный анализ

Корреляционный анализ является методом оценки линейных связей (общей пропорциональности) между переменными, т.е. насколько *согласовано* они меняются. В англоязычной литературе часто употребляется термин «линейная корреляция Пирсона». **Корреляция** Пирсона (обычно просто «корреляция») между переменными может быть положительной, отрицательной или вовсе отсутствовать.

Две переменные коррелируют положительно, если бóльшие значения одной переменной имеют тенденцию к ассоциации с бóльшими значениями другой переменной, как показано на Рис. 3.

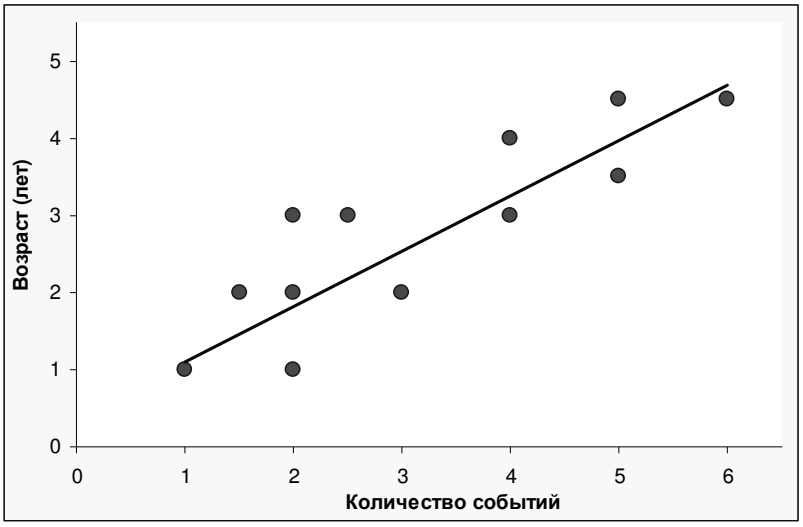

Рис 3. Пример положительной корреляции.

Напротив, если бóльшие значения одной переменной ассоциированы с меньшими значениями другой переменной, говорят об отрицательной корреляции, как показано на Рис. 4.

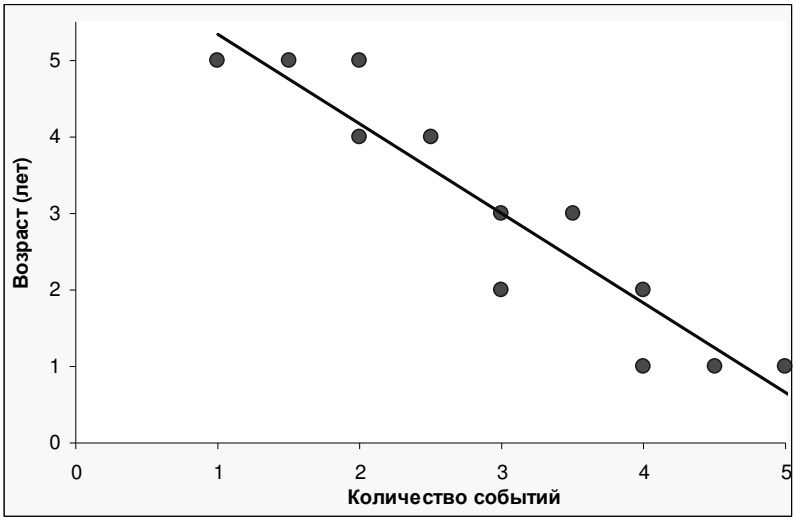

Рис 4. Пример отрицательной корреляции.

При отсутствии корреляции нет никакой закономерности взаимосвязи одних показателей с другими, как показано на Рис. 5.

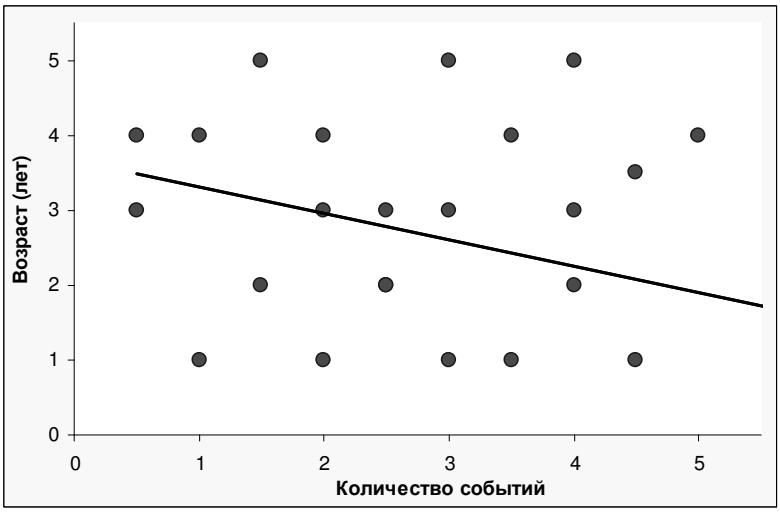

Рис. 5. Пример отсутствия корреляции.

Показателем согласованности между значениями двух переменных является **коэффициент корреляции** (correlation coefficient). Этот коэффициент является количественным, обозначается *r* (Pearson *r*), и имеет область значений от - 1 до + 1.

r = 1 означает максимально сильную положительную линейную взаимосвязь между X и Y;

r = - 1 означает максимальную отрицательную линейную взаимосвязь между X и Y;

r = 0 означает отсутствие линейной взаимосвязи между X и Y.

Для оценки того, насколько сильно линейно связаны две переменные, рекомендуется использовать **коэффициент детерминации**, который представляет собой квадрат

коэффициента корреляции Пирсона (r^2). Очевидно, что чем больше коэффициент корреляции отклоняется от 1 или - 1 (т.е. чем больше степень рассеяния точек от линии на Рис. 3 - 5), тем меньше будет значение коэффициента детерминации и тем слабее будут две переменные коррелировать между собой.

Заметим, что корреляция Пирсона основывается на предположении о том, что значения переменных распределены нормально или близко к нормальному. Если распределение значений отличается от нормального или в силу каких-то причин это невозможно оценить, то можно воспользоваться непараметрической корреляцией Спирмана, с помощью которой также можно рассчитать коэффициент корреляции *r* (англ. Spearman *r*). Статистические программы также оценивают достоверность (значение p) отличия коэффициента *r* от 0, т.е. является ли оценка корреляции достоверной. Если выборки достаточно велики (приближаются к 100 наблюдениям), форма распределения не оказывает большого воздействия на результат корреляционного анализа. Выполняется ли он с использованием стандартного (корреляция Пирсона) или непараметрического (корреляция Спирмана) метода – уже не играет большого значения.

Необходимо иметь ввиду, что наличие в выборке выбросов может сильно повысить или понизить коэффициент корреляции. Выбросы несложно обнаружить при визуализации данных на простом графике X-Y. Они представляют собой точки, далеко выступающие по одной или по обеим координатам от основного кластера, если таковой имеется. К выбросам следует относиться осторожно: они могут как обоснованно, так и необоснованно поддерживать или нарушать общую тенденцию («случайность – это непознанная закономерность»). Во всяком случае, каждый выброс рекомендуется проверить на предмет правильности записи исходных данных и исключить возможность случайной ошибки.

1.7.2. Линейный регрессионный анализ

Линейная регрессия и линейная корреляция – сходные, но не идентичные методы анализа. С помощью **линейного регрессионного анализа** определяются параметры прямой, которая наилучшим способом предсказывает значение одной переменной на основании значения другой согласно формуле:

$y = a + bx$,

где y - значение одной переменной, a – точка пересечения прямой с осью ординат (вертикальная ось, ось Y), b задает наклон линии, а x – значение другой переменной.

Линейный регрессионный анализ проводится, если корреляционный анализ выявил взаимосвязь между переменными.

Статистические программы, помимо коэффициента корреляции *r*, коэффициента

детерминации r^2, коэффициентов a и b регрессионной прямой, рассчитывают также достоверность (значение p) отклонения наклона регрессионной прямой от 0, что также является оценкой наличия значимой корреляции между двумя переменными. Некоторые программы дополнительно оценивают вероятность того, что данные отклоняются от линейного взаимоотношения. В случае, если достоверность такого отклонения оказывается высокой (т.е. получено малое значение p для этого параметра), необходимо отказаться от линейного регрессионного анализа «сырых данных» и подумать над возможностью приведения их к линейности путем преобразования (например, извлечение квадратного корня, возведения в степень, логарифмирования или описания более сложной функцией). После этого в ряде случаев линейный регрессионный анализ становится вновь возможным.

1.8. Чувствительность, специфичность и точность

Способом оценить информативность и разрешающую способность диагностического метода является оценка его *чувствительности*, *специфичности* и *точности*. Эти показатели отражают шансы поставить правильный диагноз заболевания у больных и здоровых людей. Их сравнивают с аналогичными показателями общепринятого («золотого») стандарта диагностического теста.

Чувствительность определяется как доля пациентов действительно имеющих заболевание среди тех, у кого тест был положительным. **Специфичность** определяется как доля людей, не имеющих заболевания среди всех, у кого тест оказался отрицательным. **Точность** показывает долю «правильных срабатываний теста» среди всех обследованных и является совокупным показателем информативности теста. Таблица сопряжения для проведения расчетов представлена в Таблице 3. По существу она отражает соотношение между ошибками 1 и 2 типа (см. раздел 1.4.).

Таблица 3. Организация данных для оценки информативности диагностического теста

Результат теста	Общепринятый («золотой») стандарт	
	Положительный	Отрицательный
Положительный	**а** – число имеющих заболевание и положительный результат теста	**б** – Число не имеющих заболевания и положительный результат теста
Отрицательный	**в** – число имеющих заболевание, но отрицательный результат теста	**г** – Число не имеющих заболевания и отрицательный результат теста
Итого	**а + в** = Общее число имеющих заболевание	**б + г** = Общее число не имеющих заболевания

Чувствительность = а / (а+в)
Специфичность = г / (г+б)
Точность = (**а+г**) / (**а+б+в+г**)

а - истинно положительный результат;
б - ложноположительный результат;
в - ложноотрицательный результат;
г - истинно отрицательный результат.

Высокочувствительный диагностический тест – тот, который дает наибольшее число положительных результатов при фактическом наличии заболевания. С клинической точки зрения нужно понимать, что высокочувствительный тест может отличаться гипердиагностикой, зато позволяет минимизировать риск пропустить заболевание. Это важно, например, при выявлении инфицированных людей при скрининге опасного инфекционного заболевания ввиду угрозы эпидемии. С другой стороны, высокоспецифичный тест дает отрицательные результаты при фактическом отсутствии заболевания с бóльшей вероятностью. К примеру, это важно в случаях, когда дорогостоящее лечение связано с серьезными побочными эффектами и, следовательно, гипердиагностика крайне нежелательна.

Исходя из значений чувствительности и специфичности, рекомендуется построение **характеристической кривой** (ROC-кривая; англ. **R**eceiver **O**perating **C**haracteristic (ROC) curve), которая показывает зависимость количества верно диагностированных положительных случаев от количества неверно диагностированных отрицательных случаев (ось X=специфичность, ось Y=чувствительность). Идеальный диагностический тест должен иметь Г-образную форму характеристической кривой и проходящей через верхний левый угол, в котором доля истинно положительных случаев 100% (или 1), а доля ложноположительных случаев равна 0. Чем ближе проходит характеристическая кривая к значению (0;1) (идеальная чувствительность), тем выше эффективность теста. Наоборот, чем меньше кривая напоминает форму буквы «Г», т.е. чем ближе она проходит к диагонали графика ("бесполезный тест"), тем эффективность теста меньше (см. Рис. 6.)

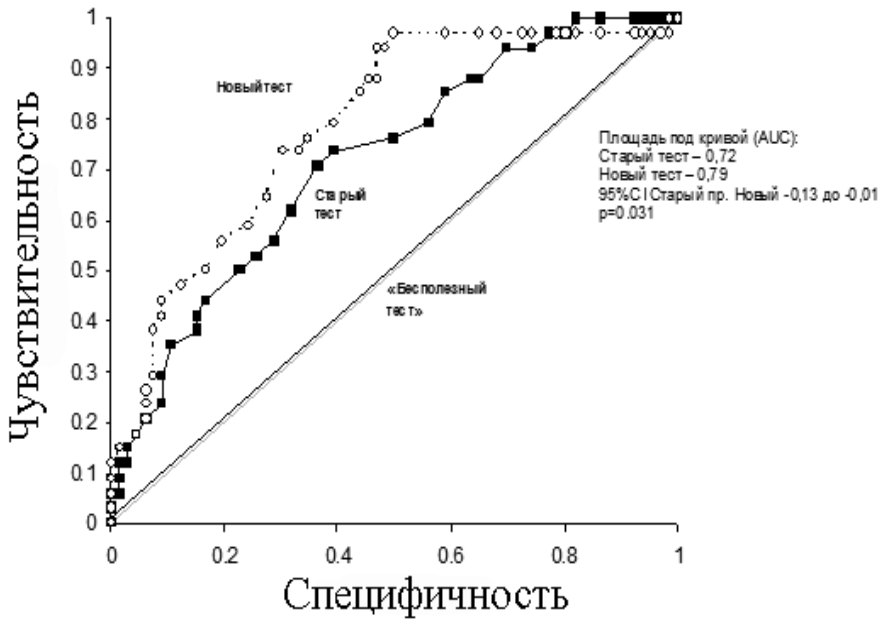

Рис. 6. Пример *характеристических (ROC) кривых* двух тестов, «старого» и «нового».

Количественную оценку характеристической кривой можно провести, рассчитав площадь под ней (англ. **A**rea **U**nder **C**urve, AUC). Приблизительная шкала значений AUC, отражающая качество диагностического теста такова:

AUC=0,9-1,0 – отличное качество

AUC=0,8-0,9 – высокое качество

AUC=0,7-0,8– хорошее качество

AUC=0,6-0,7– среднее качество

AUC=0,5-0,6– плохое (неудовлетворительное) качество.

Для того, чтобы новый диагностический метод заслужил признание, он должен продемонстрировать более высокие, чем «золотой» стандарт, значения чувствительности и специфичности.

Алгоритм построения характеристических кривых реализован во многих статистических программах, в ИНТЕРНЕТе имеется большой выбор он-лайн ROC-калькуляторов. На Рисунке 6 для примера показаны реальные расчетные характеристические кривые. Многие статпрограммы способны генерировать сглаженные кривые и возвращать необходимые статистические оценки. В рассмотренном примере «новый» тест имеет достоверно лучшие характеристики по сравнению со «старым».

Вышеизложенные методы описательной и одномерной статистики являются базовыми, с них рекомендуется начинать статистический анализ. Самостоятельное выполнение этих процедур вполне по силам исследователю, не имеющего специальной подготовки в математической статистике. С их помощью осуществляется первичная обработка и одномерный анализ имеющихся данных.

Часть 2. Анализ выживаемости и многомерная статистика

2.1. Методы анализа выживаемости

Под методами оценки выживаемости (survival) понимается изучение закономерности появления ожидаемого события у представителей наблюдаемой выборки во времени. Таким событием не обязательно является летальный исход, как можно предположить из названия анализа. Им может быть рецидив заболевания или, наоборот, выздоровление, в общем случае – происхождение определенного события. Точкой отсчета может быть дата (час) выполнения процедуры, назначения лекарственного препарата, возраст на момент диагноза и т.п. Период времени от начального события (например, постановки диагноза) до итогового (летальный

исход, рецидив, выздоровление) называется **временем до события** (time to event) или временем ожидания.

Исходно термин «выживаемость» заимствован из лексикона страховых компаний, использующих его в статистических расчетах при страховании жизни своих клиентов. С помощью определенной методики компания оценивает потенциальный риск летального исхода (страховой случай) или среднее время выживания (выживаемость, время до события) клиента с учетом сопутствующих рисков, что и определяет размер индивидуальных страховых взносов.

Исходя из общей постановки задачи, т.е. анализа среднего времени выживания и проведения на его основе, например, оценки эффективности нового метода лечения, казалось бы, можно воспользоваться параметрическими и непараметрическими статистическими методами, описанными в разделе 6 первой части обзора. В принципе это возможно, но анализ выживаемости имеет важное отличие в способе построения выборки. В то время как для рассмотренных ранее статистических методов объем и структура выборки являются постоянными, в анализе времени до события они могут меняться. Проблема заключается в том, что время до события не обязательно может быть определено для всех пациентов выборки в ходе запланированного срока наблюдения. Значение этого показателя становится определенным только среди тех лиц, у которых произошло интересующее событие. Для всех остальных объектов наблюдения показатель остается неизвестным до наступления события, которое может вообще не произойти за период наблюдения. Кроме того, пациенты могут выбывать из исследования в силу разных обстоятельств (смена места жительства и т.п.), включаться в исследование в его середине или в конце, а также ожидаемое событие может быть вызвано иной причиной (например, летальный исход не от заболевания, а в результате несчастного случая). Все это приводит к (нерегулярным) качественным и количественным изменениям в анализируемых данных, и определяет необходимость применения специальных методов, в которых можно было бы учесть и использовать неполную

информацию.

Данные, которые содержат неполную информацию, называют **цензурированными** (censored). С такими выборками приходится иметь дело, когда наблюдаемый параметр является временем до наступления события, а период наблюдения ограничен (например, у пациента рецидив заболевания не обнаружен за 6 месяцев до того, как он переехал в другой город и дальнейшая информация о нем недоступна). При анализе выживаемости, как и при других методах статистического анализа, вся информация о выборке содержится в соответствующей ей функции распределения вероятности (в данном случае – времён ожидания), но используется она не в виде плотности распределения ветоятности значений, а в виде **функции выживания** (survival function). Кумулятивная функция распределения F(t) времён ожидания отражает вероятность того, что время ожидания собятия меньше t. Соответственно, функция выживания S(t) = 1 – F(t) равна вероятности того, что событие не состоится ранее, чем по истечении времени t.

Наиболее распространенными описательными методами исследования цензурированных данных являются построение **таблиц дожития** (mortality table) и **метод Каплана-Мейера** (Kaplan-Meier method). Для анализа используют несколько подходов, из которых мы остановимся на **лог-ранк тесте** (логарифмический ранговый тест; англ. log-rank test) и **модели пропорциональных интенсивностей Кокса** (или **модель пропорциональных рисков Кокса**; англ. Cox Proportional Hazards Model).

2.1.1. Таблицы дожития

Таблицы дожития – один из наиболее традиционных методов исследования данных о выживаемости (происхождение интересующего нас события). В таблицах дожития время наступления события разбивается на интервалы, для каждого из которых определяется число и доля объектов: а) у которых событие не произошло на момент начала данного интервала времени, б) у которых событие произошло в течение данного интервала и в) которые были изъяты или цензурированы на данном интервале. По существу таблица дожития является

расширенной таблицей частот. Считается, что для получения надежных оценок основных показателей (функции выживания, плотности вероятности и интенсивности, см. ниже) размер группы должен быть не менее 30.

На основании таблицы рассчитывается ряд индикаторов. **Число изучаемых объектов** – число объектов, у которых событие не произошло на момент начала данного интервала времени минус половина числа объектов, которые были изъяты или цензурированы**. Доля «умерших» -** отношение числа объектов, у которых событие произошло в течение данного интервала, к числу изучаемых объектов на данном временном интервале. **Доля выживших -** единица минус доля «умерших». **Функция выживания** (выживаемость) – кумулятивная доля объектов, событие у которых не произошло на момент начала определенного интервала времени; ее рассчитывают как произведение долей выживших на всех предыдущих интервалах. **Плотность вероятности** – оценка вероятности наступления события в каком-либо интервале; рассчитывается как отношение разности между значениями функции выживания на любом данном и последующем интервале к продолжительности данного интервала времени. **Функция интенсивности** представляет собой вероятность того, что на данном интервале произойдет событие у того объекта, у которого оно еще не произошло на момент начала этого интервала; вычисляется как отношение числа событий, происшедших в течение данного интервала, к числу объектов, у которых событие не произошло до момента времени, находящегося в середине этого интервала. **Медиана ожидаемого времени жизни** – точка на оси времени, в которой значение функции выживания равна 0,5; медиана ожидаемого времени жизни совпадает с точкой выживания 50% наблюдений только в том случае, если до этого момента времени цензурированных наблюдений не было. Аналогично через значения функции выживания можно определить и квартили (25-й и 75-й процентили) ожидаемого времени жизни.

2.1.2. Метод Каплана-Мейера

Метод Каплана-Мейера используется для оценки доли объектов наблюдения

(пациентов), у которых событие не произошло (функция выживания, выживаемость) для любого момента времени в течение всего периода наблюдения. Поскольку разбиение данных по временным интервалам (группировка) не производится, суть метода Каплана-Мейера несколько отличается от таблиц дожития. В то же самое время результаты, получаемые с помощью этих двух методов, принципиально близки по смыслу. Оценка функции выживания в методе Каплана-Мейера представляет собой произведение выживаемости в данный момент времени на выживаемость в следующий момент времени, когда событие произошло.

Как и таблицы дожития, метод Каплана-Мейера полностью применим к цензурированным данным. Для расчетов используется истинное количество объектов, у которых событие ещё не произошло в любой момент времени, для которого производится оценка. Отметим, что цензурированность данных может оказывать влияние на оценку функции выживаемости, в связи с чем метод Каплана-Мейера использует следующие предположения: а) цензурированные объекты («выбывшие») имеют те же самые показатели выживаемости, как и те, которые продолжают наблюдаться (т.е. цензурирование не влияет на прогноз выживаемости); б) оценки выживаемости одинаковы для объектов, включенных в исследование на более ранних или более поздних сроках; в) событие происходит именно в анализируемый момент времени. Последнее предположение может искусственно завысить оценку выживаемости, если измерения производятся редко, так как определение момента времени наступления события откладывается до следующего обследования.

Метод Каплана-Мейера широко используются в клинических испытаниях, например, с целью оценки эффективности нового лекарственного препарата в изучаемой группе по сравнению с контрольной (получающей плацебо) группой. Предположим, мы хотим определить долю пациентов, у которых через 2 недели после начала применения нового препарата уровень холестерина в крови не понизился (как ожидалось) до определенного значения; определение холестерина производится один раз в неделю. Допустим, что в течение первой недели эффект наблюдался у 5 из 50 пациентов (оценка

функции выживания 0,90; рассчитывается как (50-5)/50). Из 45 пациентов, у которых снижение уровня холестерина не произошло к моменту начала второй недели, в течение нее эффект был отмечен у 9. «Выживаемость» в течение второй недели составила, таким образом, 36/45=0,80. Общая двухнедельная выживаемость (доля пациентов, у которых эффект не наблюдался) в этом случае была 0,90x0,80=0,72. Поскольку для вычислений используется операция умножения, метод Каплана-Мейера называют также множительной оценкой.

Графическое представление метода Каплана-Мейера заключается в построении кривой выживаемости, отражающей пропорцию пациентов, у которых ожидаемое событие не произошло к определенному моменту времени. Временные интервалы определяются либо периодичностью контрольных обследований или временем до события в реальном масштабе (если известен момент происхождения события). Когда у объекта наблюдения происходит ожидаемое событие, производится перерасчет пропорции оставшихся в исследовании объектов, у которых событие не произошло, что отображается «ступенькой» вниз на кривой, как показано на рисунке 1.

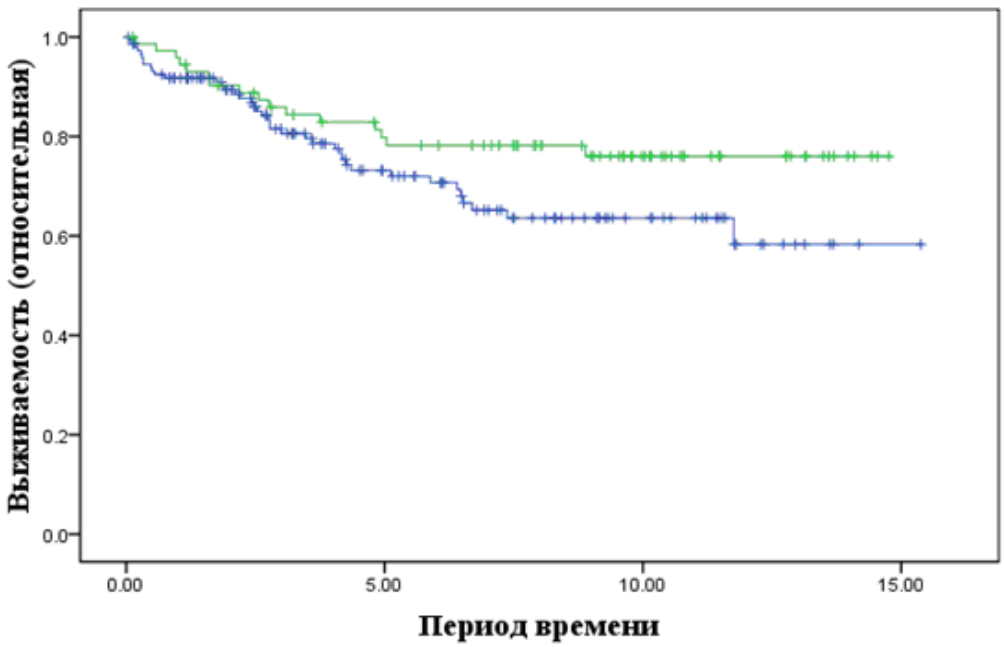

Рис. 7. Метод Каплана-Мейера (пример). Вертикальные штрихи отображают цензурированные наблюдения, «ступеньки» вниз – происхождение события.

Кривые, построенные с помощью метода Каплана-Мейера, часто используются для оценки собственно выживаемости или безрецидивной выживаемости онкологических больных.

Бесспорное преимущество метода состоит в том, что он не требует знания о предполагаемой форме кривой выживаемости или характера распределения показателей выживаемости во времени. С другой стороны, будучи описательным средством, метод Каплана-Мейера имеет тот недостаток, что он не позволяет сравнить выживаемость между группами, т.е. оценить достоверность различий кривых выживаемости.

2.1.3. Лог-ранк тест

С помощью **лог-ранк теста** (логарифмического рангового теста) можно оценить общую выживаемость в двух и более группах за весь период наблюдения, что является важным отличием от умозрительного сравнения показателей выживаемости в любой момент времени. Лог-ранк тест принимает за нулевую гипотезу то, что выживаемость в сравниваемых группах пациентов не различается. Для всего периода наблюдения определяются ожидаемые и фактические показатели выживаемости для всех моментов времени происхожения события. Дальнейшие вычисления (сравнение ожидаемых и фактических значений) производятся с помощью теста хи-квадрат с целью выявления достоверности (без оценки ее степени и без доверительных интервалов) различий. Ряд статистических программ имеют специальные модули для выполнения лог-ранк теста.

Лог-ранк тест, как и тест Каплан-Мейера, применим к цензурированным выборкам и основывается на тех же самых предположениях о влиянии цензурированности данных на результат. С помощью лог-ранк теста различия выживаемости в группах обнаружить легче если риск возникновения события в одной группе существенно и последовательно выше, чем в другой. Если кривые выживаемости вдруг пересекаются (например, при сравнении результатов хирургического лечения и тактики пассивного наблюдения прогрессирующей коронарной окклюзии), лог-ранк тест вообще не способен выявить различие. В связи с этим

при выполнении анализа с помощью этого теста необходимо вначале представлить кривые выживания на графике.

2.1.4. Модель пропорциональных интенсивностей Кокса

Модель Кокса (Cox Proportional Hazards Model), часто называемая в литературе «Пропорциональная модель Кокса», является наиболее используемым в современных публикациях и рекомендуемым инструментом анализа данных выживаемости. В ее основе лежит метод множественной регрессии (см. раздел 2.2.), и в качестве выходного параметра модель возвращает значение отношения рисков и его доверительный интервал. **Отношение рисков** (hazard ratio, **HR**) - это оценка отношения интенсивностей (показателей, уровней, функции) риска в экспериментальной и контрольной группах, рассчитанные для любого момента времени наблюдения. Модель предполагает, что отношение рисков у членов экспериментальной и контрольной групп остаются неизменными в течение всего периода наблюдения (**предположение о пропорциональности**, англ. proportionality assumption). **Интенсивность риска** представляет собой вероятность того, что событие, не произошедшее к определенному моменту времени, случится в следующий интервал времени, отнесённую к продолжительности этого интервала. Временной интервал может быть установлен очень коротким, поэтому оценку можно делать для любого момента времени. Говоря другими словами и применительно к клиническому испытанию, в котором ожидаемым результатом является, например, выздоровление пациента, отношение рисков отражает относительную вероятность быстрейшего выздоровления у больных, получающих лечение, по отношению к пациентам контольной группы для любого момента времени.

Данная модель позволяет включать в исследование всех интересующих нас пациентов, невзирая на цензурирование (частичную неполноту данных), поскольку модель использует базисное допущение о том, что выбывание пациентов происходит случайным образом и с одинаковой вероятностью как в изучаемой, так и в контрольной группе. Кроме того, изначально предполагается что пациенты, у которых произойдет или не произойдет

событие, выбывают из исследования с одинаковой вероятностью (правила пропорциональности модели).

Пропорциональная модель Кокса в последнее время получает все наибольшее признание и популярность в биомедицинских исследованиях. С точки зрения информативности выходных статистических характеристик она предоставляет возможность провести более точный и взвешенный анализ выживаемости, чем рассмотренные выше, поскольку позволяет включить в расчеты целый набор переменных влияющих или предположительно влияющих на исход.

Ввиду частого использования рассматриваемой модели, попытаемся поглубже понять интерпретацию результатов обработки данных выживаемости с ее помощью. Сразу заметим, что несмотря на то, что отношение рисков (HR) может быть применено к любому моменту времени периода наблюдения, сам по себе этот показатель не дает непосредственного представления о времени до события. Отношение рисков может показывать наличие положительного эффекта применения препарата в клиническом испытании (когда HR достоверно превышает 1), что действительно предполагает укорочение времени до выздоровления. При этом значение HR может быть меньше, больше или иногда равным отношению медиан ожидаемого времени жизни, что свидетельствует о том, что это две разные статистические характеристики.

В литературе можно встретить такие суждения, основанные на значении HR, как «ускорение периода выздоровления», «выздоровление было в столько-то раз (или на столько-то процентов) более быстрым». К примеру, HR=2 может быть истолковано исследователями (это встречается в литературе) том смысле, что пациенты, получавшие препарат, выздоравливали в *2 раза быстрее*. Определение «*в 2 раза быстрее*» может (теоретически) быть понято так, что медиана ожидаемого времени эффекта (выздоровления) снизилась в результате лечения в 2 раза; что количество выздоровевших на какой-то день было в 2 раза больше в группе, получившей лечение; или что ожидаемое количество

выздоровевших на какой-то день было в 2 раза больше в группе, получившей лечение. Ни одно из этих утверждений не является примером верной интерпретации результата анализа. Значение HR не должно восприниматься, как имеющее отношение к реальной «скорости» процесса выздоровления. HR=2 в общем предполагает более быстрое выздоровление, но его понимание должно восприниматься под специфическим «вероятностным» углом зрения. Наиболее корректной «расшифровкой» HR=2 было бы, что пациент, получающий препарат и у которого выздоровление еще не наступило до какого-то момента времени, имеет в 2 раза больший шанс выздороветь к следующему моменту времени, чем тот, кто получал плацебо. Эта интерпретация кардинально отличается от тех интуитивных формулировок, которые были приведены выше. В более широком смысле HR эквивалентно шансу того, что у члена группы высокого риска событие наступит раньше, чем у члена группы меньшего риска. Вероятность того, что событие наступит раньше, может быть рассчитана из показателя HR по формуле: p=HR/(1+HR). Таким образом, HR=2 соответствует 67% шансу более раннего наступления события (например, выздоровления) у пациента, получавшего препарат, чем у того, который получал плацебо.

Алгоритм пропорциональной модели Кокса и расчета HR, а также, что очень важно, оценка доверительных интервалов, реализован в некоторых статистических программных пакетах.

2.2. Многомерный анализ

Методы **многомерного анализа** (англ. multivariate или multivariable analysis) разработаны для оценки одновременного влияния более чем одного фактора на результат (исход). В отличие от одномерной статистики, которая дает оценку того, как каждая (одна) переменная связана с интересующим нас результатом, многомерная статистика дает информацию о степени влиянии на исход каждой из (многих) переменных, а также об эффекте взаимодействия этих переменных между собой. В отечественной литературе многомерный анализ часто называют **многофакторным анализом**.

Для примера, выдвигается предположение о том, что пациенты травматологического отделения, оперированные в 9 часов утра, имеют более высокий показатель смертности, чем пациенты, оперированные в 9 часов вечера. Анализируя смертность после операции одномерным (параметрическим или непараметрическим) статистическим методом, исследователь, допустим, действительно может обнаружить достоверную разницу. Однако было бы ошибочным полагать, что время суток является единственным определяющим фактором. В анализ должны быть включены и другие переменные, такие как тяжесть травмы, возраст больного, плановые или срочные показания к операции, и т.п. После этого время операции, скорее всего, будет исключено из набора факторов, влияющих на смертность пациентов, или его вклад окажется очень малым по сравнению с истинными причинами. Лишь используя многомерный анализ, исследователь может сделать обоснованный вывод о причинах, влияющих на вариабельность результата и оценить степень одновременного влияния на него этих, нередко (в той или иной степени) взаимосвязанных, причин.

Факторы (причины), влияющие на исход, принято называть **факторами риска** (risk factors)**, независимыми** (independent) или **объясняющими переменными** (explanatory variable), а сам **исход** (outcome) – **зависимой** (dependent) или **переменной отклика** (response variable) или эффектом.

Важным моментом, обусловливающим необходимость многомерного анализа, является именно многообразие потенциальных факторов риска, возможно связанных с исходом. Экспериментальная проверка совместного влияния многих факторов в клинической практике чаще всего просто невозможна или недопустима по этическим соображениям. Положим, нужно выяснить повышает ли курение вероятность ИБС в двух случайных выборках людей, одни из которых курят, другие – нет (заставить человека курить или не курить на протяжении долгого времени и невозможно, и неэтично). Хотя предварительный одномерный статистический анализ показывает, что ИБС у курящих развивается с бóльшей вероятностью, чем у некурящих, этот результат сам по себе еще не является доказательством

причинной связи курения с ИБС, хотя может и «намекать» на наличие таковой. В принципе нельзя исключить, что более вескими причинами развития ИБС у курящих является то, что в большинстве своем это мужчины, малообеспеченные и ведущие малоподвижный образ жизни. Привычка к курению характерна именно для такой категории лиц, а все перечисленные качества повышают риск ИБС, что известно из других исследований.

В рассмотренном примере на связь между курением и ИБС могут влиять так называемые **мещающие факторы** (конфаундеры; англ. confounders). О мешающих факторах говорят, когда на видимую связь между фактором риска и независимой переменной (результатом), оказывается влияние со стороны третьей переменной, влияющей как на фактор риска, так и непосредственно (причинно) на сам результат, как показано на Рис. 8. Мужской пол, малообеспеченность и малоподвижный образ жизни вполне могут оказаться мешающими факторами, поскольку они ассоциированы как с курением, так и с ИБС.

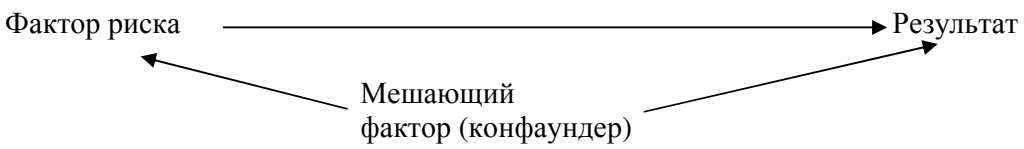

Рис. 8. Схема взаимодействия факторов, влияющих на результат (в данном случае, ИБС).

В итоге, многомерный анализ показывает, что даже с **поправкой** (учетом, нормализацией; англ. adjustment) на мужской пол, малообеспеченность или малоподвижный образ жизни, курение оказывает независимое влияние на развитие ИБС (см. Рис. 9).

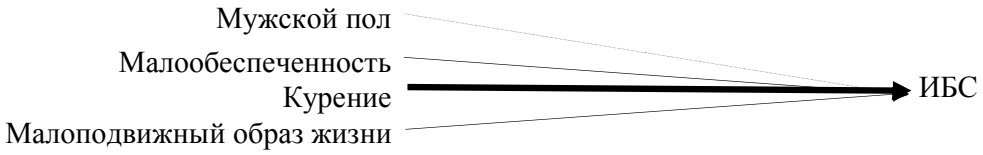

Рис. 9. Оценка степени влияния различных потенциальных факторов на результат (в данном случае - ИБС)

Определить какая переменная является независимым фактором риска, а какая конфаундером иногда невозможно ввиду того, что один и тот же фактор может оказывать как независимый эффект на результат, так и быть мешающим фактором, влияющим на другую переменную. Возвращаясь к ситуации с ИБС, малообеспеченность является мешающим фактором взаимосвязи между курением и заболеванием: вероятность того, что малообеспеченные люди злоупотребляют курением повышена, а у курящих повышена вероятность развития ИБС. С другой стороны, малообеспеченность сама по себе самостоятельно влияет на развитие ИБС: даже с учетом (поправки) курения, уровня холестерина крови, артериальной гипертензии и пр. у малообеспеченных лиц ИБС развивается с большей вероятностью, чем у людей с высоким достатком.

Заметим, что многомерный анализ является не единственным статистическим методом учета влияния или исключения конфаундеров. Оценить влияние фактора риска на исход можно также с помощью условного анализа, при котором изучаемая группа последовательно разбивается на подгруппы (страты), в которых какая-либо потенциально мешающая переменная «фиксируется». При этом модель применяется отдельно на каждой группе данных. И на каждой группе данных, вообще говоря, будет получен разный эффект и разные оценки параметров модели. Например, мужской и женский пол: далее рассматривают только курящие или некурящие мужчины и только курящие или некурящие женщины. Условный анализ эффективен когда изучается относительно небольшое число факторов (два-три). Если же потенциально мешающих факторов много, дробление приводит к образованию большого числа малочисленных выборок, в которых оценки рисков могут оказаться нестабильными (см. раздел 1.4.).

Стратификация (stratification) является способом включения в одну модель всех групп данных с разными значениями мешающих факторов таким образом, чтобы зависимость эффекта от интересующих исследователя независимых переменных подгонялась по всему множеству данных, а зависимость от мешающих факторов – только по соответствующим

подгруппам. Например, пропорциональная модель радиационного риска может состоять из произведения двух членов: фонового риска, который зависит от пола, и относительного радиационного риска, который считается одинаковым для обоих полов. В этом случае говорят, что в модели риска произведена стратификация фонового риска по полу. В общем случае, чтобы определить коэффициенты модели при интересующих независимых переменных, должна быть оценена вся модель, включая коэффициенты при мешающих факторах.

В случае применения стратификации иногда удаётся построить модель, в которой оценка интересующих исследователя переменных может быть проведена без оценки мешающих факторов. Построение таких моделей (а, соответственно, и схем исследования или выборок) с математической точки зрения весьма нетривиально. Примером является модель случай – контроль с подбором контролей к случаям по значениям мешающих параметров, разбитых на страты (matched case-control study). В этой модели отношение шансов по интересующему фактору определяется без оценки мешающих факторов. В силу математических обстоятельств, такие модели чаще всего применяются на условных выборках, которые абсолютно искажают величину эффекта в популяции, но позволяют всё-таки оценить степень статистической связи между изучаемой переменной и эффектом. Например, в условной подобранной модели случай - контроль (conditional matched case-control model) к каждому из N случаев подбирается M контролей, по полу, возрасту, месту проживания и ряду других признаков. Ясно, что в полученной выборке частота случаев не имеет никакого отношения к частоте случаев в исходной популяции. Но отношение шансов, как мера статистической связи между изучаемым фактором и эффектом, определяется независимо от мешающих факторов, по которым производился подбор контролей к случаям. Такой способ позволяет проводить анализ влияния факторов риска с высокой мощностью даже в сравнительно небольших выборках. Однако, надо отдавать себе отчёт в том, что условные модели, основанные на условных выборках – это не то же самое, что модели,

основанные на натуральных выборках (не искажающих статистических свойств популяции), в которых оценивается полный набор всех переменных. Например, отношения шансов, полученные в когортной модели заболеваемости и в условной модели случай-контроль – это разные отношения шансов, хотя и численно близкие в не слишком экзотических случаях. Напомним, что терминологически не следует путать условный анализ путём независимой подгонки модели на независимых подгруппах и условные модели, основанные на условных выборках.

2.2.1. Виды многомерного анализа

В клинических исследованиях в зависимости от задачи и типов данных чаще всего используется три метода: **множественная линейная регрессия** (multiple linear regression), **множественная логистическая регрессия** (multiple logistic regression) и **модель пропорциональных интенсивностей Кокса** (Cox proportional hazards model). В Таблице 4 представлены наиболее существенные характеристики этих трёх моделей.

Таблица 4. Основные модели многомерного анализа

Модель	Тип данных	Специфические особенности
Множественная линейная регрессия	Диапазон значений исхода (напр., показатели АД)	Коэффициенты при переменной линейно связаны с (влияют на) результатом
Логистическая регрессия	Дихотомические результаты (Да/Нет)	Модель ограничивает вероятность исхода от 0 до 1
Модель пропорциональных интенсивностей Кокса	Период времени до события (время до выздоровления/рецидива/смерти)	Применяется для продолжительных исследований, в которых объекты могут быть потеряны во время наблюдения

Множественная линейная регрессия используется для изучения изменения зависимой переменой (y) в ответ на различные значения других переменных (x_1, x_2, x_3), которые представляют собой непрерывные (численные интервальные или относительные) переменные. Модель предполагает, что с увеличением (или уменьшением) значений независимых переменных значение зависимой переменной (исхода) изменяется линейно. Для линеаризации нелинейно влияющих независимых переменных часто используется математическое приведение, такое, например, как логарифмирование. Для интервальных

переменных также предполагается, что для равных промежутков на всей шкале интервала степень влияния на исход будет одинакова. Величина коэффициента при независимой переменной и его знак в конечной модели показывают степень и характер взаимосвязи между этой переменной и исходом. Примером множественной линейной регрессии может быть модель оценки костной плотности у женщин в менопаузе, в которую входят возраст и индекс массы тела.

Логистическая регрессия используется когда значение переменной результата является бинарным, таким как выживаемость (да/нет), развитие заболевания (да/нет), положительный результат диагностического теста (да/нет) и может включать одну или более независимых переменных. К примеру, для прогнозирования летальности пациентов травматологического отделения (см. выше) может быть использована модель, которая учитывает возраст пациента, степень тяжести травмы и причину травмы. Логистическая регрессия при использовании модели пропорционального риска позволяет оценить **относительный риск** (relative risk – RR), **отношение шансов** (odds ratio - OR) и границы их доверительных интервалов (confidential intervals - CI), а также степень достоверности (величину *p*) отличия этих величин от 1 (значение при нулевой гипотезе). Значение OR и 95% CI предоставляют наглядную информацию о взаимосвязи независимой переменной с исходом. OR близкое к 1 свидетельствует о слабой взаимосвязи. Широкий CI наводит на мысль о невысокой надежности оценки и о необходимости проверки данных на предмет их аккуратности. Если границы CI включают значение 1, связь независимой переменной с исходом не может быть признана достоверной, сколь намного бы значение OR не отличалось от 1.

Модель пропорциональных интенсивностей Кокса оценивает шансы более раннего наступления события у членов изучаемой группы по сравнению с контрольной группой с помощью показателя отношения рисков (HR), как рассмотрено в разделе 1.4.

2.2.2. Включение независимых переменных в модель

Любая модель многомерного анализа должна включать как минимум один или несколько факторов риска и потенциальные мешающие факторы (конфаундеры). Однако универсального рецепта включения как факторов риска, так и потенциальных конфаундеров не существует. В связи с этим, подход к вопросу должен быть очень осторожным. В идеале в модель нужно включить все переменные, которые были определены с помощью теоретических рассуждений или установлены в предыдущих исследованиях как факторы риска или конфаундеры изучаемого исхода. Так, для выяснения влияния гиперальбуминемии на смерть от ССЗ в многомерную модель пропорциональных рисков должны быть включены возраст, пол, статус курения, наличие артериальной гипертензии, дислипидемия, диабет, ожирение, уровень креатинина в крови и др. известные как связанные с исходом.

С другой стороны, важным является не только включение в анализ всех потенциально важных переменных, но и исключение посторонних. Например, в модель изучения связи курения и рака легкого нет смысла включать наличие ИБС в список факторов риска или конфаундеров, хотя риск ИБС повышен у курящих. Ни теоретически, ни по опыту предыдущих исследований ИБС никак не была связана с развитием изучаемой онкопатологии. Если массив данных включает очевидно связанные между собой или сильно коррелирующие независимые переменные, рекомендуется произвести сознательный выбор одной как наиболее важной. Для примера, исследование причин неонатальной смертности показало, что поскольку вес при рождении и срок гестации очень взаимосвязаны, нет необходимости включать в модель обе переменные. Авторы исключили срок гестации, аргументируя это тем, что вес при рождении оказывает схожее, но более сильное влияние на исход.

На выбор включаемых в анализ переменных также влияет предназначение создаваемой модели, которая может быть **объясняющей** (explanatory) или **прогностической** (prognostic). Назначением объясняющей модели является выяснение характера и степени

влияния различных факторов на результат. Для такой модели тщательный подбор переменных и их математической формы, является крайне важным. Прогностические модели направлены на определение вероятности происхождения события. Если созданная модель хорошо воспроизводится и эффективно работает на независимых массивах данных, то вопросы подбора и ревизии переменных в ней уже не являются актуальными. В прогностической модели наиболее важным является точность предсказания результата. Например, в неё могут быть включены коррелирующие независимые переменные, и модель будет хорошо работать. Но оцененные коэффициенты модели для таких переменных не имеют самостоятельного значения и не несут информации о степени влияния на эффект каждой независимой переменной в отдельности. Такая модель не может быть объясняющей.

Количество переменных в модели можно оптимизировать с помощью автоматизированных алгоритмов их подбора. Эти алгоритмы помогают компьютеру отобрать переменные на основе критериев, определенных исследователем. Отбор осуществляется методами прямого пошагового отбора, обратного пошагового удаления и наилучшего подмножества. Для прямого пошагового отбора переменная, оказывающая наиболее сильное влияние на исход по результатам одномерного анализа, вводится первой, следом за ней добавляется переменная со следующим наиболее сильным влиянием и т.д. до тех пор, пока все переменные, влияющие на результат (с уровнем значимости, определенным исследователем; обычно в многомерном регрессионном анализе он устанавливается >90%, т.е. $p<0,1$) не будут включены в модель. Любая ранее введенная в модель переменная, которая перестает быть значимой при введении следующей переменой, последовательно исключается. В методе обратного пошагового удаления в модель сначала включаются все переменные. Затем они последовательно удаляются, начиная с переменной, имеющей наиболее слабую ассоциацию с результатом. Удаление продолжается до тех пор, пока в модели не останутся только те переменные, которые достоверно влияют на исход.

Метод наилучшего подмножества подразумевает выбор путем подстановки такого набора переменных, которые наилучшим образом удовлетворяют условиям, определенным исследователем. При автоматизированном подходе к выбору переменных может получиться так, что не все основные и мешающие переменные окажутся в модели, или в модели могут отсутствовать наиболее значимые клинические показатели. Поэтому после создания модели в автоматизированном режиме от исследователя требуется ее критическая оценка на адекватность.

2.2.3. Взаимодействие между переменными

О **взаимодействии между переменными** (interactions) в эффекте говорят, когда влияние фактора риска на исход (эффект) зависит от значения третьей синтетической переменной, составленной из двух исходных независимых переменных. При этом сама третья переменная не является независимым фактором риска или мешающей переменной. Взаимодействие между переменными также называют **эффектом модификации** (effect modification) в том случае, если одна из них рассматривается как основная с содержательной точки зрения.

Для примера, клинические испытания продемонстрировали, что некоторый препарат снижает вероятность перелома кости у пациентов с остеопорозом, но не у людей с изначально более высокой минеральной плотностью костной ткани (условный анализ на разных группах пациентов). Чтобы свести весь доступный метериал в одну модель (для увеличения достоверности и мощности исследования), был введён член взаимодействия. Исследование проводилось с помощью пропорциональной модели Кокса с учётом переменных приема препарата, результата анализа минеральной плотности костной ткани, а также синтетической переменной, которая представляла собой композицию (для двух непрерывных переменных было бы произведение) первых двух переменных. Достоверность влияния этой синтетической переменной на исход означала, что эффект препарата зависит от начальной минеральной плотности костной ткани.

Учет взаимодействия между переменными может оказаться клинически важным, однако для внесения эффекта взаимодействия в модель необходимо изначально иметь предположение о том, что переменные могут взаимодействовать. В противном случае начинается почти системный поиск взаимодействия путем разбиения групп на подгруппы, и чем больше ожидается взаимодействующих переменных, тем больше образуется подгрупп данных. Это может привести к тому, что в одной или нескольких из них взаимодействие будет обнаружено в силу случая (ошибка 1 типа).

2.2.4. Анализ качества модели

Для оценки эффективности множественной линейной регрессии используется уже известный из корреляционного анализа **коэффициент детерминации** r^2 (см. раздел 1.7.1.), который отражает степень рассеяния результата, возникающего благодаря вкладу многих переменных. Значение r^2 варьирует в пределах от 0 до 1 и чем ближе оно к 1, тем лучше модель описывает результат. Ввиду того, что в модель может входить несколько факторов риска, коэффициент рассчитывается с поправкой на их количество.

Для логистических регрессий предложено несколько статистических **критериев согласия** (goodness-of-fit test), каждый из которых имеет свои достоинства и недостатки. Чаще применяется тест Хосмера-Лемешова (Hosmer-Lemeshow test). Критерии согласия обычно используются для оценки эффективности объясняющих моделей. О надежности объясняющих моделей можно судить по их воспроизводимости на других массивах данных. Если модель надежна, то и в независимом массиве данных в модель войдут те же факторы риска с коэффициентами близкими к тем, что наблюдались в оригинальной модели.

Для прогностических моделей рекомендуется более точная количественная оценка. Для ее получения рассчитываются такие показатели, как *чувствительность*, *специфичность* и *точность* (см. раздел 1.8.) для определенных пороговых условий (например, исходя из предположения, что у всех лиц, у которых предсказанная вероятность развития заболевания была ≥40%, действительно развилось заболевание). Исходя из значений *чувствительности* и

специфичности, строится *характеристическая кривая* (ROC-кривая, см. раздел 1.8.) по форме которой и по величине площади под которой (AUC) можно судить об удачности модели. Надежность прогностической модели также может быть проверена на независимых массивах данных, в которых она должна предсказывать вероятность исхода с высокой эффективностью.

Завершая рассмотрение многомерного анализа, заметим, что хотя его алгоритмы реализованы в ряде статистических программ, он требует специальных знаний и подготовки в математической статистике. Поскольку наиболее важные для клиницистов выводы обычно делаются на основании именно многомерного анализа, его корректное выполнение и интерпретация имеют особое значение.

Заключение. В данном обзоре мы остановились на ряде наиболее часто используемых в медицинских исследованиях методов статистического анализа. Статистический анализ является неотъемлемой частью практически любого исследования, и только с его помощью можно пополнить доказательную базу. Исходя из собственного опыта, осмелимся высказать мнение, что наиболее значимые и глубокие выводы делаются на основании всестороннего и тщательного проведенного статистического анализа, в котором могут используются довольно сложные алгоритмы. В связи с этим его самостоятельное выполнение клиницисту не всегда по силам. Во многих случаях необходимо участие специалиста с профессиональной подготовкой в области математической статистики. Именно в ходе сотрудничества клинициста с математиком можно рассчитывать на проведение глубокого и корректного статистического анализа данных.

Чтобы сделать диалог более продуктивным, клиницист должен произвести ревизию информации со статистических позиций. Убедившись, что исследование обладает достаточной мощностью, необходимо составить себе четкое представление об имеющхся данных: их типах, распределению и пр. Далее необходимо охарактеризовать данные с помощью описательной статистики и произвести их одномерный статистический анализ,

после чего можно сформировать вопросы для многомерного анализа. Это будет несомненным подспорьем при обсуждении дальнейшего анализа с математиком, что значительно повышает шансы на взаимное понимание и, в итоге, на успешный результат. Мы последовательно рекомендуем и являемся сторонниками подобного сотрудничества. На наш взгляд, оно является наиболее рациональным подходом к правильной интерпретации результатов и формированию корректных выводов.

Выражаем надежду, что представленное описание методов статистического анализа окажется полезным клиницистам, особенно тем, кто находится в начале своего профессионального пути.

Список основной литературы

1. *Гланц С.* Медико-биологическая статистика. // Пер. с англ., М: Практика 1999; 459 стр.
2. *Bland J.M., Altman D.G.* Survival probabilities (the Kaplan-Meier method). // BMJ., 1998, V. 317., p. 1572.
3. *Bland J.M., Altman D.G.* The logrank test. // BMJ., 2004, V. 328., p. 1073.
4. *Bowers D., House A., Owens D.* Understanding clinical papers. // 2nd edition. England: John Wiley and Sons Ltd., 2006, 232 p.
5. *Cassidy L.D.* Basic concepts of statistical analysis for surgical research. // Journal of Surgical Research. 2005. Vol. 128, No. 2, p. 199-206.
6. *Davis C.S.* Statistical methods of the analysis of repeated measurements. // New York: Springer-Verlag, 2002, 744 p.
7. *Katz M.H.* Multivariable analysis: A primers for readers of medical research. // Annals of Internal Medicine. 2003, Vol. 138, No. 8, p. 644 – 650.
8. *Kirkwood B., Sterne J.* Essential Medical Statistics. //2nd ed, Blackwell Publishing, 2003, 501 p.
9. *Livingston E.H.* The mean and standard deviation: what does it all mean? // Journal of Surgical Research, 2005, Vol. 119, No. 2, p. 117 – 123.
10. *Livingston E.H., Cassidy L.* Statistical power and estimation of the number of required subjects for a study based on the t-test: A Surgeon's primer. // Journal of Surgical Research, 2005, Vol. 128, p. 207 – 217.
11. *Machin D., Cheung Y., Palmar M.* Survival Analysis: A practical approach, 2nd edition.: John Wiley & Son, Ltd., 2006, 278 p.
12. *Peat J., Barton B.* Medical statistics: a guide to data analysis and critical appraisal. // NY: Blackwell Publishing, 1st ed., 2005, 324 p.
13. *Petrie A., Sabin C.* Medical Statistics at a Glance: Blackwell Publishing, 2005, 157 p.
14. *Rao S.R., Schoenfeld D.A.* Survival methods. // Circulation., 2007, Vol. 115, p. 119 – 113.
15. *Royston P., Parmar M.K., Altman D.G.* Visualizing of survival in time-to-event studies: a compliment to Kaplan-Meier Plots. // J. Natl. Cancer Inst., 2008, Vol. 100, p. 92-97.
16. *Scott I., Mazhindu D.* Statistics for health care professionals. // London: SAGE Publications Ltd., 2005, 241 p.
17. *Spruance S.L., Reid J.E., Grace M, Samore M.* Minireview: Hazard ratio in clinical trial. Antimicrobial Agents and chemotherapy. 2004. Vol. 48, p. 2787 – 2792
18. *Velleman P. F., Wilkinson, L.* Nominal, ordinal, interval, and ratio typologies are misleading. // The American Statistician, 1993. Vol. 47, p. 65-72.