68. *Targher G., Bertolini L., Scala L.* et al. // Clin. Endocrinol. — 2004. — Vol. 61. — P. 700—703.
69. *Thamer C., Machann J., Tschritter O.* et al. // Horm. Metab. Res. — 2002. — Vol. 34.
70. *Thamer C., Machann J., Haap M.* et al. // Dtsch. Med. Wschr. — 2004. — Bd 129. — S. 872—875.
71. *Tilg H., Moschen A. R.* // Mol. Med. — 2008. — Vol. 14. — P. 222—231.
72. *Tomas E., Tsao T. S., Saha A. K.* et al. // Proc. Natl. Acad. Sci. USA. — 2002. — Vol. 99. — P. 16309—16313.
73. *Unger R. H., Orci L.* // Int. J. Obes. Relat. Metab. Disord. — 2000. — Vol. 24. — Suppl. 4. — P. S28—S32.
74. *Urano F., Wang X., Bertolotti A.* et al. // Science. — 2000. — Vol. 287. — P. 664—666.
75. *Utzschneider K. M., Carr D. B., Tong J.* et al. // Diabetologia. — 2005. — Vol. 48. — P. 2330—2333.
76. *Weinstein A. R., Sesso H. D., Lee I. M.* et al. // J. A. M. A. — 2004. — Vol. 292. — P. 1188—1194.
77. *Weisberg S. P.* et al. // J. Clin. Invest. — 2006. — Vol. 116. — P. 115—124.
78. *Williamson R. T.* // Br. Med. J. — 1901. — Vol. 1. — P. 760—762.
79. *Yamauchi T., Kamon J., Waki H.* et al. // Nat. Med. — 2001. — Vol. 7. — P. 941—946.
80. *Yang R. Z., Lee M. J., Hu H.* et al. // Am. J. Physiol. Endocrinol. Metab. — 2006. — Vol. 290. — P. E1253—E1261.
81. *Yin M. J., Yamamoto Y., Gaynor R. B.* // Nature. — 1998. — Vol. 396. — P. 77—80.
82. *Youn B. S., Kloting N., Kratzsch J.* et al. // Diabetes. — 2008. — Vol. 57. — P. 372—377.
83. *Yuan M., Konstantinopoulos N., Lee J.* et al. // Science. — 2001. — Vol. 293. — P. 1673—1677.
84. *Zulet M. A., Puchau B., Navarro C.* et al. // Nutr. Hosp. — 2007. — Vol. 22, N 5. — P. 511—527.

Поступила 28.10.08

♦ ЛЕКЦИЯ

УДК 616-07:312

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА В КЛИНИЧЕСКОЙ ПРАКТИКЕ

Часть I. Одномерный статистический анализ

П. О. Румянцев, В. А. Саенко, У. В. Румянцева

ГУ Медицинский радиологический научный центр РАМН, Обнинск (дир. — акад. РАМН А. Ф. Цыб)

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.*

P.O. Rumyantsev, V.A. Saenko, U.V. Rumyantseva

STATISTICAL METHODS FOR THE ANALYSES IN CLINICAL PRACTICE. PART 1. UNIVARIATE STATISTICAL ANALYSIS

Medical Research Radiological Centre, Russian Academy of Medical Sciences, Obninsk

Statistical analysis is an integral component of any clinical study. The objective of the present work was to assist clinicians in understanding various methods of statistical treatment of medical findings without going into details of mathematical computation. The most strongly sought for and popular analytical procedures are considered in application to clinical and experimental medical research. Part 1 of this communication is focused on descriptive statistics and methods of univariate statistical analysis. Part II will be concerned with the analysis of survivorship and multivariate statistics.

Key words: *methods of statistical analysis, medicine, descriptive statistics, algorithm, distribution, parametric and non-parametric statistics, confidence, statistical strength, linear regression, diagnostic information value*

На протяжении всей своей истории медицина искала пути повышения эффективности результатов диагностики и лечения. Начиная с интуитивных 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

Сведения об авторах

Для контактов:
***Румянцев Павел Олегович*, канд. мед. наук, вед. науч. сотр.**
Адрес: 249036, Обнинск, ул. Королева, 6
Телефон: 8(48439) 93241
roum@mrrc.obninsk.ru
Саенко Владимир Александрович, канд. биол. наук, ст. науч. сотр.
Румянцева Ульяна Викторовна, канд. мед. наук, ст. науч. сотр.

протокола результаты исследования могут считаться доказательными.

Ввиду того, что объемы данных и размеры групп (выборок) могут сильно варьировать, а данные могут 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, комплексная статистическая обработка представляет собой сложную задачу. Во многих случаях, если не в большинстве, для глубокого анализа клинических данных необходимо участие специалиста с профессиональной подготовкой в области математической статистики. 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 рассматриваем базисные концепции наиболее востребованных в медицине методов статистического анализа.

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 нее. К примеру, для изучения эффективности какой-либо операции невозможно собрать данные на всех пациентов, когда-либо ей подвергавшихся. Вместо этого подбирают и анализируют репрезентативную выборку. 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 дает 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_0) 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.

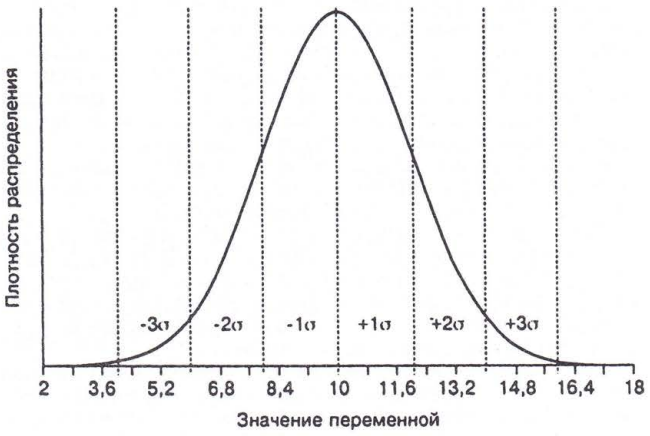

Рис. 1. Плотность нормального распределения.

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 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.

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 свое 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 недопустимы; 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 С. Стивенс ввел следующие уровни классификации переменных: номинальный (nominal), порядковый (ordinal) и непрерывный (continuous), причем последний делился на подуровни: интервальный (interval) и относительный (ratio).

Дискуссия о "правильной" классификации переменных в статистике продолжается до сих пор. На сегодняшний день 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 3 типа переменных: непрерывные, дискретные и категориальные (номинальные). Непрерывные переменные (continuous variables) могут принимать любые численные значения, которые естественным образом упорядочены на числовой оси (например, рост, масса тела, артериальное давление (АД), СОЭ).

Дискретные переменные (discrete variables) способны принимать счетное 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 определенные значения.

Нормальное (или гауссово) распределение имеет колоколообразную форму, абсолютно симметричную относительно оси, проходящей через среднее значение (рис. 1) и математически описывается формулой, включающей 2 параметра — среднее и стандартное отклонение (см. раздел 3).

Оценка соответствия распределения данных гауссову 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 АД в группе мужчин по сравнению с группой женщин: АД у мужчин не зависит от аналогичного показателя у женщин. Примером зависимых выборок являются показатели АД, измеренного у пациентов в 9 ч утра и измеренного у них же в 5 ч вечера. Результаты этих измерений для каждого человека и в целом между выборками скорее всего будут коррелировать, поэтому они считаются парными и оцениваются как зависимые.

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 массы 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 сигм. Оно гласит, что практически все наблюдения укладываются в интервал "среднее $\pm$ 3σ". Действительно, в интервал "$\pm 3\sigma$" попадают 99,7% наблюдений, "$\pm 2\sigma$" включает 95,4% всех наблюдений, а "$\pm 1\sigma$ — всего 68,3. 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). Стандартную ошибку рассчитывают 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 выше 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% 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 $\pm 1{,}96$ 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 АД 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 $\pm 9{,}8$ ($5 \cdot 1{,}96$) 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 $\pm$ стандартное отклонение или стандартную ошибку? Это зависит от того, разброс чего — исходной случайной величины или оценки ее среднего значения (медианы) — изучает исследователь. Если непрерывные переменные распределены нормально (или близко к таковому) и разброс данных обусловлен естественными причинами (люди разного роста, массы тела и т. п.), то принято указывать среднее $\pm$ стандартное отклонение. Если же рассеяние связано с неточностью измерения (например, техническое ограничение или погрешность прибора), то рекомендуется приводить среднее $\pm$ (95%) доверительный интервал или стандартная ошибка. Во всяком случае необходимо указать, какие именно характеристики представлены. Когда непрерывные данные не подчиняются нормальному распределению, для их описания обычно используют медиану 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.

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-го и 2-го типа.

При проверке гипотезы принимается во внимание возможность ошибок измерений, что может стать причиной ложного результата. В зависимости от характера возможного ложного результата, ошибки бывают 1-го и 2-го типа. Ошибка 1-го типа (обозначается α) определяется как вероятность обнаружить различие, которое в действительности отсутствует ("ложноположительный результат"). Другими словами, это вероятность неправомерно отбросить гипотезу (H_0) в пользу гипотезы H_a. Ошибка 2-го типа (обозначается β) — это вероятность сделать вывод об

Таблица 1

Типы ошибок и статистическая мощность исследования

Результаты проверки гипотезы	Истинный, но неизвестный характер взаимодействия	
	гипотеза H_0 неверна	гипотеза H_0 верна
Отвергнуть гипотезу H_0	Корректное решение (достаточная статистическая мощность)	Ошибка 1-го типа (α)
Принять гипотезу H_0	Ошибка 2-го типа (β)	Корректное решение

отсутствии различия, в то время как фактически оно имеется ("ложноотрицательный результат"), т. е. неправомерно принять гипотезу H_0. В биомедицинских исследованиях предельно допустимый предел ошибки 1-го типа обычно устанавливается на уровне 5%, а ошибки 2-го типа — не более 20% ($\alpha = 0,05$; $\beta \leqslant 0,2$). Ошибка 1-го типа рассматривается как более критическая, потому что менее всего хотелось бы неправомерно отвергнуть общепринятую гипотезу (H_0). 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).

Понимая природу ошибок 1-го и 2-го 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"). В табл. 1 показана взаимосвязь между ошибками 1-го и 2-го 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 незначительных различий потребуется более многочисленная выборка. Если планируемая численность выборки не обеспечивает приемлемого уровня статистической мощности ($\geqslant 80\%$), чтобы убедительно отвергнуть гипотезу H_0 или согласиться с ней, результаты исследования не будут доказательными. Например, если исследователь хочет определить различие в средней массе тела между двумя группами (получавшими и не получавшими препарат, снижающий аппетит) и доказать разницу в 1 кг при стандартном отклонении 10 кг в контрольной и изучаемой группах, то при $\alpha = 0,05$ 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). В случае использования 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.

5. Статистическая достоверность

При сравнении групп мы изначально исходим из того, что они не различаются (это — H_0). 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 еще 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 на множестве выборок (причем с равномерным распределением на интервале 0—1).

С помощью статистических расчетов вычисляют значение *p*, которое затем сравнивают с заранее выбранным уровнем значимости, часто обозначаемым греческой буквой α (не путать с ошибкой 1-го типа). Обычно в биомедицинских исследованиях уровень значимости устанавливается на уровне $\alpha \leqslant 0,05$ ($\leqslant 5\%$). Если выбран уровень значимости $\alpha = 0,05$, то все выборки, которые для выдвинутой гипотезы возвращают величину $p \leqslant 0,05$, отвергают эту гипотезу, а выборки с величиной $p > 0,05$, не дают оснований для того, чтобы ее отвергнуть. Величину уровня значимости следует понимать так: мы задаем, что не более чем в 5% попыток сравнения (какого-либо параметра в разных группах) обнаруженная разница может быть обусловлена чистой случайностью, а не ее реальным существованием. Иными словами, мы задаем вероятность ложного отказа от гипотезы H_0 (стандартной) в пользу гипотезы H_0 (изучаемой). В итоге, повторимся, если статистический анализ показывает, что $p \leqslant 0,05$, 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 $\pm 1,96$ 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 статистически недостоверно. 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 уверенности в наличии разницы по какому-либо параметру между сравниваемыми группами нужно, чтобы "усы ошибок" (границы доверительных интервалов) не пересекались. В определенном смысле это верно: непересечение "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.

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 — задача нетривиальная, 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); 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 групп и более с целью сравнения лишь одной переменной. Многомерная статистика используется для анализа двух групп и более, 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 позволит 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.

6.1. Параметрическая статистика

Параметрическая статистика используется для анализа непрерывных (численных) переменных, значения которых распределены нормально. Наиболее часто используется так назы-

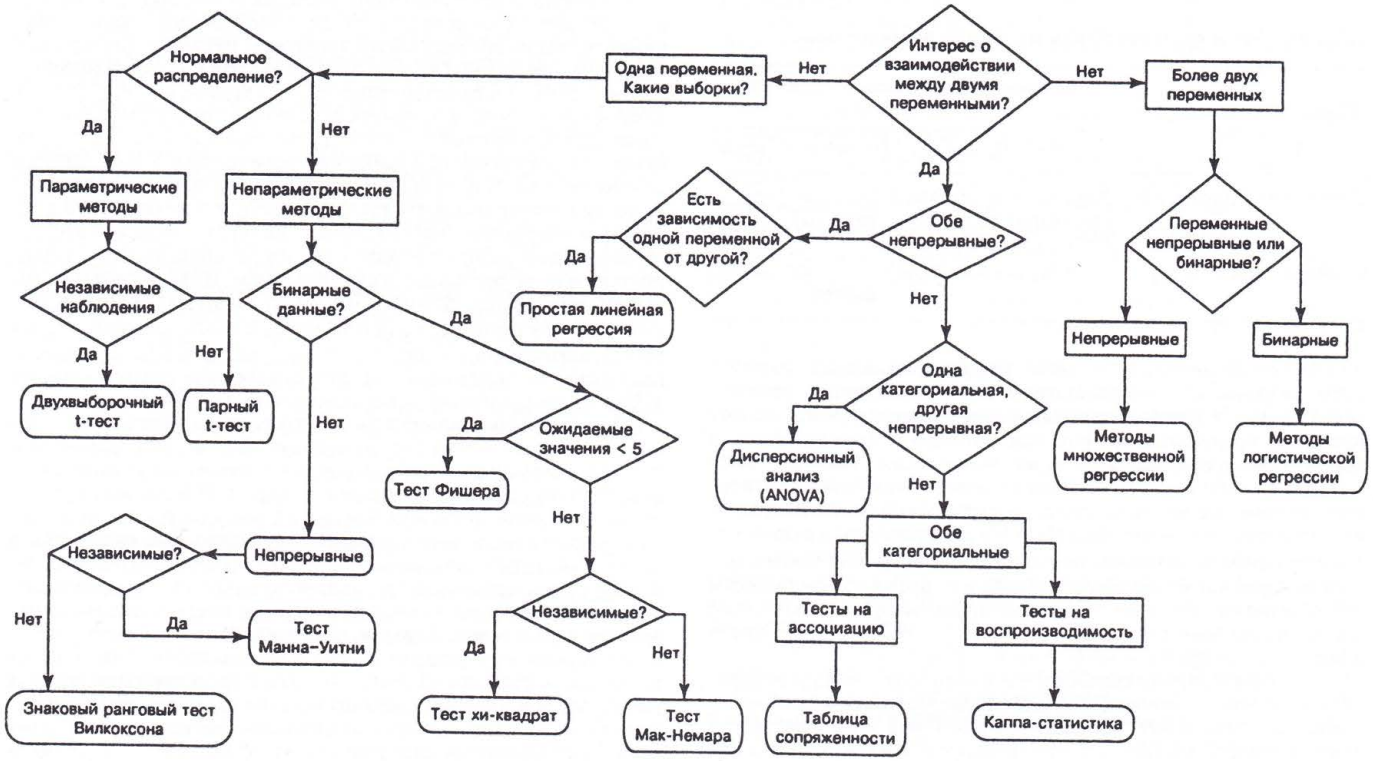

Рис. 2. Алгоритм выбора метода одномерного статистического анализа [5].

ваемый непарный t-тест (распространенное название — "тест Стьюдента"; t-test), с помощью которого возможно провести проверку гипотезы (H_0) 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.

Распространенной ошибкой является применение t-тестов к показателям состояния пациентов (пациента) до и после применения 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 используют 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 двусторонние 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 (ANalysis Of Variance, ANOVA). К примеру, его можно применить для выявления разницы среднего систолического АД в различных возрастных группах. Для зависимых данных, оцениваемых в более чем двух группах, используется дисперсионный анализ с повторным измерением (Repeated-Measures ANOVA, RM-ANOVA).

6.2. Непараметрическая статистика

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

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 остается логичным выбором. Проблемы с интерпретацией теста, как и в случае t-тестов, возникают, когда распределения для двух выборок различаются по форме, например, имеют сильно отличающиеся дисперсии.

Для иллюстрации важности адекватного выбора статистического теста предположим, что исследователь сравнивает массу тела в двух независимых группах пациентов. В 1-й группе, помимо людей с "нормальной" массой тела, имеется два полных человека; средняя масса тела в группе составила 100,3 кг, а медиана — 75,1 кг. Во 2-й группе, напротив, есть несколько худощавых людей; средняя масса тела в группе — 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 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 оправданно в том случае, если распределение переменной массы тела в обеих группах оказалось 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 различия между группами.

При сравнении переменной более чем в двух независимых группах непараметрическим аналогом дисперсионного анализа является тест Краскела—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.

Таблица 2

Таблица сопряжения непарных дискретных данных

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

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).

χ^2-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 × 2 (точный тест Фишера изначально разработан для таблиц сопряжения размерностью χ^2), однако многие исследователи традиционно предпочитают статистику χ-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 χ^2-тест основываются на принципиально разной идеологии расчета. Точный тест Фишера использует перебор вариантов заполнения таблицы сопряженности (перестановочный тест), в то время как χ^2-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 × 2 предпочтителен точный тест Фишера, поскольку он дает более точную оценку, чем χ^2-тест. Однако применение и χ^2-теста как для таблиц 2 × 2, так и для таблиц большей размерности, также правомерно.

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 χ^2-тест, как правило, непротиворечиво выдают значение *p*, которое будет либо больше, либо меньше установленного порогового уровня значимости, например, на уровне 0,05.

Пример данных, организованных в таблицу сопряжения размерностью 2 × 2, приведен в табл. 2, В ней рассматривается абстрактная ситуация возникновения побочного эффекта (например, тахикардии) после применения какого-либо препарата.

Расчеты, проведенные с помощью точного теста Фишера и χ^2-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 $(А \cdot Г)/(Б \cdot В)$. 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 табл. 3, отношение шансов составляет 0,93, а 95% доверительный интервал от 0,56 до 1,53. В англоязычной литературе показатель часто записывается в таком виде: 0,93 [0,56—15,3] (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, используют тест Мак-Немара (McNemar test), который представляет собой модификацию χ^2-теста для парных или соотнесенных данных. Примером уместного использования теста Мак-Немара 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 Мак-Немара часто используется в исследованиях типа "случай-контроль" (case-control study), в которых каждому случаю противопоставляется конкретный контроль. Для расчетов с помощью теста Мак-Немара составляют таблицу сопряжения, подобную табл. 3, однако в каждой ячейке указывают не количество лиц, соответствующих какому-либо исходу, а количество пар (до/после лечения, случай/контроль).

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.

7. Корреляционный и регрессионный анализ

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

7.1. Корреляционный анализ

Корреляционный анализ является методом оценки линейных связей (общей пропорциональности) между переменными, т. е. определяет, насколько согласованно они меняются. В англоязычной литературе часто употребляется термин "линейная корреляция Пирсона". Корреляция Пирсона (обычно просто

Таблица 3

Организация данных для оценки информативности диагностического теста

Результат теста	Общепринятый ("золотой") стандарт	
	положительный	отрицательный
Положительный	а — число пациентов, имеющих заболевание и положительный результат теста	б — число пациентов, не имеющих заболевания и положительного результата теста
Отрицательный	в — число пациентов, имеющих заболевание и отрицательный результат теста	г — число пациентов, не имеющих заболевания и отрицательного результата теста
Итого...	а + в = общее число имеющих заболевание	б + г = общее число не имеющих заболевания

Примечание. 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.

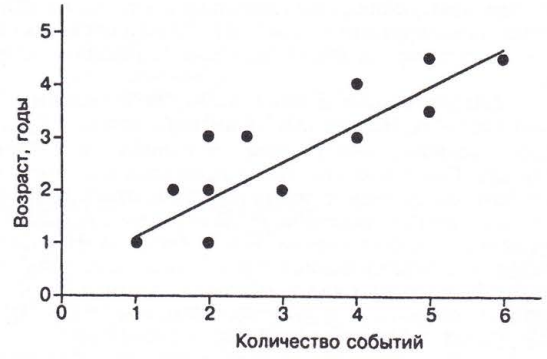

Рис. 3. Пример положительной корреляции.

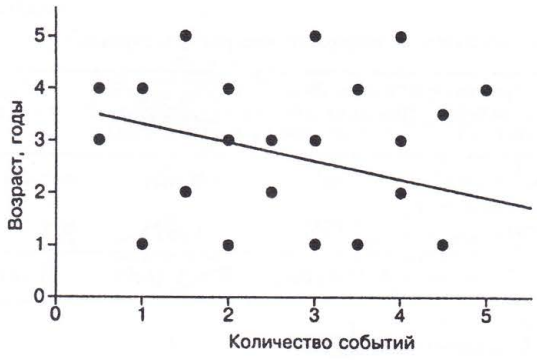

Рис. 5. Пример отсутствия корреляции.

"корреляция") между переменными может быть положительной, отрицательной или вовсе отсутствовать.

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

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

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

Показателем согласованности между значениями двух переменных служит 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. определяют, 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 имеет большого значения.

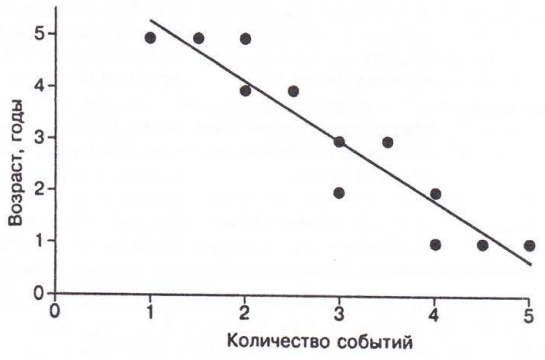

Рис. 4. Пример отрицательной корреляции.

Необходимо иметь в виду, 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.

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.

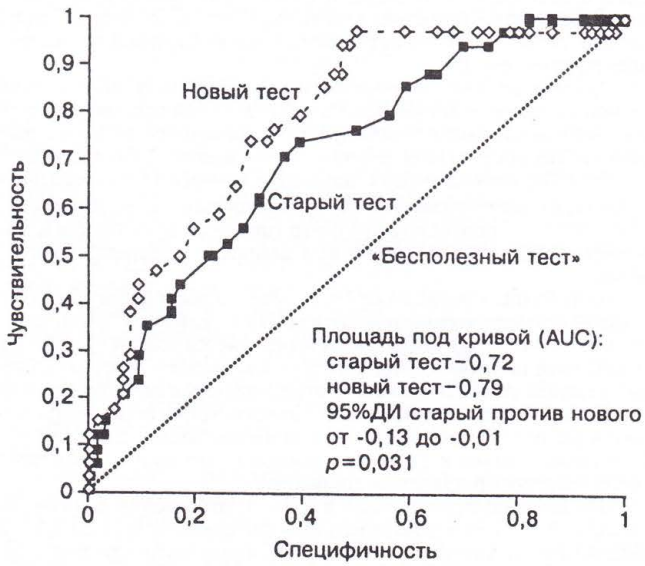

Рис. 6. Пример характеристических (ROC) кривых двух тестов — "старого" и "нового".

Статистические программы, помимо коэффициента корреляции r, коэффициента детерминации r_2, коэффициентов a и b регрессионной прямой, рассчитывают 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, возведение в степень, логарифмирование или описание более сложной функцией). После этого в ряде случаев линейный регрессионный анализ становится вновь возможным.

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. Модель таблицы сопряжения для проведения расчетов представлена в табл. 2. По существу, она отражает соотношение между ошибками 1-го и 2-го типа (см. раздел 4).

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. Receiver Operating Characteristic (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, 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 (рис. 6).

Количественную оценку характеристической кривой можно провести, рассчитав площадь под ней (англ. Area Under Curve, AUC). Приблизительная шкала значений AUC, отражающая качество диагностического теста, такова:

AUC = 0,91—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 онлайн ROC-калькуляторов. На рис. 6 для примера показаны реальные расчетные характеристические кривые. Многие статистические программы способны генерировать сглаженные кривые и возвращать необходимые статистические оценки. В рассмотренном примере "новый" тест имеет достоверно лучшие характеристики по сравнению со "старым".

Заключение

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

Во второй части обзора будут рассмотрены принципы анализа выживаемости и методы многомерной статистики.

Авторский коллектив выражает благодарность С. Ю. Чекину (МРНЦ РАМН) за конструктивную помощь и критические замечания при подготовке данной работы.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Гланц С. Медико-биологическая статистика: Пер. с англ. — М., 1999.
Cassidy L. D. // J. Surg. Res. — 2005. — Vol. 128, N 2. — P. 199—206.
Davis C. S. Statistical Methods of the Analysis of Repeated Measurements. — New York, 2002.
Kirkwood B. R., Sterne J. A. Essential Medical Statistics. — 2-nd Ed. — New York, 2003.
Livingston E. H. // J. Surg. Res. — 2004. — Vol. 119, N 2. — P. 117—123.
Livingston E. H., Cassidy L. // J. Surg. Res. — 2005. — Vol. 126,N 2. — P. 207—217.
Machin D., Cheung Y., Palmar M. Survival Analysis: A Practical Approach. — 2-nd Ed. — London, 2006.
Petrie A., Sabin C. Medical Statistics at a Glance. — New York, 2005.
Spruance S. L., Reid J. E., Grace M., Samore M. // Antimicrob. Agents Chemother. — 2004. — Vol. 48, N 8. — P. 2787—2792.
Stevens S. S. // Science. — 1946. — Vol. 103. — P. 677—680.
Velleman P. F., Wilkinson L. // Am. Statist. — 1993. — Vol. 47, N 1. — P. 65—72.

Поступила 16.12.08