УДК 616-071/-078

Проверка гипотезы: что необходимо знать о достоверности различия

Г. П. Тихова

Карельский научный центр РАН, 185910, Петрозаводск

Testing of hypothesis: what is necessary to know about statistically significant difference

G. P. Tikhova

FSBI "Karelian Research Center of Russian Academy of Science", 185910, Petrozavodsk

В статье рассматривается алгоритм сравнения двух выборок, выбор подходящего критерия, а также разъясняется смысл некоторых статистических терминов, используемых в критериях сравнения, в частности, нуль-гипотеза, ошибки первого и второго рода, уровень значимости и мощность критерия. *Ключевые слова*: проверка статистической гипотезы, уровень значимости.

The article is dedicated to correct choosing of criteria for comparison of two groups in scientific research. Several important terms and notations are comprehensively considered and clarified such as null hypothesis, significance level, type I and type II error and criteria power. *Key words*: testing of statistical hypothesis, significance level.

Наибольшее количество исследовательских работ, публикуемых в современных медицинских изданиях, посвящено, если сказать в общих словах, двум проблемам: сравнению и влиянию.

Сравнивают методы лечения, препараты, их дозировки и схемы введения, сравнивают процедуры, режимы терапии, манипуляции в условиях той или иной патологии (или комплекса заболеваний). Что касается изучения влияния, то здесь спектр исследований столь же обширен. Изучается влияние факторов риска (обсервационные, когортные исследования) или воздействия (клинические испытания) назначаемой терапии, влияние условий среды, особенностей популяции и т. п. на возникновение и/или течение заболевания, осложнения, качество анестезии, скорость реабилитации пациента после операции и т. д. Во всех таких исследованиях присутствуют две группы пациентов, иногда и больше, но никогда не меньше. В этих двух группах регистрируются идентичные показатели в одни и те же моменты времени (или приблизительно так, если изучаемый процесс допускает некоторую неточность по времени) и в результате исследователь получает две (или больше) выборки. В данной статье мы рассмотрим наиболее часто встречающийся и наиболее простой случай сравнения только двух групп пациентов между собой. Такой дизайн является стандартом для клинических рандомизированных контролируемых испытаний, которые обладают самым высоким уровнем доказательности, конечно, при условии минимальных отступлений от протокола их проведения.

Итак, в результате проведенного испытания или исследования доктор имеет две выборки, два набора чисел, которые он должен как-то сравнить. Остановимся на этом месте и зададимся вопросом: что описывают эти два набора чисел? Как правило, обе выборки представляют собой два множества значений изучаемого показателя, зарегистрированных у пациентов в основной группе (или группе исследования) и в группе сравнения (или группе контроля). Доктор выбрал этот показатель, поскольку предполагает, что он должен продемонстрировать различие между двумя группами наилучшим образом, наиболее ярко. Что это значит? Вероятно, именно эти два набора чисел должны более всего отличаться друг от друга при наличии эффекта от исследуемой процедуры (или фактора риска). Но что в данном случае означает их отличие? Мы легко можем понять, когда друг от друга отличаются два числа, т. е. когда они не равны друг другу, но что означает отличие двух выборок (наборов чисел), полученных в наших двух группах? Какие критерии сравнения здесь применимы и почему в результате мы имеем право аргументированно говорить о различии или идентичности этих двух групп по данному показателю? Здесь придется немного углубиться в теорию вероятностей. Дело в том, что любой показатель, который врач измеряет и регистрирует в ходе своего исследования, является случайной величиной. Случайная величина – это

не число и не одно значение, как можно заключить из ее названия. Случайная величина – это такой необычный математический объект, который обладает интервалом своих возможных значений и вероятностью появления каждого из этих значений при разовом измерении. Например, ЧСС молодого здорового человека в покое может быть от 60 до 80 ударов в минуту, но наиболее вероятно значение, допустим, 72. Если мы начнем измерять ЧСС у реальных добровольцев с соблюдением указанных условий (молодой, здоровый, в покое), то мы, конечно, не получим только значение 72, будут встречаться и другие, например: 65, 76, 69 и т. д., но при большом количестве таких добровольцев число 72 будет наиболее частым, а 60 или 80 будут регистрировать намного реже. Почему я утверждаю это с такой уверенностью, даже не проведя подобного опыта? Потому что ЧСС (как и большинство исследуемых в медицине измеряемых показателей) является случайной величиной, у нее есть интервал значений, и каждое число из этого интервала обладает определенной вероятностью, которая и отвечает за то, с какой частотой это значение встречается нам при массовом измерении этого показателя.

То, что в биомедицинских исследованиях мы всегда имеем дело со случайными величинами – большая удача, потому что это дает нам право использовать для обработки данных в таких исследованиях мощные методы теории вероятности и математической статистики. Ведь все эти методы разработаны исключительно для изучения и описания случайных величин самых разных типов. Как мы уже выяснили в прошлой статье [1], каждая случайная величина (СВ) имеет свои интегральные характеристики, которые ее полностью определяют и описывают. В список таких интегральных характеристик СВ входят, например, среднее значение, среднеквадратическое отклонение (но не ошибка среднего! [1]), медиана и некоторые другие. Каждое существенное, системное изменение СВ (иначе говоря, регистрируемого показателя) немедленно отражается на значениях его интегральных характеристик. Чем сильнее изменение, тем ярче выражено отличие интегральных характеристик. Теперь, возвращаясь к нашим выборкам, легко понять, что чем сильнее влияние исследуемого фактора, тем ярче отличие интегральных характеристик показателя при сравнении двух групп. Но, как мы помним, этих характеристик несколько. Какую из них взять за основу для выяснения вопроса о наличии или отсутствии различия? А кроме того, как определить, достаточно ли велико это различие, чтобы не считать его случайным, обусловленным лишь флуктуациями нашей выборки? [1–3] Наиболее легко воспринимаются и интерпретируются такие статистики, как среднее значение, среднеквадратическое отклонение и относительная частота. Мы подробно обсуждали их в предыдущей статье [1], поэтому повторю лишь основное. Среднее значение – это своего рода центр тяжести всей выборки, среднеквадратическое отклонение (еще раз, не путать с ошибкой среднего!) – это характеристика вариабельности показателя в данной группе, а относительная частота – это параметр, отражающий распределение вероятности по всему интервалу допустимых значений исследуемого показателя в данной выборке. Если полученные в исследовании данные распределены по нормальному закону, то два параметра – среднее значение и среднеквадратическое отклонение аккумулируют в себе всю информацию об этой выборке, всю без потерь. Математик, никогда не видевший результатов исследования, может в точности их воспроизвести, зная только среднее значение и среднеквадратическое отклонение полученной выборки, как если бы он сам получал эти данные в эксперименте. Но, повторюсь, такой фокус возможен только в случае распределения этих данных строго по нормальному закону. Итак, есть три основные характеристики выборки, которые выражены числами. Вычислив, например, среднее значение в одной группе и в другой, мы можем их сравнить. Как сравнить два числа, никому объяснять не надо. Глубокий смысл такого сравнения заключается в том, что это не просто какие-то числа, а своего рода полпреды двух выборок, а на самом деле даже целых популяций, которые аккумулируют в себе особенности последних. Сравнение трех названных статистик (среднего, среднеквадратического отклонения и частоты), полученных в группе исследования и группе контроля, покрывает 90% потребностей исследователя при доказательстве или опровержении межгруппового различия. В самом деле, посмотрим, какие изменения могут произойти с показателем – случайной величиной под воздействием изучаемого фактора:

– интервал значений может сдвинуться вправо/влево (рис. 1);

– интервал значений может стать шире/короче (рис. 2);

– вероятности по интервалу значений могут перераспределиться (рис. 3).

Все эти изменения продемонстрированы на рис. 1–3 по отдельности, хотя может случиться, и такие случаи не редкость, что изменение затронет одновременно все три позиции. Но все же, если рассматривать их по отдельности, то очевидно, что первое изменение затронет среднее значение, оно увеличится или уменьшится

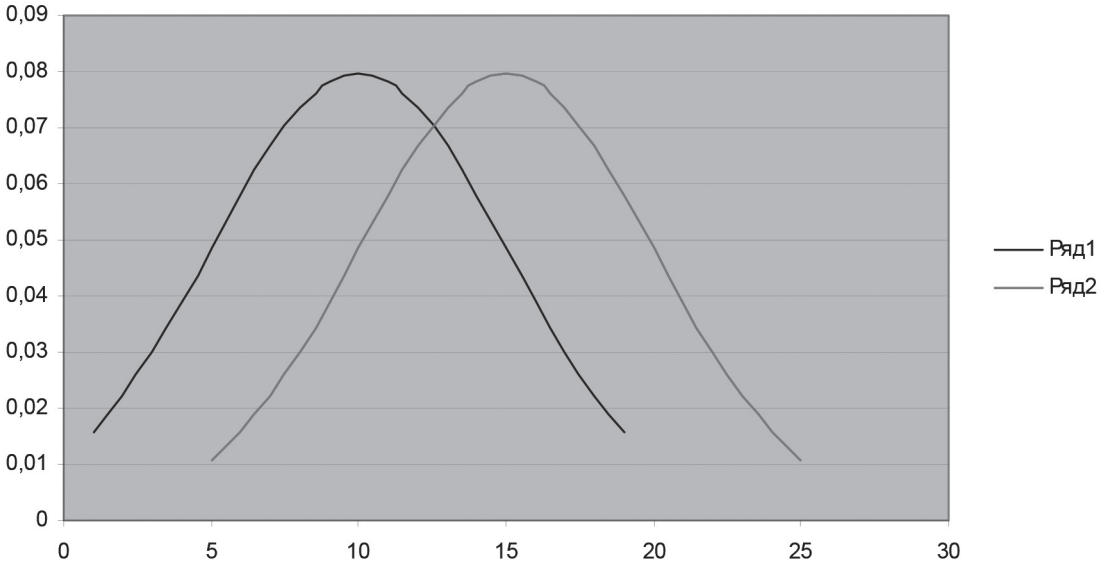

Рис. 1. Графическое отображение различия двух выборок по только средним значениям

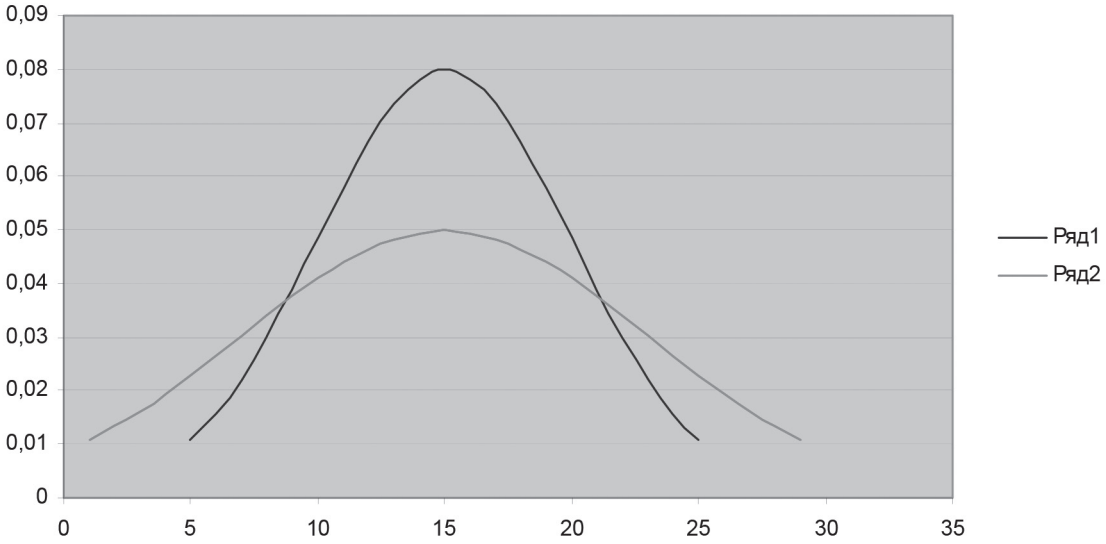

Рис. 2. Графическое отображение различия двух выборок по только среднеквадратическим отклонениям

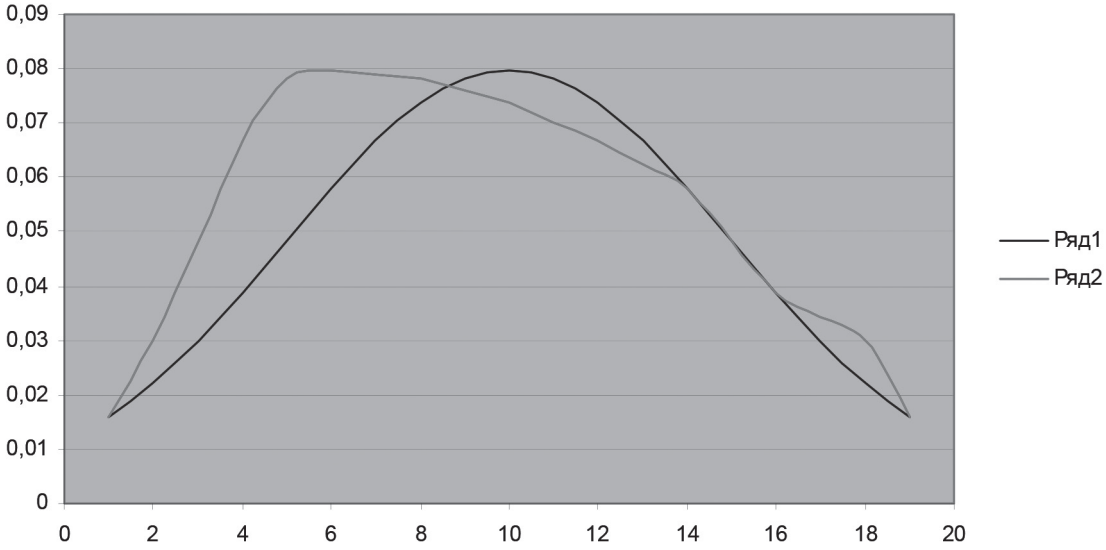

Рис. 3. Графическое отображение различия двух выборок по распределению частот

в зависимости от направления сдвига. Второе изменение проявится в характеристике вариабельности: если интервал удлинится, среднеквадратическое отклонение увеличится, если интервал станет короче, то уменьшится. Перераспределение частот по интервалу приведет к изменению относительных частот регистрируемых значений. Таким образом, выводы о межгрупповом различии исследуемого показателя мы можем строить на основе сравнения его средних значений и/или среднеквадратических отклонений и/или выборочного распределения частот (т. е. частот, рассчитанных в наших двух группах). Выбор критерия сравнения начинается именно с этого момента. Что для нас важно выяснить, какие изменения под влиянием исследуемого фактора для нас существенны или полезны с практической точки зрения, а на какие мы можем не обращать внимания? Отвечая на эти вопросы, мы вполне конкретно определяем выбор статистического критерия сравнения двух групп. Чаще всего для сравнения выбирают средние значения, потому что они наиболее показательны и легко интерпретируемы в клинической практике. Однако просто сравнить два средних значения, как обычные числа, мы не можем, поскольку эти величины, рассчитаны на выборках, и, следовательно, являются не точными значениями, а лишь выборочными оценками точных средних значений в одной и во второй группе [1]. Для сравнения выборочных оценок разработаны специальные статистические критерии достоверности различий. В таблице 1 кратко описаны 3 таких критерия, которые используются наиболее часто при сравнении двух выборок. Весь алгоритм сравнения двух выборок с помощью одного из указанных критериев заключается в следующих шагах:

1. Выбираем тот статистический параметр, по которому будем сравнивать наши выборки

– среднее значение (изменение расположения центра тяжести выборки) или

– среднеквадратичное отклонение (изменение вариабельности показателя) или

– частоту (изменение закона распределения частот).

2. Абстрагируясь от содержательной стороны вопроса, уже исключительно в терминах математической статистики четко формулируем гипотезу, которую будем проверять. Это будет наша, так называемая, нуль-гипотеза H_0. Если мы выбрали для сравнения средние значения, то нуль-гипотеза будет формулироваться так: X1 = X2, где X1 и X2 – средние, полученные в 1-й и 2-й группах, соответственно. Если мы сравниваем среднеквадратические отклонения, то H_0: S1=S2, где S1 и S2 – указанные параметры из 1-й и 2-й групп. Обратите внимание, с помощью статистического критерия мы проверяем не ту гипотезу, которая сформулирована в цели нашего исследования и даже не то предположение об эффективности терапии, которое описываем в терминах предметной области, а именно статистическую гипотезу о равенстве двух средних (или любого другого статистического параметра, выбранного для сравнения). Как мы их рассчитали, какие это средние, какого показателя – на данный момент это совершенно неважно. Мы перешли на территорию математической статистики и в ее пределах имеет значение только то, что это два средних значения какой-то случайной величины, мы предполагаем, что они равны и хотим доказать или опровергнуть наше предположение.

3. Даже в обычной жизни мы далеко не всегда уверены в своем доказательстве на все 100%. В этом смысле статистика очень близка к нашему жизненному опыту [2, 3]. Разница состоит лишь в том, что в статистике требуется точно указать меру нашей уверенности, с которой мы будем считать нашу

Таблица 1. **Наиболее часто применяющиеся критерии сравнения двух выборок**

Название критерия	Сравниваемые статистические параметры	Проверяемая статистическая гипотеза (H_0)	Формула расчета	Комментарии
Т-критерий Стьюдента	Средние значения двух выборок	X1 = X2, где X1, X2 – средние значения	$T = (X1–X2)/(S^2/N)^{1/2}$	Применяется только для однократного сравнения двух групп
F-критерий Фишера	Среднеквадратические отклонения двух выборок	S1 = S2, Где S1, S2 – среднеквадратические отклонения	$F = S1^2/S2^2$	Критическое значение зависит от числа степеней свободы
Критерий согласия Пирсона (хи-квадрат)	Выборочные частоты	f1(x) = f2(x), где f1(x), f2(x) – дискретные функции распределения частот в группах 1 и 2	$\chi^2 = \Sigma[(f1–f2)^2/f1]$	Используется для сравнения именно функций распределения частот двух групп, а не каждой частоты в отдельности

гипотезу доказанной. Это число заключено между 0 и 1 (или в процентах между 0 и 100), но никогда не достигает пределов своего допустимого интервала, т. е. оно всегда больше 0 и меньше 1. Чтобы разобраться в этой проблеме, рассмотрим все возможные варианты результатов нашего тестирования (табл. 2). Итак, реальность предлагает нам только две ситуации: либо средние равны, либо нет. Наше решение тоже ограничено лишь двумя вариантами: средние равны или не равны. Если в действительности средние равны, и мы принимаем нуль-гипотезу, то мы правы, однако если мы при равенстве средних отвергаем нуль-гипотезу, то мы совершаем ошибку. Такая ошибка в эхолокации называется «ложная тревога» или ошибка первого рода и обозначается греческой буквой α. Однако это еще не все неприятности, которые ожидают нас на пути принятия решения по поводу нашей нуль-гипотезы. Если в действительности средние не равны, а мы считаем, что они равны, мы совершаем другую ошибку, «пропуск сигнала», ошибку второго рода, обозначаемую β. Конечно, нам бы хотелось свести к минимуму обе ошибки, но, к сожалению, это невозможно, так как они зависят друг от друга достаточно неприятным образом: чем меньше одна, тем больше вторая. Поэтому мы должны найти некоторый баланс между ними, который бы нас устроил. В биомедицинских исследованиях обычно более строго относятся к вероятности «ложной тревоги», предел этой ошибки принимают значительно меньше, чем второй, как правило, не выше 0,05. Этот предел называется уровнем значимости. Если в результате расчетов мы получили значение ошибки первого рода (α) больше, чем 0,05, то мы не отвергаем нуль-гипотезу, т. е. считаем, что различия между группами по данному показателю статистически не доказаны. Если же полученная ошибка меньше заданного уровня значимости, то различие между группами считается статистически достоверным. В этом случае надо обязательно указать вероятность «ложной тревоги», т. е. вероятность того, что мы приняли различие двух средних, тогда как они на самом деле они равны. Итак, если в результате применения, например, критерия Стьюдента для сравнения двух средних мы получили ошибку первого рода, ее обозначают *p*, равную 0,02, а уровень значимости в своем исследовании мы задали равным 0,05, то мы имеем право утверждать, что на уровне значимости, равном 0,05, средние исследуемого показателя наших двух групп статистически значимо (или достоверно) различаются. При этом надо очень четко осознавать 2 важнейших момента:

– Мы не говорим, что два средних точно не равны, т. к. мы на самом деле этого не знаем на 100%. Просто мы установили некоторый порог вероятности ошибки (уровень значимости), рассчитываемой в ходе применения критерия, ниже которого условились считать, что эти средние значения не равны или, другими словами, статистически значимо различаются.

– Когда мы принимаем решение, что наши средние не равны, мы имеем в виду, что мы можем ошибаться, и вероятность этой ошибки составляет *p*, которое мы получили, т. е. в нашем примере 0,02, следовательно, наша уверенность в своей правоте составляет 98% (100%–0,02*100%).

Если же мы получили ошибку *p* больше принятого уровня значимости, то мы принимаем нуль-гипотезу, т. е. считаем, что наши средние статистически достоверно не различаются. Опять же это не значит, что они точно равны, мы вполне можем ошибаться, и такая ошибка, как было сказано выше, называется ошибкой второго рода – пропуск сигнала или в нашем случае пропуск различия. Вероятность этой ошибки в исследованиях не указывается явно, но о ней следует обязательно помнить, утверждая, что различий между группами нет. Величина ошибки второго рода составляет до 15–20% при выборке среднего объема. Эта ошибка снижается с ростом числа пациентов в обеих группах, но здесь не все так просто. Если у вас в исследовании группы не равны по численности, что особенно характерно для обсервационных исследований, изучающих какие-либо редкие события (осложнения, состояния, патологии), то ошибка β будет зависеть от выборки меньшего объема. Это значит, что при очень маленькой основной группе, которую увеличить достаточно сложно, потому что исследуемое заболевание встречается редко, ошибка «пропуска сигнала» может быть довольно большой, несмотря на то, что контрольная группа может включать достаточно много пациентов. Кроме того, необходимо помнить, что, начиная с некоторого объема выборки, уменьшение ошибки второго рода дается с большим трудом и фактически трудозатраты на ее снижение просто не оправдываются, порождая настоящий «девятый вал» чисел, не приносящих существенных уточнений в расчеты. Не стоит очень сильно увлекаться ее снижением.

Итак, все выше сказанное подводит нас к тому, что еще до расчета критерия, мы должны

Таблица 2. **Вероятности ошибок первого и второго рода при проверке гипотезы о равенстве двух средних значений**

		Действительность	
		X1 = X2	***X1 ≠X2***
Решение исследователя	X1 = X2	1 - α	b
	X1 ≠ X2	a	1 - β

установить порог ошибки первого рода – уровень значимости, который позволит нам однозначно принимать решения о межгрупповом равенстве средних (или других статистических параметров). Если в результате применения критерия мы получаем ошибку *p* менее установленного уровня значимости, то мы считаем, что средние не равны и две группы различны. При этом мы помним, что можем ошибаться, и вероятность этой ошибки равна *p*. Если же полученное *p* больше уровня значимости альфа, то мы считаем, что средние равны (две выборки взяты из одной популяции), но и тут мы не можем дать никаких 100%-ных гарантий нашему заключению. Вероятность того, что мы, принимая это решение, совершаем ошибку составляет, допустим, 15% (или то же самое по-другому $\beta = 0{,}15$). Величина (1-бета) называется мощностью критерия.

Еще раз подчеркну, что уровень значимость альфа должен быть установлен до того, как мы начнем вычислять значение критерия по его формуле.

4. Когда выбран статистический параметр для сравнения двух групп, четко сформулирована статистическая гипотеза и принят уровень значимости (максимальное значение ошибки первого рода), то по определенной формуле, которую диктует выбранный тест, вычисляется значение статистического критерия. В этой формуле обязательно участвуют выбранные для сравнения статистические параметры, рассчитанные в обеих выборках. В результате вычисления критерия получается некоторое число. По данному числу раньше с помощью специальных таблиц, сейчас чаще всего с использованием специализированных компьютерных программ, определяется ошибка *p*.

5. Следующий шаг – просто сравнить полученную ошибку с заранее заданным уровнем значимости и принять решение о статистической значимости различия средних или других статистических параметров, которые были выбраны для сравнения двух групп.

Теперь, когда мы знаем, что при сравнении двух групп должны помнить не только об ошибке *p*, но и об ошибке второго рода β, которую в научных публикациях указывают редко, становится понятно, что отсутствие различия, полученное на маленьких выборках, даже при проведении по всем правилам клинических испытаний, еще не доказывает однозначно, что различия действительно нет. У исследований с выборками маленького объема достаточно велика вероятность ошибки второго рода β, «пропуск сигнала», особенно, если величина различия (эффекта) небольшая. Необходимо обязательно помнить об этой второй ошибке и в случае утверждения об отсутствии эффекта обязательно обращать внимание на объем выборок, на основе которых делается это заключение. Наилучший способ усилить доказательную базу вывода об отсутствии различия между группами – это провести несколько исследований и убедиться, что результаты воспроизводимы. Если они будут отличаться, то процедура метаанализа поможет принять сбалансированное объективное решение в такой неоднозначной ситуации.

Необходимо добавить, что если исследователь хочет доказать наличие определенного эффекта (конкретную величину различия) на заданных фиксированных уровнях обеих ошибок, то ему надо воспользоваться процедурами расчета соответствующих объемов выборок, которых будет достаточно, чтобы решить эту задачу. В следующей публикации мы подробно остановимся на этой проблеме и ее решении.

Литература

1. *Тихова Г. П.* Значение и интерпретация ошибки среднего в клиническом исследовании и эксперименте. *Регионарная анестезия и лечение острой боли* 2013; 3:50–53
2. *Тихова Г. П.* Четырехпольная таблица – Бритва Оккама в мире статистики. Часть 1. Как рассчитать относительный риск и другие параметры из четырехпольной частотной таблицы? *Регионарная анестезия и лечение острой боли* 2013; 4:69–75
3. *Тихова Г. П.* Четырехпольная таблица – Бритва Оккама в мире статистики. Часть 2. Как увидеть лес за деревьями? *Регионарная анестезия и лечение острой боли* 2013; 4:54–60

References

1. *Tikhova G. P.* Importance and interpretation of standard error of mean in clinical study and trial. *Regionarnaya anesteziya i lechenie ostroy boli.* 2013; 3:50–53 (In Russia).
2. *Tikhova G. P.* Fourfold frequency table – Occam's Razor in the world of statistics. Part 1. Calculating relative risk and other parameters from fourfold frequency table. *Regionarnaya anesteziya i lechenie ostroy boli.* 2013; 4:69–75.
3. *Tikhova G. P.* Fourfold frequency table – Occam's Razor in the world of statistics. Part 2. How can we see the forest among the trees? *Regionarnaya anesteziya i lechenie ostroy boli.* 2013; 4:54–60.