Британский журнал математической и статистической психологии (2014), 67, 213–230 ©
Британское психологическое общество, 2013 г.

The British Psychological Society

www.wileyonlinelibrary.com

Оценка клинической значимости: включение надежных статистических данных с нормативными сравнительными тестами

Катрина ван Виринген и Роберт А. Крибби* Факультет психологии, Йоркский университет, Торонто, Канада

Цель этого исследования состояла в том, чтобы оценить модифицированный тест эквивалентности для проведения нормативных сравнений, когда формы распределения ненормальны, а дисперсии неравны. Исследование Монте-Карло использовалось для сравнения эмпирической частоты ошибок типа I и мощности предложенного теста эквивалентности Шуирмана-Юэна, в котором используются усеченные средние, с ранее рекомендованными тестами эквивалентности Шуирмана и Шуирмана-Уэлча, когда предположения о нормальности и дисперсионной однородности выполняются, а также когда они не выполняются. Эмпирические коэффициенты ошибок типа I по шкале Шуирмана-Юэна были намного ближе к номинальному уровню а, чем по шкале Шуирмана. или критерий Шуирмана-Уэлча, а мощность критерия Шуирмана-Юэна была значительно выше, чем у критериев Шуирмана или Шуирмана-Уэлча, когда распределения были асимметричными или присутствовали выбросы. Тест Шуирманна-Юэна рекомендуется для оценки клинической значимости с нормативными сравнениями.

1. Введение

В области клинической психологии большое количество исследований посвящено оценке эффективности различных вмешательств. Важным аспектом эффективности вмешательства является клиническая значимость, которую можно определить как практическую или прикладную значимость вмешательства. Эта важность обычно описывается с точки зрения того, действительно ли вмешательство меняет повседневную жизнь клиента или людей, которые взаимодействуют с клиентом, или способно ли вмешательство вернуть клиента в состояние нормального функционирования. Следуя многочисленным рекомендациям по отчетности о клинической значимости (например, Kendall, 1997), исследователи-клиницисты начинают делать измерения клинической значимости важной частью своих клинических интервенционных исследований (например, S anchez-Ortuno & Edinger, 2010; Wallach, Safir). и Бар-Цви, 2009).

Было предложено множество методов для оценки эффективности вмешательства; однако есть несколько проблем с этими методами. Например, одна из проблем заключается в том, что большинство методов не анализируют, возвращается ли обработанная группа к состоянию нормального функционирования. Другими словами, традиционные методы оценки вмешательств могут оценить, претерпела ли лечащая группа существенные изменения, но могут не указать, насколько сильно изменилась группа, что означает это изменение или вернуло ли это изменение клиентов в группе к прежним изменениям. уровень нормального функционирования. Вторая проблема заключается в том, что многие методы оценивают клиническую значимость на индивидуальном, а не на групповом уровне.

*Корреспонденцию следует направлять Роберту А. Крибби, факультет психологии, Йоркский университет, Торонто, Онтарио, M3J 1P3, Канада (электронная почта: cribbie@yorku.ca).

DOI: 10.1111/bmsp.12015

что может затруднить глобальную оценку эффективности вмешательства. Например, какая доля улучшенных клиентов укажет на то, что вмешательство было эффективным?
Хотя оценки на индивидуальном уровне важны для клиницистов, чтобы иметь возможность идентифицировать клиентов с экстремальной реакцией на лечение, исследователи, изучающие интервенционные исследования, часто интересуются глобальными оценками эффективности вмешательства. Еще одна важная проблема заключается в том, что большинство методов игнорируют тот факт, что распределение баллов в группах лечения, контроля и нормального сравнения часто не является нормальным, и что дисперсии часто сильно различаются между этими группами.

Один из наиболее многообещающих недавних методов оценки клинической значимости включает в себя оценку эквивалентности группы пациентов, прошедших лечение, и нормальной группы сравнения по интересующему исходу (например, симптомам депрессии) (Kendall, Marrs-Garcia, Nath & Sheldrick, 1999).). Преимущества этого подхода заключаются в том, что оценки проводятся на групповом уровне, а метод непосредственно решает вопрос о том, вернулись ли клиенты в группе к состоянию нормального функционирования. Цель этого исследования состояла в том, чтобы исследовать усовершенствования существующей тестовой статистики для оценки эквивалентности группы пролеченных клиентов и нормальной группы сравнения для ситуаций, в которых распределения групп ненормальны и/или дисперсии групп неравный.

2. Традиционные методы определения эффективности вмешательства

Традиционно статистический анализ интервенционных исследований включал только сравнение данных до и после лечения, чтобы определить, было ли лечение причиной наблюдаемых изменений. Эти сравнения обычно проводились по отношению к контрольной группе.
Традиционные тесты статистической значимости, например, критерий Стьюдента для двух связанных выборок или критерий ANOVA F, используются для сравнения данных экспериментальной группы или экспериментальной и контрольной групп до и после лечения. Использование традиционных тестов статистической значимости для оценки эффективности лечения ограничено тем, что не дает информации о силе взаимосвязи или о том, имеет ли она клиническое значение или значимость, например, возвращаются ли клиенты к состоянию нормального функционирования (Jacobson & Truax). , 1991; Кремер, Морган, Лич, Глинер, Васке и Хармон, 2003).

Показатели величины эффекта (например, коэффициент Коэна d) могут использоваться для помощи в интерпретации практической значимости, поскольку они обеспечивают меру силы взаимосвязи. Однако Кремер и соавт. (2003) обсуждают важное ограничение величины эффекта, связанное с их способностью действовать как мера клинической значимости. Проблема в том, что величины эффекта нельзя интерпретировать с точки зрения того, насколько люди затронуты лечением, потому что изначально они не были разработаны как показатели клинической значимости. Однако более современные показатели величины эффекта для клинических вмешательств, например, разница показателей успеха (SRD), обеспечивают более подходящие показатели эффекта для рандомизированных исследований. SRD представляет собой разницу между вероятностью того, что у клиента в лечебной группе результат лечения предпочтительнее, чем у клиента в контрольной группе, и вероятностью того, что у клиента в контрольной группе результат лечения предпочтительнее, чем у клиента в лечебной группе. Кремер и Купфер, 2006). Точнее говоря, хотя величины эффекта оценивают силу ассоциации (например, насколько больше улучшилось состояние экспериментальной группы по сравнению с контрольной группой), они не говорят нам, вернулись ли клиенты к состоянию нормального функционирования.

Популярные методы оценки клинической значимости, такие как метод Якобсона и Труакса (1991), измеряют изменения на индивидуальном уровне путем определения того,

пролеченные клиенты выходят за пределы дисфункциональной популяции или в пределах функциональной популяции. Первым шагом в подходе Джейкобсона и Труакса является расчет точки отсечения для клинически значимого изменения, которая представляет собой точку, которую клиент должен пересечь во время оценки после лечения, чтобы быть классифицированным как изменение на клинически значимое. значительная степень. Джейкобсон и Труакс предложили три способа расчета порогового показателя: (1) уровень функционирования после вмешательства должен выходить за пределы диапазона неблагополучной популяции, где диапазон определяется как расширение до двух стандартных отклонений за пределы среднего значения неблагополучное население; (2) уровень функционирования после вмешательства должен находиться в пределах диапазона функциональной популяции, где диапазон определяется в пределах двух стандартных отклонений от среднего значения этой популяции; или (3) уровень функционирования после вмешательства ставит этого клиента ближе к среднему значению функциональной популяции, чем к среднему значению неблагополучной популяции. Вторым шагом является расчет индекса надежных изменений (RCI), чтобы определить, является ли переход клиента от предварительного теста к послетестовому надежному и не связан ли он с ошибкой измерения, поскольку послетестовые оценки могут пересечь границу. -off point еще не будет статистически надежным.

$$RCI\ ¼\ \frac{x2\ x1}{Сдифф};$$

Здесь x1 представляет собой оценку клиента до теста, x2 представляет собой оценку того же клиента после теста, а Sdiff представляет собой стандартную ошибку разницы между двумя оценками теста. Sdiff это традиционно рассчитываемое как spre(1 r)1/2, где spre — дотестовая оценка, стандартное отклонение, а r — внутренняя согласованность или надежность повторных тестов дотестовой оценки (см. Martinovich, Saunders & Howard, 1996). , преимущества и недостатки каждого).

Есть пара важных критических замечаний, связанных с использованием метода Джейкобсона и Труакса. Во-первых, неясно, насколько устойчив метод, когда нарушается допущение о том, что дисфункциональное и функциональное распределения являются нормальными, поскольку метод предполагает нормальное распределение. Это особенно важно, так как несколько предыдущих обзоров показали, что распределения в психологии редко бывают нормальными (например, Micceri, 1989). Во-вторых, этот метод предназначен для изучения клиентов на индивидуальном уровне (т. е. каждый человек исследуется отдельно), что затрудняет общие утверждения об эффективности вмешательства. Во многих случаях исследователи будут вычислять долю людей, которые выздоровели, улучшились, не изменились или ухудшились, а другие могут даже сравнить пропорции в каждой категории; однако эти тесты не дают прямого ответа на вопрос о том, эквивалентны ли результаты ранее клинической группы по сравнению с нормальной контрольной группой.

3. Нормативные сравнения

Нормативные сравнения представляют собой процедуру оценки клинической значимости терапевтических вмешательств. Например, представьте, что вмешательство, как было показано, дает значительно большее улучшение, чем плацебо, другое лечение и т. д., но неясно, насколько хорошо функционируют пролеченные клиенты по сравнению с нормальной контрольной группой. Нормативные сравнения включают сравнение среднего значения группы, получавшей лечение (т. е. клинической группы после вмешательства), со средним значением нормальной группы сравнения, чтобы определить клиническую значимость исследования. В частности, цель исследования состоит в том, чтобы определить, имеют ли две группы одинаковые баллы по

интересующая мера или поведение. Кендалл и др. (1999) поднимают два важных вопроса, связанных с клинической значимостью вмешательства. Во-первых, достаточно ли велико количество изменений, произошедших в результате лечения, чтобы их можно было считать значимыми? Во-вторых, отличимы ли обработанные люди от нормальных людей? Нормативные сравнительные тесты решают эти вопросы на групповом уровне. Этот метод позволяет оценить эффективность вмешательства по эталону, не зависящему от лиц с исходным расстройством. Однако, поскольку статистический анализ включает в себя демонстрацию эквивалентности, а не различий между обработанными и нормальными группами сравнения, требуются специальные методы проверки эквивалентности.

Тем не менее, есть несколько важных предостережений относительно нормативных сравнений, прежде чем подробно обсуждать методы. Во-первых, методы, описанные в этой статье, предназначены для вмешательств, реальной целью которых является возвращение клиентов в состояние функционирования, находящееся в пределах нормы. Хотя это не относится ко всем вмешательствам (например, лечение поведенческих проблем у аутистов), мы полагаем, что это относится ко многим вмешательствам. Как обсуждал анонимный рецензент этой статьи, использование психиатрических препаратов (например, антидепрессантов) также может быть аспектом вмешательства, которое помогает вернуть поведение людей в состояние нормального функционирования; нормативные сравнения все еще были бы полезны в этих условиях, если только группа, получавшая лечение, не принимала лекарство на краткосрочной основе. Второе предостережение, также выдвинутое анонимным рецензентом, заключается в том, что, поскольку нормативные сравнения используют только меру поведения в один момент времени, они не могут учитывать возможность рецидивов в будущем. Наконец, следует также подчеркнуть, что популяция сравнения (обычно называемая нормальной группой сравнения) должна быть выбрана таким образом, чтобы максимально точно соответствовать соответствующим характеристикам группы, подвергаемой лечени Например, представьте, что группа людей, у которых диагностировано большое депрессивное расстройство, лечилась когнитивно-поведенческой терапией. Если исследователь хотел провести нормативные сравнения, было бы неуместно сравнивать выборку сообщества, получавшего лечение, с, например, нормативной популяцией, состоящей из студентов университета, потому что хорошо известно, что последние имеют более высокие, чем обычно, показатели депрессии (и, следовательно, леченные). выборка может показаться эквивалентной университетской выборке по депрессии, даже если они не эквивалентны выборке сообщества по депрессии). Без сопоставимой нормативной группы сделанные сравнения могут ввести в заблуждение.

4. Тестирование

эквивалентности Тестирование эквивалентности – это статистический метод, часто используемый в биофармацевтических исследованиях для определения эквивалентности двух экспериментальных препаратов. Роджерс, Ховард и Весси (1993) объясняют, что методы проверки эквивалентности можно использовать для оценки многих важных гипотез в психологии и смежных областях. В недавней литературе (например, Cribbie & Arpin Cribbie, 2009; Gruman, Cribbie & Arpin-Cribbie, 2007; Rogers et al., 1993; Seaman & Serlin, 1998) подчеркивается, что традиционные тесты, основанные на различиях, не подходят, когда целью является чтобы определить, эквивалентны ли две группы по переменной результата, и поощряет использование тестов эквивалентности для оценки этих типов гипотез. Когда исследователь пытается продемонстрировать эквивалентность групп с помощью традиционного критерия значимости, он или она часто находится в ситуации, когда он или она пытается подтвердить, а не отвергнуть нулевую гипотезу. Широко известно, что при достаточно большом размере выборки даже незначительные различия будут статистически значимыми (Rogers et al., 1993). Далее

размеры малы, средства почти всегда будут объявлены эквивалентными (т. е. нулевой гипотеза об отсутствии различий редко отвергается).

Нулевая гипотеза тестов эквивалентности утверждает, что разница между группами выходит за пределы интервала эквивалентности, указанного исследователем, и альтернатива гипотеза утверждает, что разница между группами попадает в указанные интервал (Rogers et al., 1993). Шуирманн (1987) предложил проверку эквивалентности метод, в котором исследователь должен сначала определить интервал эквивалентности, а затем выполнить две одновременные односторонние проверки гипотез. При определении эквивалентности интервал, необходимо учитывать, какая наименьшая значимая разница будет дана характер исследования. Диапазон должен быть ограничен двумя значениями: нижнее значение является отрицательная дельта (d), а верхнее значение – положительная дельта (d). Выбор большого d будет повышает вероятность объявления групп равнозначными, но в то же время снижает вероятность того, что различия между группами можно считать бессмысленными. А меньшее d затрудняет установление эквивалентности двух групп, но в то же время время есть больше уверенности в заявлениях об эквивалентности (Cribbie & Arpin Cribbie, 2009). Пусть l1 и l2 представляют два сравниваемых средних значения генеральной совокупности, и пусть d представляет собой наименьшую разницу между средними значениями, которые можно было бы рассматривать важный. Статистические гипотезы определяются как:

$$H01 : l1 \; l2 \; д; \; H02 : l1 \; l2 \; д;$$
$$Ha1 : l1 \; l2 \backslash d; \; Ha2 : l1 \; l2 \; [\; д:$$

Для установления эквивалентности средств проводятся два одновременных односторонних теста. используется для тестирования H01 и H02. В тесте 1 цель состоит в том, чтобы отвергнуть нулевую гипотезу, утверждающую, что разница между средними больше или равна d. В тесте 2 цель состоит в том, чтобы отклонить нулевая гипотеза, утверждающая, что разница меньше или равна d. H01: л1 л2 д отвергается, если t1 ta,df, где

$$t1 \; ¼ \; \frac{ðM1 \; M2Þ \; д}{д \; \frac{ðn1þn2Þ½ðn1 \; 1ÞS2 \; 1þðn2 \; 1ÞS2}{n1n2ðn1þn2 \; 2Þ}};$$

и H02: л1 л2 d отвергается, если t2 та,дф, где

$$т2 \; ¼ \; \frac{ðM1 \; M2Þ \; ð \; dÞ}{д \; \frac{ðn1þn2Þ½ðn1 \; 1ÞS2 \; 1þðn2 \; 1ÞS2}{n1n2ðn1þn2 \; 2Þ}};$$

M1 и M2 — групповые средние значения, n1 и n2 — объемы групповой выборки, s1 и s2 — групповые стандартные отклонения, а ta,df — верхнехвостое критическое значение a-уровня t с n1 + n2 2 степени свободы. Важно отметить, что альтернативный метод проведение этих анализов заключается в том, чтобы определить, является ли доверительный интервал 1 2a полностью содержится в пределах интервала эквивалентности. Другими словами, отклоните H01 и H02 , если:

$$M1 \; M2 \; t2a;df \; c \; \frac{ðn1 \; þ \; n2Þ½ðn1 \; 1ÞS2 \quad 1 \; þ \; Pn2 \; 1ÞS2 \quad 2}{n1n2ðn1 + n2 \; 2Þ};$$

полностью содержится внутри (d, d).

Для целей этой статьи основное внимание в тестах эквивалентности уделяется тому, является ли вмешательство эффективным, путем сравнения результатов после тестирования группы, получавшей лечение, с нормальной группой сравнения. Использование тестов эквивалентности для сравнения обработанных и нормативных популяций называется «нормативными сравнениями», и эти сравнения описаны в следующем разделе.

5. Нормативная процедура сравнения Кендалла Первый шаг в

методе Кендалла и соавт. (1999) нормативный подход к сравнению заключается в определении диапазона близости, в пределах которого две группы будут считаться клинически эквивалентными (т. е. интервал эквивалентности). Кендалл и др. (1999) отмечают, что, хотя одно стандартное отклонение может быть ориентиром для опубликованных норм, определенный диапазон может быть изменен по-разному в зависимости от конкретного сравнения. Например, при выборе d можно руководствоваться пороговыми значениями из опубликованных показателей, процентами от среднего значения в группе здоровых или получавших лечение или оценками размера эффекта различий, которые не являются клинически значимыми (Kendall et al., 1999). Крибби и Арпин-Крибби (2009) обсуждают, как в случаях, когда у исследователя нет четко определенного интервала эквивалентности, единственное значение d не позволяет ему или ей количественно оценить уровень близости, установленный терапией. Крибби и Арпин-Крибби (2009) предлагают в тех случаях, когда исследователи, оценивающие эквивалентность обработанных и нормативных групп, имеют мало информации для выбора подходящего единственного интервала, использовать несколько интервалов эквивалентности. В качестве рекомендаций предлагаются следующие уровни d: (1) d = 0,5s в норме; (2) d = нормальный; и (3) d = 1,5snormal, где snormal — стандартное отклонение нормальной группы сравнения. Опять же, уровни d следует выбирать в зависимости от характера конкретного исследования, поскольку трудности, связанные с возвращением поведения к нормальному функционированию, сильно различаются от поведения к поведению и от исследования к исследованию (Крибби и Арпин-Крибби, 2009). Крибби и Арпин-Крибби также обсуждают важность проведения предварительного теста, который оценивает, отличается ли клиническая группа в предварительном тесте от нормальной группы сравнения (с использованием традиционного теста, основанного на различиях), поскольку, если группы не отличаются до вмешательства, ценность демонстрации их эквивалентности после вмешательства существенно снижается.

Следующим шагом является проведение теста на эквивалентность, чтобы определить, можно ли считать обработанную и нормальную группы сравнения эквивалентными. При проверке эквивалентности двумя популярными подходами являются оригинальная процедура Schuirmann (1987) с двумя односторонними тестами (TOST) (представленная выше) и тест Schuirmann-Welch (Dannenberg, Dette & Munk, 1994; Gruman et al., 2007). Первоначальная процедура Шуирмана предполагает, что обработанная и нормальная сравниваемые дисперсии генеральной совокупности равны, тогда как Шуирманн-Уэлч не требует, чтобы дисперсии генеральной совокупности были равными. Хотя Кендалл и соавт. (1999) и другие (например, Golinski & Cribbie, 2009; Keselman et al., 1998) показали, что в психологических исследованиях дисперсии населения редко бывают одинаковыми, Kendall et al. представьте гомоскедастическую процедуру TOST, предложенную Schuirmann (1987) в качестве метода проведения нормативных сравнений.

Данненберг и др. (1994) предложили модификацию исходного критерия эквивалентности Шуирмана, которая включала гетероскедастическую стандартную ошибку и степени свободы, предложенные Уэлчем (1938) и Саттертуэйтом (1946). Эта модификация была сделана потому, что исходный критерий эквивалентности Шуирмана использует ту же стандартную ошибку и степени свободы, что и t-критерий для независимых выборок, и, таким образом, проблемы размера выборки и неравенства дисперсии, которые влияют на t-критерий для независимых выборок, также влияют на критерий Шуирмана.

тест эквивалентности (Cribbie & Arpin-Cribbie, 2009). Было обнаружено, что эмпирические коэффициенты ошибок типа I для теста эквивалентности Шуирманна существенно отклоняются от номинального уровня, когда размеры выборки и дисперсии не равны (Gruman et al., 2007). критерий эквивалентности Шуирмана–Уэлча, H01: l1 l2 ta;dfW d одностороннего теста эквивалентности Уэлча, dfW ; H02: l1 l2, где

$$= \frac{\delta(M1 - M2) - d}{\sqrt{\frac{c2}{n1} + \frac{c2}{n2}}};$$

$$\frac{\delta(M1 - M2) - \delta - d}{\sqrt{\frac{c2}{n1} + \frac{c2}{n2}}};$$

а также

$$dfW = \frac{\left(\frac{c2}{n1} + \frac{c2}{n2}\right)^2}{\frac{c4}{n2\,1(n1 - 1)} + \frac{c4}{n2\,2(n2 - 1)}}$$

6. Пример нормативных сравнений

Manzoni, Cribbie, Villa, Arpin-Cribbie, Gondoni and Castelnuovo (2010) использовали обсервационное предварительное и последующее исследование для изучения эффективности 4-недельной программы кардиореабилитации для улучшения психологического благополучия пациентов с ожирением. . Пациенты заполняли опросник общего психологического благополучия (PGWBI) при поступлении и при выписке. Исследователи использовали нормативную процедуру сравнения, чтобы определить клиническую значимость 4-недельной программы проживания. Манцони и др. выдвинули гипотезу о том, что пациенты с ожирением и сердечными заболеваниями будут иметь более низкие баллы, чем обычно, по показателю качества жизни на исходном уровне и что кардиореабилитационное вмешательство будет эффективным для улучшения психического здоровья до нормального уровня. Проверка эквивалентности с нормативными сравнениями использовалась для оценки того, была ли 4-недельная программа реабилитации эффективной для улучшения нарушенных параметров PGWBI до нормаль Результаты показали, что средние баллы пациентов по шкалам PGWBI, которые были снижены на исходном уровне по сравнению с нормой (общий балл, самоконтроль и жизнеспособность), были эквивалентны нормальным средним значениям при выписке. Даже на уровне подгруппы шкалы PGWBI, которые были нарушены в начале исследования, значительно улучшились до нормального уровня в конце вмешатель

7. Улучшение нормативных сравнений

Важным вопросом при оценке эквивалентности обработанных и нормальных групп сравнения является то, что распределения групп часто не являются нормальными и/или содержат посторонние случаи. Например, при лечении клинических популяций часто у большей части группы наблюдается улучшение, но у небольшой части группы либо не улучшается, либо даже ухудшается состояние. Это может привести к отрицательно асимметричному распределению с побочными случаями. Кроме того, нормальные оценки группы сравнения по популярным психологическим шкалам, измеряющим депрессию, тревогу и т. д., часто дают сильно перекошенные распределения с положительными отклонениями в верхней части хвоста. Это всего лишь несколько примеров распределений, которые часто бывают искажены в клинических исследованиях.

Несмотря на то, что тест Шуирмана – Уэлча устойчив к нарушениям дисперсии однородности, неясно, насколько надежным будет этот тест, когда распределения искажены или содержат выбросы. Например, как t Стьюдента (который предполагает равные дисперсии) и t Уэлча (который не предполагает равных дисперсий) плохо контролируют Частота ошибок типа I, когда распределения ненормальны, а дисперсии неравны (Кесельман, Отман, Уилкокс и Фрадетт, 2004 г.). Популярный подход к работе с ненормальными распределениями и/или выбросами заключается в удалении нетипичных значений путем обрезки. данные. Уилкокс (1997), Кесельман и др. (2004) и другие указывают, что показатели Типа I ошибка и способность обнаруживать эффекты гораздо менее подвержены влиянию, когда усеченные средние заменены обычные групповые средства. Прошлые исследования рекомендовали устанавливать количество обрезки на уровне 20% (например, Кесельман, Уилкокс, Ликс, Альгина и Фрадетт, 2007; Уилкокс и Кесельман, 2001; Wilcox, 1997), так что анализы рассчитываются после удаления самые крайние 20% случаев из каждого хвоста (подробности см. ниже). Юэнь (1974) предложил усеченный двухвыборочный t-критерий, основанный на стандартной ошибке и степенях свободы Welch (1938), который оказался более мощным и имел более точные эмпирические данные. Частота ошибок типа I, чем исходный t Стьюдента, когда распределения были ненормальными.

Чтобы улучшить подход к нормативным сравнениям для ситуаций, в которых распределения ненормальны или содержат выбросы, критерий Шуирмана-Уэлча для эквивалентность была изменена путем замены исходных средних значений и дисперсий усеченными средние значения и Winsorized дисперсии (названные процедурой Шуирманна-Юэна). H01 это отклонено, если tY1 ta;dfY и H02 отклоняется, если tY2 ta;dfY ; куда

$$\frac{Mt1 - Mt2}{\sqrt{d1 - d2}} = tY1;$$

$$\frac{Mt1 - Mt2}{\sqrt{d1 - d2}} = tY2;$$

а также

$$df_y = \frac{(d1 + d2)^2}{\frac{(d1)^2}{(h1 - 1)} + \frac{(d2)^2}{(h2 - 1)}}.$$

h1 и h2 представляют размеры выборки после обрезки, а Mt1 и Mt2 — совокупность. обрезанные средства. Чтобы получить усеченные средние значения, пусть Y(1)j Y(2)j ... Y(nj)j и пусть gj = [cnj] указывает, что cnj округляется до ближайшего целого числа в меньшую сторону; с представляет доля наблюдений, которые должны быть усечены в каждом хвосте распределения. эффективный объем выборки для j-й группы становится hj = nj 2gj. Поэтому образец усеченное среднее

$$Mtj = \frac{1}{hj}\sum_{i=gj+1}^{n-gj} Xj \quad \text{Идж:}$$

Далее, d определяется как

$$DJ = \frac{(n-1)r^2_{wj}}{(hj)(hj-1)};$$

где r^2_{wj} является Winsorized дисперсией. Образец Winsorized дисперсии, который требуется чтобы получить теоретически достоверную оценку стандартной ошибки усеченных средних, затем дается по

$$r^{\wedge}_{2wj} = \frac{1}{\text{номер } 1} \sum_{i=1} Xnj \quad Xij\ l^{\wedge}wj \quad ^{2}$$

Выборочное среднее Winsorized, необходимое для вычисления Winsorized дисперсии, вычисляется как

$$l^{\wedge}wj = \frac{1}{nj} \sum_{i=1} Xnj\ Xij;$$

куда

$$Xij = Yðgjþ1Þj \text{ , если } Yij < Yðgjþ1Þj$$
$$= Yij \text{ , если } Yðgjþ1Þj \le Yij \le Yðnj gjÞ$$
$$= Yðnj gjÞj \text{ , если } Yij > Yðnj gjÞj$$

Другими словами, среднее Winsorized вычисляется путем замены усеченных наблюдений самым экстремальным необрезанным значением из соответствующего хвоста.

Таким образом, цель этого проекта состоит в том, чтобы определить, является ли тест Шуирманна-Юэна, предложенный в этой статье, более точным в реальных условиях данных, чем ранее предложенные тесты Шуирмана и Шуирмана-Уэлча для определения, когда обработанные и нормальные группы сравнения эквивалентны. Что касается исследованных условий данных, в дополнение к неравным дисперсиям и распределениям, которые не являются нормальными, но одинаковыми по форме, это исследование расширит литературу по надежности, исследуя свойства тестов, когда базовые распределения совокупности различаются (например, одно перекошенный, один нормальный). Это важный аспект данного исследования, поскольку формы распределения в клинических и нормальных группах сравнения часто существенно различаются.

8. Метод

Моделирование Монте-Карло использовалось для сравнения вероятности обнаружения эквивалентности с тестом эквивалентности Шуирманна, тестом эквивалентности Шуирмана-Уэлча и тестом эквивалентности Шуирмана-Юэна. В этом исследовании манипулировали несколькими переменными, включая размер выборки, стандартные отклонения населения и форму распределения.
Условия, использованные в этом исследовании, приведены в таблице 1. Были выбраны размеры выборки n = 40, 100 и 400, причем использовались как равные, так и неравные размеры выборки. Было пять условий стандартного отклонения: в дополнение к одинаковым стандартным отклонениям по группам мы исследовали два уровня неравных стандартных отклонений, которые были либо положительно, либо отрицательно связаны с неравными размерами выборки. Положительно спаренные размеры выборки и стандартные отклонения означают, что больший размер выборки сочетается с большим стандартным отклонением, а меньший размер выборки сочетается с меньшим стандартным отклонением. Отрицательная пара размеров выборки и стандартных отклонений означает, что больший размер выборки сочетается с меньшим стандартным отклонением, а меньший размер выборки сочетается с большим стандартным отклонением.

Всего было использовано девять комбинаций форм распределения, основанных на нормальном распределении, асимметричном распределении и распределениях, содержащих выбросы в одном или обоих хвостах. Распределения с положительной и отрицательной асимметрией были сгенерированы с использованием распределения g и h (Hoaglin, 1985), где g представляет собой параметр асимметрии, а h

Таблица 1. Условия для исследования методом Монте-Карло

п1, п2	р1, р2	с1, с2
20, 20	1, 1	Нормально, нормально
15, 25	0,7, 1,3	Нормальный, + перекос
25, 15	0,5, 1,5	Нормальный, выброс (+/)
50, 50	1,3, .7 0,7,	Нормальный, выброс (+)
25, 75	1,3	+перекос, +перекос
75, 25		+ перекос, выброс (+/)
200, 200		перекос, + перекос
150, 250		Выброс (+/), выброс (+/)
250, 150		Выброс (+), выброс (+)

Примечание. с = форма распределения; + асимметрия = положительно асимметричная; перекос = отрицательный перекос; выброс (+) = выбросы только в верхней части хвоста; выброс (+/) = выбросы в верхнем и нижнем хвостах.

представляет параметр эксцесса. В частности, сильно асимметричное распределение было генерируется с использованием g = 1 и h = 0. Чтобы сгенерировать данные из g- и h-распределения, стандартные единичные нормальные переменные (Zij) были преобразованы в случайную величину

$$Xij = \frac{e^{gZij} - 1}{\text{грамм}} e^{\frac{hZ2}{ij}};$$

по выбранным для исследования значениям g и h. Отрицательно перекошенный распределение создается путем отражения положительно асимметричного распределения перед изменением среднее значение распределения. Чтобы получить распределение со стандартным отклонением rj, каждый Xij умножается на значение rj. Важно отметить, что это не влияет на стоимость нулевая гипотеза при g = 0 (см. Wilcox, 1994). Однако при g > 0 популяция среднее значение для g- и h-переменной равно

$$\text{лгх} = \text{г} \frac{1}{\text{ð1 чÞ р}} \quad 1\text{Þ}\frac{g2}{3}\text{h} \quad 1 :$$

Таким образом, для тех условий, где g > 0, lgh сначала вычиталось из Xij перед умножение на rj. При работе с усеченными средними совокупность усеченных средних за j-я группа была вычтена из переменной перед умножением на rj.

Распределение, содержащее выбросы, было создано путем добавления выбросов к одному или обоим из хвосты нормального распределения. Следуя методу, описанному Циммерманом (1994), выбросы были взяты из нормального распределения со стандартным отклонением в пять раз больше, чем больше, чем у оригинального дистрибутива. Как для одностороннего, так и для двустороннего выброса условиях, 10% от общего числа случаев были взяты из распределения выбросов. Как и в случае данных с g- и h-распределением, распределения, содержащие выбросы только в одном хвосте, были скорректированы таким образом, чтобы средние значения совокупности (или усеченные средние значения) были равны 0. Различные условия формы распределения пересекались с размером выборки и стандартным отклонением условий, в результате чего было получено 405 уникальных условий, каждое из которых оценивалось, когда значение null гипотеза была верной (ошибка первого рода) и когда нулевая гипотеза была ложной (мощность).

Для каждого условия было проведено десять тысяч симуляций с использованием номинального уровень значимости a = 0,05 и интервал эквивалентности (1, 1). Для ошибки первого рода

условиях, когда средние значения были установлены равными 0 и 1 (т. е. разница между средними находится на границах интервала эквивалентности). Для условий мощности средние значения были установлены равными 0 и 0,66 (т. е. разница между средними значениями попадает в интервал эквивалентности).

9. Результаты

Эмпирические частоты ошибок первого рода в интервале 0,025–0,075 (т. е. a 0,5a) считаются робастными. На рис. 1 и 2 приведены эмпирические частоты ошибок типа I для условий, включающих распределения одинаковой формы и распределения неидентичной формы соответственно. Таблицы 2 и 3 содержат эмпирические значения мощности для

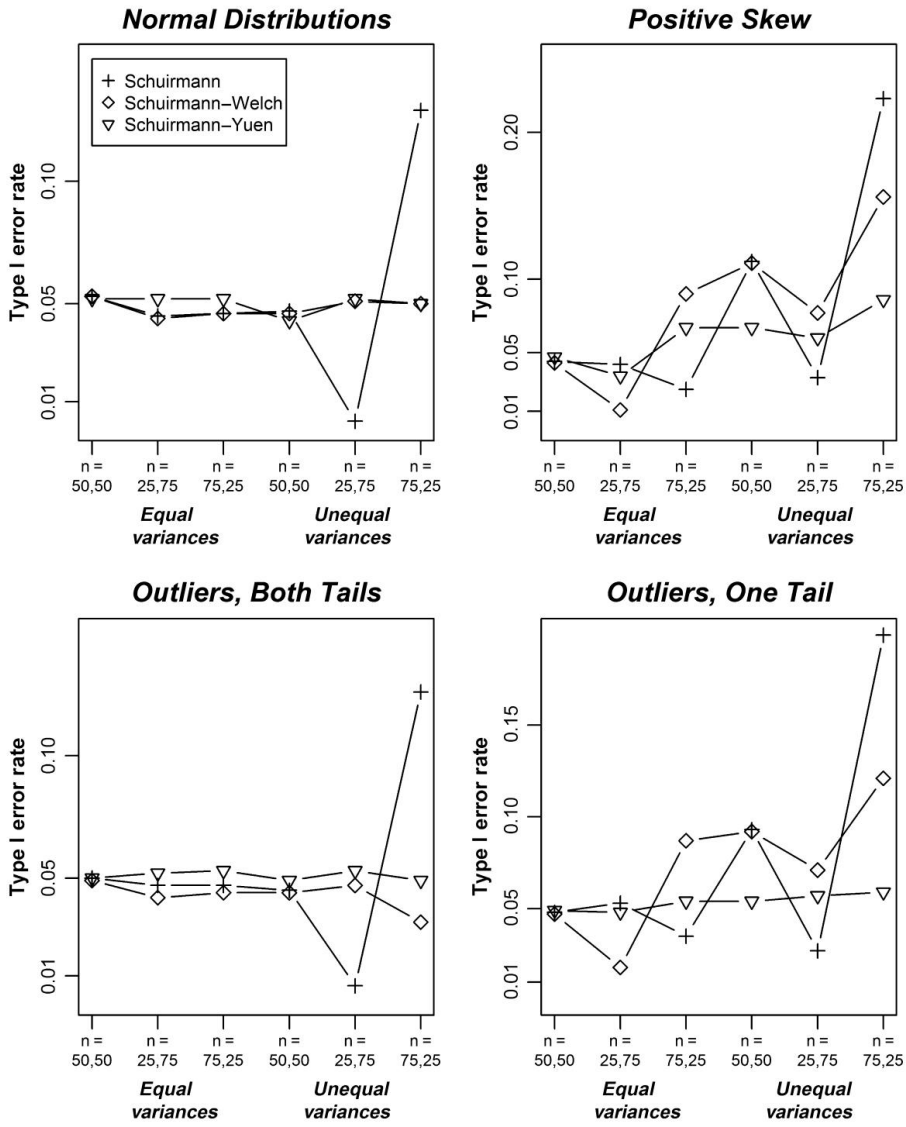

Рисунок 1. Коэффициенты ошибок типа I с одинаковыми формами распределения

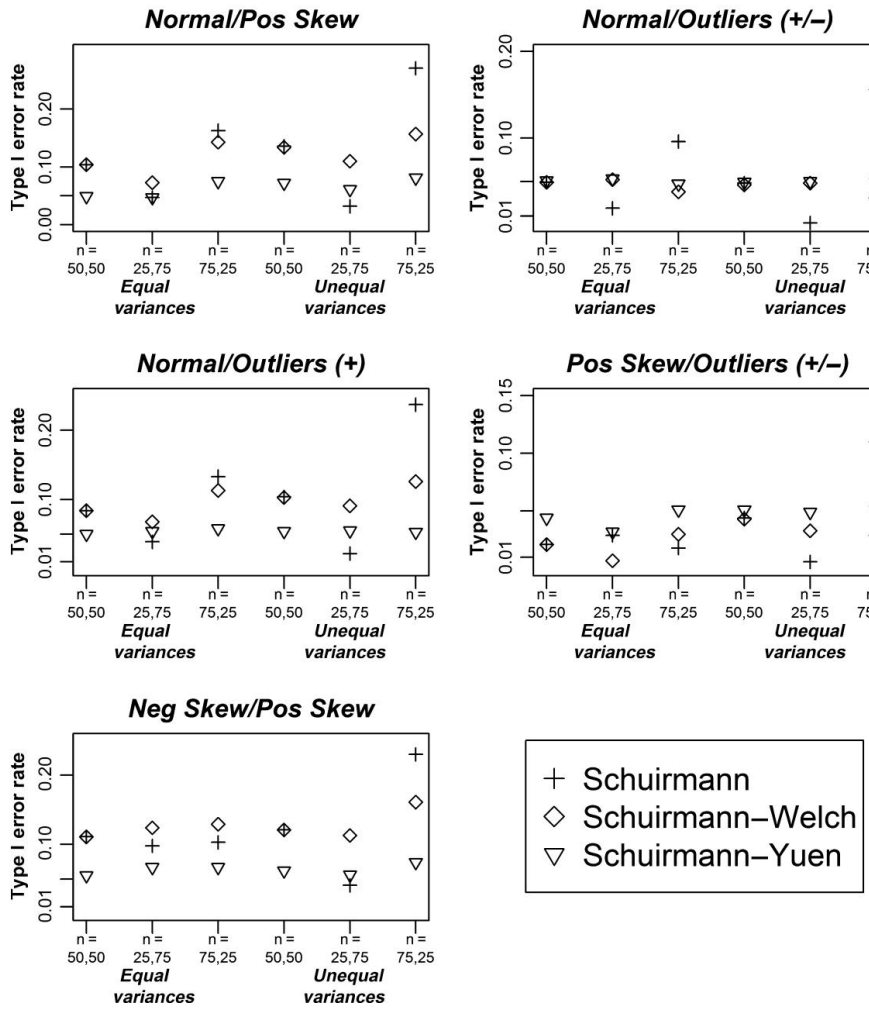

Рис. 2. Коэффициенты ошибок типа I при различных формах распределения.

Примечание: Pos Skew = положительная асимметрия; Neg Skew = отрицательный перекос; Выбросы (+/) = выбросы в верхнем и нижнем хвостах; Выбросы (+) = выбросы только в верхней части хвоста

условия, включающие распределения одинаковой формы и распределения неидентичной формы соответственно. Обратите внимание, что когда формы распределения одинаковы, нет необходимости менять порядок стандартных отклонений генеральной совокупности, потому что это просто воспроизведет предыдущие условия, но когда формы распределений различны, изменение порядка стандартных отклонений генеральной совокупности приводит к уникальным результатам. условия.

Поскольку картина результатов была очень похожей в условиях умеренного и крайне неравного стандартного отклонения, мы сообщаем только результаты для условий крайне неравного стандартного отклонения, поскольку это условие оказывает наибольшее влияние на статистику теста. Кроме того, поскольку структура результатов для N = 40, N = 100 и N = 400 была очень похожей, отображаются только коэффициенты ошибок типа I для N = 100 и коэффициенты мощности для N = 40 и 100. Тем не менее, мы обсуждаем результаты ниже, и полные таблицы результатов доступны, связавшись с авторами.

Таблица 2. Мощности при одинаковых формах распределения

п1, п2	Ш	Ш-В	Щ-й
	г1 = 1, г1 = 0,5, г2 = 1 г2 = 1,5	г1 = 1, г1 = 0,5, г2 = 1 г2 = 1,5	г1 = 1, г1 = 0,5, г2 = 1 г2 = 1,5
c1 = c2 = нормальный			
20, 20	0,282 0,243	0,281 0,238	0,254 0,217
15, 25	0,274 0,170	0,270 0,277	0,247 0,248
25, 15	0,265 0,298	0,261 0,195	0,255 0,172
50, 50	0,511 0,238	0,510 0,445	0,467 0,387
25, 75	0,438 0,277	0,436 0,513	0,409 0,446
75, 25	0,414 0,195	0,409 0,291	0,362 0,274
c1 = c2 = + перекос			
20, 20	0,089 0,212	0,085 0,208	0,207 0,250
15, 25	0,077 0,141	0,042 0,182	0,123 0,247
25, 15	0,098 0,222	0,133 0,170	0,225 0,225
50, 50	0,203 0,302	0,203 0,298	0,394 0,357
25, 75	0,140 0,110	0,069 0,262	0,275 0,399
75, 25 0,157 0,422 c1 = c2 = выброс (+/)		0,283 0,314	0,355 0,294
20, 20 15, 25 0,087 0,0	0,089 0,096	0,087 0,093	0,223 0,183
0,109 50, 50 0,251 0,22	0,174 0,088	0,093 0,103	0,197 0,218
c1 = c2 = выброс (+)		0,087 0,058	0,194 0,117
		0,250 0,218	0,402 0,319
		0,203 0,260	0,332 0,385
	0,183 0,297	0,194 0,120	0,329 0,229
20, 20	0,100 0,178	0,099 0,176	0,219 0,184
15, 25	0,091 0,121	0,067 0,192	0,194 0,221
25, 15	0,079 0,238	0,121 0,159	0,196 0,149
50, 50	0,243 0,291	0,243 0,290	0,398 0,355
25, 75	0,188 0,129	0,129 0,264	0,309 0,403
75, 25	0,215 0,394	0,295 0,246	0,337 0,235

Примечание. Sch = Шуирманн; Sch-W = Шуирманн-Велч; Sch-Y = Шуирманн-Юэнь; c = форма распределения; + асимметрия = положительно асимметричная; выброс (+/) = выбросы в верхнем и нижнем хвостах выброса; (+) = выбросы только в верхней части хвоста; серый цвет = частота ошибок типа I не контролируется в пределах 0,5a (0,025–0,075).

9.1. Ошибка I типа

9.1.1. Идентичные формы распределения

Результаты показали, что, когда как в группе лечения, так и в группе сравнения нормальное распределение, частоты ошибок I рода поддерживались в интервале 0,025–0,075 для всех трех тестов эквивалентности. Исключение составляли случаи, когда стандартные отклонения и размеры выборки были неодинаковыми, и в этом случае частота ошибок типа I для исходного Процедура Шуирмана была завышена в ситуации отрицательного спаривания (например, 0,129 для N = 100) и занижается в ситуации положительного спаривания (например, 0,002 для N = 100).

Когда формы распределения были искажены или содержали выбросы, ошибка типа I ставки поддерживались в консервативных пределах наиболее эффективно при Критерий Шуирманна–Юэна. Хотя у Шуирманна-Юэна была некоторая частота ошибок типа I. которое вышло за допустимые пределы, когда оба распределения были чрезвычайно асимметричными

Таблица 3. Мощности при различных формах распределения

n1, n2	Ш	Ш-В	Щ-й
	г1 = 1, г1 = 0,5, г2 = 1 г2 = 1,5	г1 = 1, г1 = 0,5, г2 = 1 г2 = 1,5	г1 = 1, г1 = 0,5, г2 = 1 г2 = 1,5
c1 = нормальный; c2 = + перекос			
20, 20 15, 25 0,238 0,188 0,281 0,262		0,296 0,265	0,249 0,240
25, 15 0,376 0,322 50, 50 0,253 0,438		0,263 0,283	0,210 0,238
c1 = норма; c2 = выброс (+/-)	25, 25	0,294 0,245	0,244 0,224
	0,450 0,675	0,376 0,321	0,418 0,355
	75, 25	0,339 0,352	0,365 0,422
		0,352 0,293	0,349 0,259
20, 20 15, 25 0,124 0,054 0,219 0,169		0,191 0,109	0,243 0,166
0,134 50, 50 0,307 0,239 25, 75 0,368		0,172 0,119	0,223 0,210
0,358 75, 25 c1 = норма; c2 = выброс	(+)	0,157 0,060	0,192 0,140
		0,307 0,233	0,395 0,341
		0,300 0,273	0,363 0,428
		0,233 0,134	0,323 0,222
20, 20 15, 25 0,223 0,148 0,287 0,262		0,286 0,256	0,243 0,193
25, 15 0,373 0,306 50, 50 0,253 0,432		0,242 0,244	0,216 0,174
c1 = +косая; c2 = выброс (+/-)	25, 25	0,285 0,211	0,225 0,180
	0,450 0,675	0,371 0,302	0,444 0,331
	75, 25	0,328 0,311	0,351 0,403
		0,336 0,283	0,340 0,218
20, 20 15, 25 25, 15 50, 50 25, 75 75,		0,026 0,060	0,185 0,123
25 0,103 0,288 c1 = косой; c2 = +		0,024 0,056	0,206 0,115
перекос	0,028 0,086	0,042 0,048	0,102 0,121
	0,145 0,201	0,145 0,198	0,333 0,368
	0,128 0,039	0,078 0,170	0,387 0,314
		0,170 0,120	0,211 0,298
20, 20 0,242 0,153 15, 25 25, 15 50, 50		0,244 0,225	0,224 0,224
	25, 75 75, 25	0,249 0,240	0,228 0,232
	0,218 0,271	0,237 0,187	0,221 0,234
	0,310 0,303	0,310 0,302	0,375 0,345
	0,299 0,145	0,299 0,304	0,321 0,373
	0,299 0,407	0,299 0,282	0,325 0,266

Примечание. Sch = Шуирманн; Sch-W = Шуирманн-Велч; Sch-Y = Шуирманн-Юэнь; c = форма распределения; + асимметрия = положительно асимметричная; выброс (+/) = выбросы в верхнем и нижнем хвостах выброса; (+) = выбросы только в верхней части хвоста; серый цвет = частота ошибок типа I не контролируется в пределах 0,5a (0,025–0,075).

и дисперсии были неравными, исходные процедуры Шуирмана и Шуирмана-Уэлча потерпели неудачу во многих других условиях, чем Schuirmann-Yuen, и когда они действительно потерпели неудачу ставки часто сильно отличались от номинального уровня. Например, при n1 = 25 и n2 = 15 и неравные дисперсии, частота ошибок типа I для теста Шуирмана составила 0,177. а для Шуирманна-Уэлча было 0,145. Кроме того, в условиях большого размера выборки

(N = 400) , показатели для Шуирманна-Юэна всегда поддерживались в пределах допустимых границ ошибки Типа I, тогда как картина для процедур Шуирмана и Шуирмана-Уэлча была аналогична таковой для небольших размеров выборки, при этом показатели регулярно падали ниже или выше допустимых уровней.

9.1.2. Различные формы
распределения. Когда две группы имели разные формы распределения, эмпирический контроль ошибок типа I по процедуре Шуирманна-Юэна был значительно лучше, чем у оригинальной процедуры Шуирмана или Шуирманна-Уэлча. В частности, хотя у процедуры Шуирмана-Юэна коэффициенты ошибок типа I выходили за допустимые пределы в одном из 30 условий, указанных в таблице 3 (нормальное и асимметричное распределение и неравные дисперсии), каждая из процедур Шуирмана и Шуирмана-Уэлча имела Частота ошибок типа I, которые выходили за допустимые пределы более чем в половине условий. Кроме того, хотя коэффициенты для всех условий варьировались от 0,010 до 0,091 для теста Шуирманна-Юэна, коэффициенты для коэффициента Шуирмана варьировались от 0,000 до 0,271, а коэффициенты для критерия Шуирманна-Уэлча варьировались от 0,000 до 0,180. Когда размеры выборки были большими, как и при одинаковых формах распределения, эмпирические частоты ошибок типа I хорошо контролировались критерием Шуирмана-Юна во всех условиях, но обычно выходили за допустимые пределы с помощью критерия Шуирмана и Шуирмана-Уэлча.

9.2. Мощность

Значения мощности, выделенные серым цветом в таблицах 2 и 3, представляют собой условия, при которых частота ошибок типа I не находилась в пределах интервала устойчивости, и, таким образом, показатели мощности являются смещенными (и их не следует интерпретировать).

9.2.1. Идентичные формы
распределения Когда формы распределения были нормальными, стандартные отклонения были равными, а размеры выборки были одинаковыми, показатели мощности были несколько выше для тестов Шуирманна и Шуирмана-Уэлча, чем для тестов Шуирмана-Юна. Этот результат обусловлен тем, что Schuirmann-Yuen использует усеченные средние значения, и поэтому статистика теста была основана на меньшем размере выборки. Когда формы распределения были ненормальными, но идентичными, мощности для критерия Шуирмана-Юэна, как правило, были намного выше, чем для критерия Шуирмана или Шуирмана-Велча (и также важно отметить, что многие коэффициенты мощности для критерия Шуирмана и Шуирмана –Welch не поддаются интерпретации, поскольку частота ошибок типа I вышла за пределы допустимого диапазона). Например, при равных размерах выборки и дисперсии и N = 100 мощность как для процедур Шуирмана, так и для процедур Шуирмана-Уэлча составила 0,203, тогда как мощность процедуры Шуирмана-Юэна составила 0,394.

Когда размеры выборки были большими (N = 400), различия в мощности с ненормальными распределениями были еще более преувеличены, с показателями мощности для Шуирманна и Шуирмана-Уэлча в диапазоне от примерно 0,3 до 0,6, а коэффициенты для Шуирманна-Юэна варьировались примерно от 0,3 до 0,6. 0,7 до 0,9. Например, когда оба распределения были сильно перекошены в положительную сторону, размеры выборки составляли 150 и 250, а дисперсии были неравными, коэффициенты мощности для критерия Шуирмана и Шуирмана-Уэлча составляли 0,385 и 0,481 соответственно, а мощность для критерия Шуирмана-Юэна составляла 0,812. .

9.2.2. Различные формы распределения

Когда распределения имели разные формы, появлялась одна и та же картина результатов; однако из-за плохого контроля ошибок типа I процедур Шуирмана и Шуирмана-Уэлча в этих условиях было очень мало возможностей для беспристрастных сравнений с процедурой Шуирмана-Юна. Однако для условий, в которых были возможны сравнения, Шуирманн-Юэн всегда был более мощным. Например, когда одно распределение было асимметричным и одно распределение содержало выбросы в обоих хвостах, n1 = 50 и n2 = 50, а дисперсии были неравными, мощность для коэффициента Шуирмана составила 0,201, для коэффициента Шуирмана–Уэлча – 0,198, а для коэффициента Шуирмана–Юэна был 0,368.

Когда размеры выборки были большими (N = 400), появлялась та же картина, что и для выборок меньшего размера, при этом показатели мощности для теста Шуирмана-Юэна часто были значительно выше, чем показатели для других процедур. Например, когда одно распределение было асимметричным и одно распределение содержало выбросы в обоих хвостах, n1 = 150 и n2 = 250, и дисперсии были равными, мощность для Шюрмана составила 0,475, для Шуирмана-Уэлча была 0,455, а для Шуирмана-Юна был 0,876.

10. Обсуждение

Когда клиент начинает терапию по определенной проблеме, такой как депрессия, конечная цель во многих случаях состоит в том, чтобы поведение клиента по этой проблеме вернулось к состоянию нормального функционирования. Нормативные сравнительные тесты позволяют исследователям определить, претерпела ли группа, прошедшая лечение, возвращение к нормальному функционированию, сравнивая группу, прошедшую лечение, с нормативной группой по интересующему показателю. Kendall et al. (1999) использовали два односторонних критерия эквивалентности Schuirmann (1987), чтобы определить, эквивалентны ли обработанная и нормативная группы сравнения по конкретному показателю. Было показано, что тесты эквивалентности более эффективны, чем традиционные тесты гипотез (например, критерий Стьюдента) для определения эквивалентности, но традиционные тесты эквивалентности не являются надежными, когда нарушаются предположения о нормальности и/или однородности дисперсии. Данненберг и др. (1994) включили статистику гетероскедастического теста (критерий Шуирманна-Уэлча), которая оказалась надежной, когда предположение об однородности дисперсии было нарушено. Однако оставались опасения по поводу устойчивости этого теста к ненормальности распределения или наличию выбросов.

Цель этой статьи состояла в том, чтобы сравнить надежность исходных критериев эквивалентности Шуирманна (1987) и Шуирманна-Уэлча (Dannenberg et al., 1994), когда стандартные отклонения были неравными, а распределения населения были ненормальными, с надежностью критерия эквивалентности процедура Шуирмана-Юэна, основанная на усеченных средних, описанная в этой статье. Ожидалось, что тест Шуирманна-Юэна, который включает статистику гетероскедастического теста с усеченными средними, обеспечит лучший контроль ошибок типа I и мощность в условиях дисперсионного неравенства и ненормальности распределения.

Результаты этого исследования показывают, что тест Шуирмана-Юэна обеспечивает гораздо лучший контроль эмпирической частоты ошибок типа I и более высокую мощность, чем тесты Шуирмана или Шуирмана-Уэлча. Хотя метод Шуирманна-Юэна не всегда обеспечивал приемлемый контроль ошибок типа I при небольших размерах выборки, показатели при больших размерах выборки хорошо контролировались, а показатели во всех условиях контролировались намного лучше, чем исходные процедуры Шуирмана или Шуирмана-Уэлча. Кроме того, мощность процедуры Шуирмана-Юэна, когда хотя бы одно из распределений было ненормальным, была регулярно выше, чем у процедур Шуирмана или Шуирмана-Уэлча, т.е.

преимущество, которое имеет первостепенное значение для исследователей, проводящих нормативные сравнения. Одно предостережение, высказанное анонимным рецензентом, заключается в том, что, хотя критерий Шуирманна-Юэна сравнивает усеченные средние значения и, таким образом, является эффективным методом сравнения типичного индивидуума в одной группе с типичным индивидуумом в другой группе, возможно, что лежащие в основе формы распределения могут быть очень разными. другой. Таким образом, исследователям рекомендуется исследовать формы распределения своих переменных и проявлять осторожность при сравнении центральных тенденций, когда формы распределения различаются. Чтобы улучшить доступность методов, описанных в этом документе, функция R (R Development Core Team, 2010) для проведения нормативных сравнительных тестов, описанных в этом документе, доступна по адресу http://www.psych.yorku.ca/ cribbie/ norm_comparisons_rprogram_web.txt. R — это статистическая программа с открытым исходным кодом, доступная по а

Потребность в надежных статистических тестах с годами возросла с осознанием того, что традиционные статистические тесты ненадежны, когда нарушаются предположения о нормальности и однородности дисперсии. Это особенно важно для исследователей в области клинической психологии, где эти предположения редко выполняются. В заключение, предложенный тест эквивалентности Шуирмана-Юна рекомендуется для проведения нормативных сравнений, поскольку он обеспечивает лучший контроль ошибок типа I и большую мощность, чем исходные тесты эквивалентности Шуирмана или Шуирмана-Уэлча.

использованная литература

Крибби, Р.А., и Арпин-Крибби, Калифорния (2009 г.). Оценка клинической значимости посредством тестирования эквивалентности: расширение подхода к нормативным сравнениям. Psychotherapy Research, 19, 677–686.

Данненберг, О., Детте, Х., и Мунк, А. (1994). Распространение приближенного t-решения Уэлча на сравнительные испытания биоэквивалентности. Биометрика, 81, 91–101.

Голински, К., и Крибби, Р.А. (2009). Растущая роль количественных методологов в продвижении психология. Канадская психология, 50, 83–90.

Груман, Дж. А., Крибби, Р. А., и Арпин-Крибби, Калифорния (2007). Влияние гетероскедастичности на тесты эквивалентности. Журнал современных прикладных статистических методов, 6, 132–140.

Хоаглин, округ Колумбия (1985). Численное суммирование формы: g- и h-распределения. В Д. Хоаглин, Ф. Мостеллер и Дж. Тьюки (редакторы), Изучение таблиц данных, тенденций и форм (стр. 461–513). Нью-Йорк: Уайли.

Джейкобсон, Н.С., и Труакс, П. (1991). Клиническая значимость: статистический подход к определению изменений в психотерапевтических исследованиях. Журнал консалтинга и клинической психологии, 59, 12–19.

Кендалл, ПК (1997). Редакция. Журнал консалтинга и клинической психологии, 15, 3–5.

Кендалл, ПК, Маррс-Гарсия, А., Нэт, С.Р., и Шелдрик, Р.К. (1999). Нормативные сравнения для оценки клинической значимости. Журнал консультирования и клинической психологии, 67, 285–299.

Кесельман, Х. Дж., Хьюберти, С. Дж., Ликс, Л. М., Олейник, С., Крибби, Р., Донахью, Б., ... Левин, младший (1998). Статистическая практика исследователей в области образования: анализ их ANOVA, MANOVA и ANCOVA. Обзор образовательных исследований, 68, 350–386.

Кесельман, Х.Дж., Отман, А.Р., Уилкокс, Р.Р., и Фрадетт, К. (2004). Новые и улучшенные два образец t-теста. Психологическая наука, 15, 47–51.

Кесельман, Х.Дж., Уилкокс, Р.Р., Ликс, Л.М., Альгина, Дж., и Фрадетт, К. (2007). Адаптивная робастная оценка и тестирование. Британский журнал математической и статистической психологии, 60, 267–293.

Кремер, ХК, и Купфер, DJ (2006). Размер эффектов лечения и их важность для клинических исследований и практики. Биологическая психиатрия, 59, 990–996.

Крамер, Х.К., Морган, Г.А., Пиявка, Н.Л., Глинер, Дж.А., Васке, Дж.Дж., и Хармон, Р.Дж. (2003). Меры клинической значимости. Журнал Американской академии детской и подростковой психиатрии, 42 (12), 1524–1529.

Мандзони, Г.М., Крибби, Р.А., Вилла, В., Арпин-Крибби, К.Х., Гондони, Л., и Кастельнуово, Г. (2010). Психологическое благополучие у пациентов с ишемической болезнью сердца, страдающих ожирением, при поступлении и при выписке из четырехнедельной программы кардиореабилитации. Границы психологии, 1, 1–7.

Мартинович З., Сондерс С. и Ховард К. (1996). Некоторые комментарии к «оценке клинических значение'. Психотерапевтические исследования, 6, 124–132.

Микчери, Т. (1989). Единорог, нормальная кривая и другие невероятные существа. психологический Бюллетень, 105, 156–166.

Основная группа разработки R (2010 г.). R: Язык и среда для статистических вычислений. Вена, Австрия: R Foundation for Statistical Computing.

Роджерс, Дж. Л., Ховард, К. И. и Весси, Дж. Т. (1993). Использование критериев значимости для оценки эквивалентности между двумя экспериментальными группами. Психологический бюллетень, 113, 553–565.

Санчес-Ортуно, М.М., и Эдингер, Д.Д. (2010). Копейка за ваши мысли: модели убеждений, связанных со сном, симптомы бессонницы и результаты лечения. Поведенческие исследования и терапия, 48, 125–133.

Саттертуэйт, FE (1946). Приближенное распределение оценок компонентов дисперсии. Бюллетень биометрии, 2, 110–114.

Шуирманн, ди-джей (1987). Сравнение процедуры двусторонних тестов и степенного подхода для определения эквивалентности средней биодоступности. Журнал фармакокинетики и биофармацевтики, 15, 657–680.

Моряк, Массачусетс, и Серлин, Р.К. (1998). Доверительные интервалы эквивалентности для двухгрупповых сравнений средних. Психологические методы, 3, 403–411.

Уоллах, Х.С., Сафир, М., и Бар-Цви, М. (2009). Когнитивно-поведенческая терапия виртуальной реальности при тревоге публичных выступлений: рандомизированное клиническое исследование. Модификация поведения, 33, 314–333.

Уэлч, Б.Л. (1938). Значение разницы между двумя средними значениями при населении дисперсии не равны. Биометрика, 29, 350–362.

Уилкокс, Р.Р. (1994). Модель односторонних случайных эффектов для усеченных средних. Психометрика, 59, 289–306.

Уилкокс, Р. Р. (1997). Введение в надежную оценку и проверку гипотез. Сан-Диего, Калифорния: Академическая пресса.

Уилкокс, Р. Р., и Кесельман, Х. Дж. (2001). Использование усеченных средств для сравнения мер K, соответствующих двум независимым группам. Многомерные поведенческие исследования, 36, 421–444.

Юэн, К.К. (1974). Две выборки обрезаны по неравным дисперсиям генеральной совокупности. Биометрика, 61(1), 165–170.

Циммерман, Д. В. (1994). Примечание о влиянии выбросов на параметрические и непараметрические тесты. Журнал общей психологии, 121, 391–401.

Поступила в редакцию 10 апреля 2012 г.; исправленная версия получена 21 марта 2013 г.