



Изучение восприятия осмысленности в визуальных представлениях двумерных отношений

Натали Берибски, Хизер Дэвидсон и Роберт А. Крибби

Факультет психологии, Йоркский университет, Торонто, Канада

АННОТАЦИЯ

Исследователям часто приходится учитывать практическое значение отношений. Например, интерпретация величины величины эффекта или установление границ при проверке эквивалентности требует знания осмысленности отношения. Однако было проведено мало исследований, изучающих степень взаимосвязи между переменными (например, корреляция, разность средних), необходимую для того, чтобы связь интерпретировалась как значимая или практически значимая. В этом исследовании мы представили статистически подготовленным и неподготовленным участникам набор цифр, которые отображали различную степень средней разницы между группами или корреляции между переменными, и участники указывали, было ли каждое отношение значимым. Результаты показывают, что статистически обученные и неподготовленные участники различаются в своей квалификации значимых отношений и что существует значительная вариabельность в том, насколько большими должны быть отношения прежде чем они будут признаны значимыми. Результаты также показали некоторый скепсис, какая степень отношений считается значимой для людей в условиях, несвязанных с контекстом.

Предметы: Психология, психология, Статистика

Ключевые слова: Величина эффекта, Перекрывающиеся истогаммы, Диаграммы рассеяния, Практическая значимость

ВВЕДЕНИЕ

Во многих случаях исследователям необходимо учитывать практическое значение отношений.

Это может быть необходимо при интерпретации величины величины эффекта или количественной оценке минимально важных отношений при проверке эквивалентности. Несмотря на это, было проведено мало исследований, изучающих, когда отношения (например, корреляция или разность средних) начинают интерпретироваться как значимые или практически значимые. Из-за этого недостатка информации многие исследователи по умолчанию использовали рекомендацию Кэна (1988) о размере эффекта, которые определяют нижние границы для интерпретации отношения как «малого», независимо от контекста исследования. Еще меньше известно о вариативности интерпретации величины эффекта суждения о размерах эффекта могут сильно различаться от человека к человеку. Например, у людей могут быть разные моменты, когда они начинают воспринимать корреляции между переменными или различия между двумя средними распределениями как значимые.

В этой статье мы с начала обсудим два важных параметра, в которых практическое значение имеет значение. важную роль: интерпретация размера эффекта и проверка эквивалентности. После этих приложений, имеющих практическое значение, будут рассмотрены рекомендации предыдущих исследований, касающиеся практической значимости. Наконец, мы проводим исследование для изучения отдельных

Подано 16 июля 2018 г.
Принято 27 марта 2019 г.
Опубликовано 14 мая 2019 г.

Автор, ответственный за переписку Роберт А. Крибби, cribbie@yorku.ca

Академический редактор

Эндрю Грей

Дополнительная информация и
Объявления можно найти на
стр. 16

DOI 10.7717/peerj.6853

Copyright

2019 Beribisky et al.

Распространяется
под Creative Commons CC-BY 4.0

ОТКРЫТЫЙ ДОСТУП

восприятие значимости в визуальных представлениях корреляции и стандартизированной разности средних.

Интерпретация величины эффекта

Величина эффекта может быть определена как сила или величина логического результата, важного для исследователя в рамках исследования или как количественное свидетельство величины явления (Cumming, 2012; Kelley & Preacher, 2012). Величину эффекта и доверительные интервалы для величины эффекта следует сообщать вместе с результатами традиционной проверки значимости нулевой гипотезы (NHST), чтобы представить читателю уровень практической значимости (Thompson, 2001; Cumming, 2012). Кроме того, наименьший значимый размер эффекта является рекомендуемым значением для использования в анализе мощности, чтобы обеспечить достаточную мощность для обнаружения размеров эффекта, потенциально представляющих интерес для исследователя.

Величина эффекта может быть представлена в не стандартизированных (например, разница между двумя средними значениями) и/или в стандартизированных единицах. В этой статье мы сосредоточимся на стандартизированных размерах эффекта, поскольку их можно интерпретировать без привязки к конкретной переменной результата. Коэффициент (с стандартизированной разницей) d (или g для доли дисперсии) обычно используется в качестве размера эффекта широкой меры средних разностей, а r и f^2 — используемые меры размера эффекта для корреляции или регрессии. Эти четыре меры стандартизируют размер эффекта в зависимости от степени изменчивости выборки. Тем не менее, было проведено мало исследований относительно того, когда размеры эффекта начинают представлять значимую взаимосвязь. Хотя размер эффекта, представляющий наименьшую значимую взаимосвязь, будет зависеть от характера исследования, понимание того, что люди воспринимают как наименьший значимый эффект в условиях, явно не зависящих от контекста, полезно для понимания того, как исследователи воспринимают и интерпретируют величину эффекта. Например, исследователи, интерпретирующие величину величины эффекта или определяющие наименьшую значимую взаимосвязь для использования в анализе мощности, выигрывают от доступных руководств, которые можно использовать в сочетании с контекстом их исследования.

Традиционное тестирование на основе различий в сравнении с тестированием на эквивалентность Традиционный NHST используется для оценки доказательства отсутствия связи между переменными (например, различий в средних значениях популяций). Если у исследователя есть гипотеза о наличии взаимосвязи между переменными, то подходит традиционный NHST. Однако, когда целью исследования является обнаружение отсутствия взаимосвязи между переменными (например, незначительная разница в средних значениях, отсутствие связи между переменными и т. д.), существуют альтернативы традиционной NHST, основанной на различиях, такие как проверка эквивалентности, которая следует использовать для исследования отсутствия связи (например, Kruschke, 2011; Rogers, Howard & Vessey, 1993; Rouder et al., 2009; Wellek, 2010).

В традиционном NHST, основанном на различиях, отказ отклонить нулевую гипотезу часто ошибочно интерпретируется как доказательство нулевой гипотезы (Allan & Cribbie, 2013). Поскольку нуль никогда не может быть принят, исследователи, стремящиеся обнаружить отсуствие взаимосвязи между переменными, должны использовать тесты, созданные для обнаружения эквивалентности. Например, нулевая гипотеза для двустороннего теста эквивалентности утверждает, что величина отношения больше или равна минимальному значимому значению (например, $H_0: \rho \geq \delta$ или $\rho \leq -\delta$), в то время как альтернативная гипотеза предполагает, что величина отношения меньше минимального значимого значения (например, $H_1: \delta < \rho < -\delta$).

Определение интервала эквивалентности (т. е. минимальной значимой разницы) перед проведением исследования является важным шагом в тестировании эквивалентности, непосредственно влияющим на интерпретацию результатов. Rusticus & Eva (2016), Rogers и коллеги (1993) и многие другие утверждают, что определение интервала эквивалентности, относящегося к контексту и целям исследования, может быть чрезвычайно сложной задачей. Те же проблемы возникают, когда исследователи применяют байесовский анализ для получения доказательств в пользу традиционной нулевой гипотезы. Например, Kruschke (2013) описывает «область практической эквивалентности» как байесовскую версию интервала эквивалентности, а Morey & Rouder (2011) обсуждают освоенные на интервалах гипотезы для проведения байесовского факторного анализа. Соответственно, слишком большой интервал эквивалентности может привести к тому, что исследователь сделает неправильный вывод об эквивалентности (когда между группами есть значимые различия), а неоправданно малый интервал может привести к тому, что исследователь упустит наличие незначительной связи.

Хотя контекст чрезвычайно важен при установлении границ эквивалентности, в литературе существует множество общих рекомендаций по созданию этих границ. Например, для ситуационных связанных с отсутствием различий между групповыми средними значениями, рекомендации широко варьировались, варьируясь от 5% (Barker et al., 2002; Rusticus & Lovato, 2011) до 20% (Роджерс, Ховард и Веси, 1993). Рекомендации Коэна относительно того, что составляет «малую» связь, также цитируются, когда исследователи изучают соответствующие границы для проверки эквивалентности (Davidson & Cribbie, 2016). Кроме того, Wellek (2010) представил несколько рекомендаций по установлению границ эквивалентности для проведения тестов эквивалентности в широком диапазоне исследовательских усилий. Это исследование надеется расширить эти не зависящие от контекста рекомендации, эмпирически оценив, насколько большими должны быть отношения, прежде чем они будут признаны значимыми, а также насколько изменчивы люди в своем восприятии того, что составляет значимые отношения. Вместе с контекстом исследования эта информация может помочь исследователям, которым необходимо установить соответствующий интервал/границу эквивалентности.

Предложения относительно наименьшей значимой взаимосвязи. Существует множество существующих рекомендаций по выбору наименьшей значимой взаимосвязи, такой как минимальное клинически значимое различие (MCID) в исследованиях в области здравоохранения. В клинических исследованиях Fayers et al. (2000) утверждали, что наименьшие значимые различия, используемые при расчете размера выборки, должны быть как клинически важными (например, для групп заинтересованных сторон, которых касается различие), так и клинически реалистичными (например, не предполагать более высокую выживаемость для препарата, чем в настоящее время). Методы выбора важности или реалистичного различия включают (а) обзор текущего состояния исследований по вопросу с помощью метаанализа или систематического обзора или (б) сбор мнений соответствующих заинтересованных сторон с помощью панелей или фокус-групп (Cook et al., 2018).

Хотя эти методы идеальны, они не всегда могут быть практичными или выполнимыми, особенно при поисковых исследованиях. В этих случаях Cook et al. (2018) заявляют, что пилотное исследование может быть полезным способом получения знаний о том, что может быть реалистичным MCID между группами. Пилотные исследования обычно проводятся в несколько этапов для оценки соответствующей информации, необходимой для будущих более масштабных исследований (Cook et al., 2014).

Кроме того, существует три метода определения важных различий, которые могут служить основой для MCID в клинических исследованиях (Cook et al., 2014). Сюда входят метод прививки, метод расщепления и метод экономики здоровья. Метод прививки использует суждение с соответствующей заинтересованной стороны, чтобы сообщить, в чем может заключаться важное различие. Например, Jaeschke, Singer & Guyatt (1989) отмечают, что клиницисты часто способны уловить смысл клинически значимой разницы во времени, что делает разумным использование их опыта при определении того, что такое MCID для клинического инструмента. Метод расщепления предполагает, что MCID может быть больше, чем максимальная неточность инструмента измерения (Cook et al., 2018). Подходы, основанные на методе расщепления, обычно основаны на (а) стандартной ошибке измерения, (б) наименьшей разнице, обнаруживаемой статистическим тестом, или даже (в) «эмпирическом правиле», которое фокусируется на конкретном статистическом показателе (Кук и др., др., 2014). В экономических методах здравоохранения используется ожидаемая чистая выгода для оптимизации стоимости единицы воздействия на здоровье с суммой, которую лицо, принимающее решение, желает заплатить, что позволяет исследователям количественно оценить ожидаемую чистую прибыль от конкретного исследования или исследовательского протокола.

Дополнительный способ принять решение о подходе с минимально значимым эффектом при отсутствии какой-либо ценной информации — использовать стандартизированный размер эффекта вне контекста, опираясь на рекомендации предыдущих исследований (Weber & Popova, 2012). Например, Rogers, Howard & Vessey (1993) приводят пример 20-процентной разницы как наименьшее значимое расхождение между двумя средними значениями, в то время как Cohen (1988) предложил нижнюю границу для стандартизированной малой величины эффекта $d = 0,2$ или $r = .1$. Однако, как и Коэн, Роджерс, Ховард и Вессе (1993) заявили, что наименьшая значимая ассоциация обязательно будет варьироваться в зависимости от рассматриваемого вопроса исследования, и отговорили исследователей от легкого принятия предложенных ими руководящих принципов.

Некоторые исследователи предположили в контексте поиска хорошего исследования, что можно было бы включить ряд возможных минимально значимых различий (Крибби и Арпин-Крибби, 2009; Вебер и Попова, 2012). Например, в клинических условиях Крибби и Арпин-Крибби (2009) утверждали, что средняя эквивалентность групп может быть зафиксирована различными уровнями «близости». Точно так же в контексте исследования размера эффекта Вебер и Попова (2012) заявили, что полученный эффект можно сравнить с диапазоном размеров эффекта, чтобы проиллюстрировать, как различные предположения об эффекте могут повлиять на решение теста. Однако они также предупредили, что их результаты следует использовать в качестве рекомендаций только при установлении минимальной значимой разницы в исследованиях.

Фергюсон (2009) составил список предложений по интерпретации результатов социальными науками о новых рекомендациях из предыдущих обзоров. Для Коэна и Гирса минимальные размеры эффекта, представляющие практически значимый эффект, составляли 0,41 и 0,20 соответственно.

В одном недавнем исследовании изучалась точка, в которой люди начинают ощущать наличие отношений. Rusticus & Eva (2016) исследовали восприятие людьми эквивалентности, чтобы определить порог эквивалентности для данных медицинских школ образования. Исследователи попросили участников сравнить учебные заведения медицинских школ на предмет их сходства по четырем различным результатам: средний балл бакалавриата, опрос студентов, структурированный клинический экзамен и экзамен на получение лицензии медицинских школ. Участники видели эти сравнения на

ряд гистогрaмм, отображающих широкий диапазон стандартизированных разностей средних (Коэн d от 0,10 до 1,20). В исследовании был установлен порог эквивалентности, при котором 50% участников считали слайты неэквивалентными. Основываясь на своих результатах, Рустикус и Ева установили $d = 0,50$ как порог, при котором люди различают изменения между распределениями.

[Rusticus & Eva \(2016\)](#) был примером исследования, в котором интерпретация размера эффекта проводилась в заранее определенном контексте. Как обсуждалось ранее, минимально значимые эффекты следует использовать с учетом определенного контекста. Хотя идеальное решение о конкретном значимом эффекте должно приниматься в ходе многогранного процесса принятия решений, стандартизированные размеры эффектов, не зависящие от контекста, предоставляют полезную дополнительную информацию, когда нет других жизнеспособных альтернатив.

Настоящее исследование: количественная оценка минимально важных отношений. Настоящее исследование направлено на предоставление информации о центральной тенденции и изменчивости восприятия важности отношений между переменными с помощью контекстно-свободных визуальных представлений отношений. Понимание величины и изменчивости в интерпретациях того, что квалифицируется как значимое отношение в контекстно-свободных условиях, полезно для внесения вклада в базу знаний для исследователей, принимающих решения относительно соответствующих интервалов для проверки эквивалентности, интерпретаций величины величины эффекта, установления наименьшего значения значимый размер эффекта в анализе мощности и т. д. в условиях, когда важен контекст. В исследовании, используя методологию, аналогичную [Rusticus & Eva \(2016\)](#), участники просы указывать, иллюстрируют ли представленные цифры (которые либо изображают две перекрывающиеся групповые гистогрaммы, либо корреляцию между двумя непрерывными переменными) значимую взаимосвязь. Это исследование также направлено на равнение того, как восприятие того, что представляет собой значимые отношения, различается между статистически обученными и статистически неподготовленными участниками. Мы надеемся, что это исследование предложит некоторые первые шаги к предоставлению исследователям ценных рекомендаций о том, как количественно определить, когда отношения становятся значимыми; вместе с конкретным контекстом исследования эта информация может помочь исследователям понять, что представляет собой минимальную значимую ассоциацию.

МЕТОД

Участники. Это

исследование получило этическое одобрение Подкомитета по рассмотрению участников-людей (Сертификат: 2017 - 001) и с соответствовало руководящим принципам Канады по этике исследований Tri-Council. Участники были набраны с использованием двух различных методов. Сначала были набраны 252 студента, изучающих курс «Введение в психологию» в крупном канадском университете. Эти участники получили кредит курса за участие в исследовании. Мы также наняли 65 аспирантов и преподавателей факультетов психологии Канады по электронной почте. Из этих участников 18 были зачислены в магистерскую программу по психологии, 4 указали, что они завершили магистерскую программу по психологии, 23 были зачислены в докторскую программу по психологии, а 16 указали, что они получили степень доктора психологии. Выборка состояла из 85 участников мужского пола и 229

Таблица 1. Демографические характеристики выборки.

	N	Возраст (медиана [IQR])	Пол, n	Количество Курсов статистики <small>Среднее (среднее ± стандартное отклонение)</small>	Средняя продолжительность Опыт (Режим)
Статистически обученный	76	27 [10.8]	Ф: 54 М: 20	5,7 ± 1,8	докторантура продолжается
Статистически необученный	241	19 [2,0]	Ф: 175 М: 65	1,3 ± 0,6	Первый год
Общий	317	19 [5.0]	Ф: 229 М: 85	2,4 ± 2,1	Первый год Бакалавриат

Заметки.

Ф, женщина; М, муж.

участники женского пола (три участника предпочли не указывать свой пол). Возраст выборки варьировался от 18 до 69 лет с средним и медианным значением 22,5 года и 19 лет соответственно.

Чтобы изучить потенциальные различия между статистически подготовленными и статистически неподготовленными участниками, всех участников попросили указать количество односеместровых курсов статистики, которые они прошли во время послесреднего образования. Участники считались статистически неподготовленными, если они сообщили о прохождении менее трех односеместровых статистических курсов, и статистически подготовленными, если они сообщили о трех или более односеместровых статистических курсах. Соответственно, из 317 участников 75 участников считались статистически обученными, а 242 участника считались статистически неподготовленными. Как правило, классифицировать непрерывную переменную никогда не бывает хорошей идеей, однако на основе статистической подготовки, полученной в канадских университетах, была создана дихотомия (Davidson et al., в печати). В этом случае почти у всех студентов бакалавриата было от нуля до двух курсов статистики (обычно один курс введения в статистику), в то время как у большинства остальных участников (аспирантов/преподавателей) было значительно больше двух курсов статистики, и, следовательно, естественно присуща дихотомия. В частности, для участников бакалавриата 176 участников заявили, что не посещали курсы статистики, 44 заявили, что прошли один курс, 14 участников прошли два курса, восемь — четыре, шесть — пять, и только четыре участника прошли шесть или более курсов. Для аспирантов и преподавателей: 2 участника заявили, что прошли один курс статистики, пятеро заявили, что прошли два, 19 заявили, что прошли три, 17 участников заявили, что прошли четыре, 11 заявили, что прошли пять, а девять — шесть или больше. Дополнительная демографическая информация о выборке представлена в таблице 1.

Критерии исключения были созданы для выявления невнимательных ответов, трудностей с пониманием задания и т. д. Участники исключались из исследования, если они неправильно отвечали на четыре или более из тринадцати возможных контрольных вопросов. Отборочные вопросы включали четыре примера вопросов, которые давали участникам инструкции по проведению опроса; эти примеры вопросов явно направляли участников к выбору правильного варианта. Остальные девять вопросов, использованных как часть критериев исключения, включали крайние примеры ассоциаций или отсутствия ассоциаций между переменными в исследовании. В частности, вопросы из исследования, которые использовались при скрининге невнимательности, включали: (а) очень большую стандартизированную разницу средних (перекрывающиеся)

гистограммы, изображающие d Кэна 1,95 или 2,00), (b) нулевая/минимальная разницас стандартизированных средних (перекрывающиеся гистограммы, изображающие d Кэна 0,00 или 0,05), (c) очень большая корреляция между переменными (диаграммы рассеяния, изображающие корреляции 0,60 или -0,60) и (d) нулевая/минимальная корреляция между переменными (диаграммы рассеяния, изображающие корреляции -0,05, 0,00 и 0,05). Основываясь на этих контрольных вопросах, 105 участников были исключены из анализа данных, в результате чего осталось 212 участников (54 в статистически обученной группе и 158 в статистически необученной группе).

Меры Во-первых,

участникам были представлены цифры, отображающие разную степень взаимосвязи, и их попросили классифицировать взаимосвязь как значащую или нет. Слово «значащий» использовалось в наших измерениях, чтобы побудить участников оценить величину различия или отношения в фигуре. Были показаны два разных типа рисунков: (1) рядом расположенные гистограммы; и (2) диаграммы рассеяния.

Гистограммы отображали два распределения, которые были разделены популяционным коэффициентом Кэна в диапазоне от 0,00 до 2,00 с шагом 0,05 (всего 41 параллельная гистограмма), с размером выборки на группу $N = 1000$. Первая гистограмма в каждой паре имела среднее значение совокупности, равное 0, в то время как вторая имела среднее значение совокупности, установленное для каждого из следующих значений d Кэна. Формы распределения были нормальными, и все они имели стандартное отклонение популяции, равное единице. Эти цифры были сгенерированы с помощью функции `hist` из графического пакета в R (R Core Team, 2016). Для каждой цифры участники просили выбрать один из двух вариантов: (1) существует значимая разница между оценками групп или (2) нет значимой разницы между оценками групп. Пример фигуры гистограммы, которую видели участники, представлен на рис. 1.

Диаграммы рассеяния отображали корреляции на сечениях в диапазоне от -0,60 до 0,60 с шагом 0,05 (всего 25 диаграмм рассеяния) с размером выборки $N = 1000$. Эти рисунки были сгенерированы с помощью функции `plot` в графическом пакете R (R Core Team, 2016). В каждом случае переменные X и Y были выбраны из двумерного нормального распределения с средним значением, равным нулю, и стандартным отклонением, равным единице. Для каждой диаграммы рассеяния участники просили выбрать один из двух вариантов: (1) существует значимая связь между переменными X и Y ; или (2) Между переменными X и Y нет значимой связи. Пример диаграммы рассеяния представлен на рис. 2.

Чтобы гарантировать, что отдельные цифры не были загрязнены выбросами, статистическими моделями случаев и т. д., пять цифр были сгенерированы для каждой величины эффекта (т. е. для каждого значения d и r), и участникам случайным образом была назначена одна из пяти цифр для каждого из условий размера эффекта.

Процедура Все

участники получили досутк опросу с помощью онлайн-программного обеспечения SurveyMonkey (<http://www.Surveymonkey.com>). После предоставления информированного согласия на свое участие участники попросили предоставить демографическую информацию.

Затем участникам были показаны четыре примера типов фигур, с которыми они столкнутся на протяжении оставшейся части опроса. Два примера гистограмм изображены

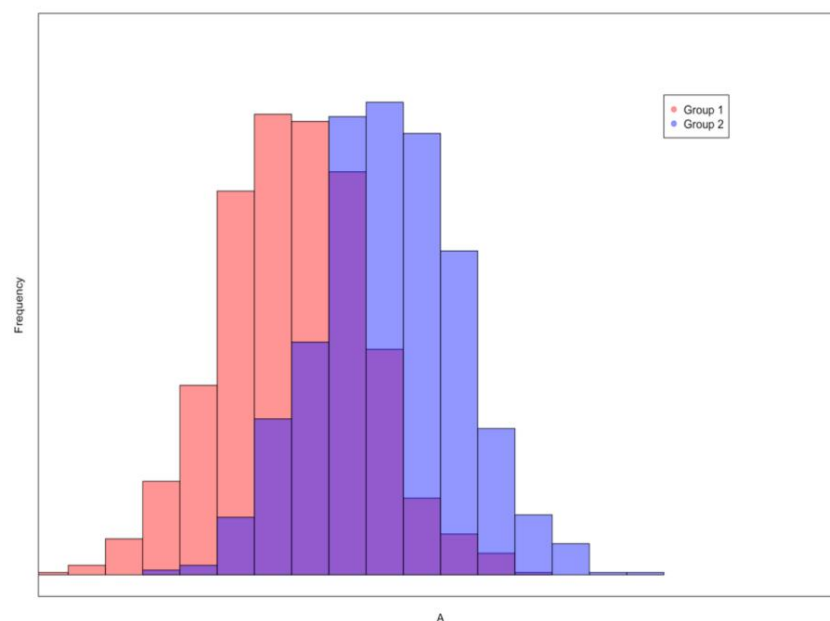


Рисунок 1 Пример перекрывающихся гистограмм, которую видели участники. Красное распределение представляет собой частоту баллов из группы 1, синее распределение представляет частоту баллов из группы 2. Фиолетовая окраска указывает, где распределения перекрываются. В этом примере $d_K = 1.3$.

Полноформатный DOI: [10.7717/peerj.6853/fig-1](https://doi.org/10.7717/peerj.6853/fig-1)

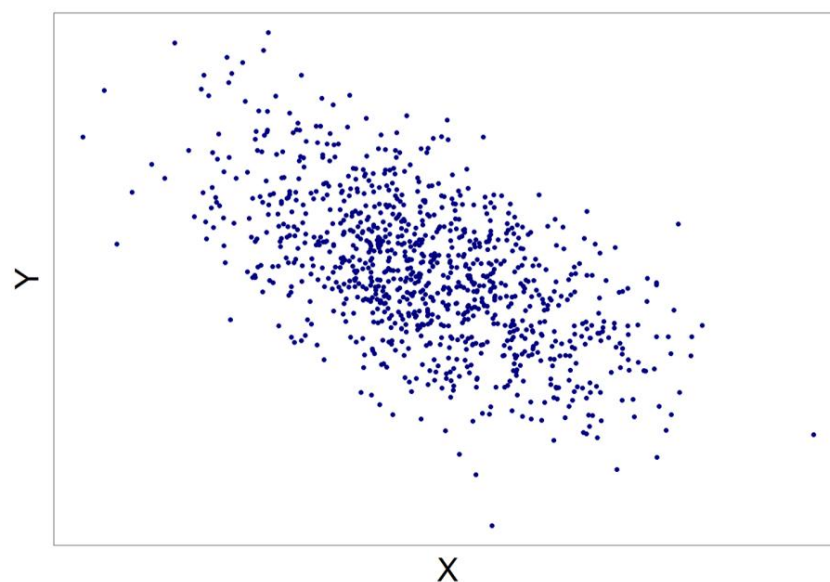


Рисунок 2 Пример диаграммы рассеяния, которую видели участники. В этом случае $r = -0.55$.

Полноформатный DOI: [10.7717/peerj.6853/fig-2](https://doi.org/10.7717/peerj.6853/fig-2)

распределения баллов для двух групп, которые (1) очень мало перекрываются и (2) почти полностью перекрываются. Точно так же примеры диаграмм рассеяния показали участникам (1) почти идеальную корреляцию и (2) нулевую корреляцию. Каждый из примеров также содержал абзац с описанием рисунка, обращая внимание на соответствующие особенности графиков (например, перекрытие в случае гистограмм или связь между показателями X и Y в случае диаграмм рассеяния).

Другими словами, участники провели краткое вводное обучение тому, как интерпретировать параллельные гистограммы и диаграммы рассеяния. Чтобы убедиться, что участники могли точно интерпретировать отображаемые отношения, участников попросили решить, иллюстрирует ли каждый пример значимое различие (гистограммы) или значимую связь (диаграммы рассеяния). На рис. 3 показаны четыре примера рисунков вместе с учебным материалом и вопросами, на которые ответили участники.

После просмотра примеров участникам были показаны перекрывающиеся гистограммы и диаграммы рассеяния, как и на вводном обучении, их попросили оценить каждую из них как отображающую либо значимую, либо незначимую стандартизованную среднюю разницу или корреляцию соответственно. Эти вопросы были идентичны четырем примерам вводного тренинга, однако никаких инструкций над рисунками представлено не было. Порядок представления рисунков был случайным, чтобы контролировать любое потенциальное смещение порядка. Участники могли перемещаться назад и вперед по вопросу по желанию и даже возвращаться к вводным слайдам обучения, которым предшествовали объяснения.

ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

Приведенные ниже результаты основаны на данных, представленных 212 участниками, которые прошли критерии включения. Результаты нашего анализа были в значительной степени схожими независимо от того, использовали ли данные всей выборки или данные 212 участников.

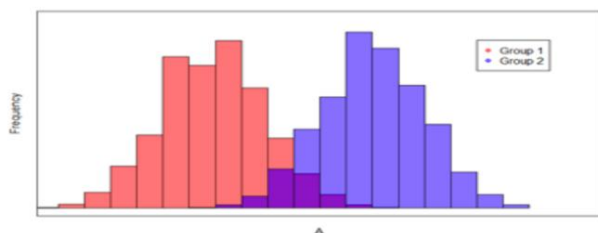
Алгоритм

Коэна был создан для количественной оценки уникального порога каждого человека для обнаружения того, когда оценки двух групп стали значительно отличаться друг от друга. Точка отсечки каждого участника была определена как среднее между (1) максимальными двумя последовательными размерами эффекта, когда участник указал, что оценки групп не отличались значимо, и (2) минимальными двумя последовательными размерами эффекта, когда участник указал, что оценки групп значимо различались. Например, предположим, что участник указал, что оценки групп в последний раз подряд не отличались друг от друга при d Коэна, равном 0,75 и 0,80, а затем оценил оценки групп как разные в первый раз последовательно при d Коэна, равном 0,85 и 0,85, 0,90. Соответственно, точка отсечения участника оценивалась бы как d Коэна 0,825 (среднее значение 0,80 и 0,85).

Уникальные пороговые значения для статистически необученных и статистически обученных групп были примерно нормально распределены. Диапазон уникальных пороговых значений варьировался от $d = 0,22$ до $d = 1,73$ для 156 статистически неподготовленных участников (обратите внимание, что пороговые значения не могли быть рассчитаны для двух статистически обученных и двух статистически неподготовленных участников, поскольку алгоритм требовал определенного образца ответов, которого не было в ответах этих участников). Пороговые значения для статистически неподготовленных участников

A. Example 1: Differences in the Means of Groups

In the figure below, the frequency of each score on the Variable 'A' is displayed for each of two groups (Red = Group 1, Blue = Group 2). Here there are many more lower scores for Group 1 and many more higher scores for Group 2. The purple color indicates where there is overlap between the scores of the groups. In this example, there is very little overlap in the scores between the two groups and thus we would probably indicate that there is a meaningful difference between the scores of the groups. For practice, please answer the question below.

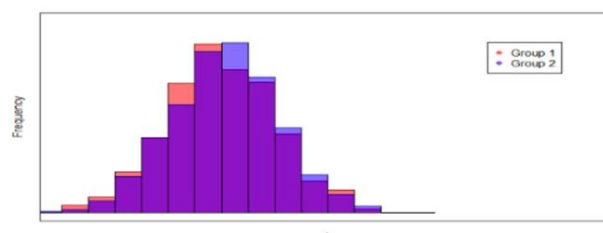


With respect to the figure above, which of the following statements is most applicable?

- There is a meaningful difference between the scores of the groups
- There is NOT a meaningful difference between the scores of the groups

B. Example 2: Differences in the Means of Groups

In the figure below, although the distributions do not overlap completely, there appears to be very little difference between the scores of the participants in each group. In this case, we would probably indicate there is no meaningful difference between the scores of the groups. For practice, please answer the question below.

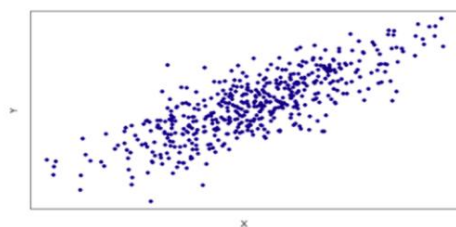


With respect to the figure above, which of the following statements is most applicable?

- There is a meaningful difference between the scores of the groups
- There is NOT a meaningful difference between the scores of the groups

C. Example 3: Associations Among Variables

In the figure below, each blue dot represents where a single individual's score on X and score on Y meet (i.e., where the X score and Y score intersect). Variables are related when high scores on X relate to high scores on Y, high scores on X relate to lower scores on Y, etc. In this example, higher scores on X tend to correspond to higher scores on Y, and lower scores on X tend to correspond to lower scores on Y. Since scores on X tend to relate strongly to scores on Y, we would probably indicate that there is a meaningful association. For practice, please answer the question below.

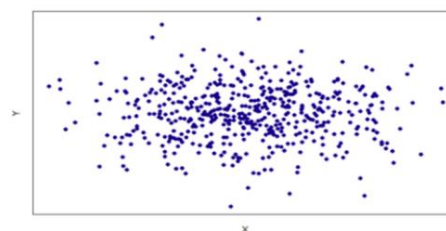


With respect to the figure above, which of the following statements is most applicable?

- There is a meaningful relationship between the variables X and Y
- There is NOT a meaningful relationship between the variables X and Y

D. Example 4: Associations Among Variables

In the figure below, scores on X do not seem to be related to higher or lower scores on Y. Since scores on X do not appear to be related to scores on Y, we would probably indicate that there is no meaningful association. For practice, please answer the question below.



With respect to the figure above, which of the following statements is most applicable?

- There is a meaningful relationship between the variables X and Y
- There is NOT a meaningful relationship between the variables X and Y

Рисунок 3. Примеры вопросов, показанных участникам перед началом исследования. В частности, (А) изображает различия в групповых средних с распределениями, которые имеют очень малое перекрытие, (В) изображает различия в групповых средних с распределениями, которые имеют значительное перекрытие, (С) изображает связь между переменными, где появляются X и Y. быть сильно связанными, а (D) изображает связь между переменными, где X и Y не кажутся сильно связанными.

Полноразмерный DOI: 10.7717/peerj.6853/fig-3

имел среднее значение $d = 0,95$ ($SD = 0,30$), 95% ДИ [0,36, 1,54]. Для 52 статистически подготовленных участников диапазон уникальных пороговых значений варьировался от $d = 0,38$ до $d = 1,92$.

Обученные участники имели средний порог $d = 1,21$ ($SD = 0,42$), 95% ДИ [0,41, 2,01].

Примечательно, что существует значительная вариабельность среди участников с точки зрения того, что представляют собой значимые отношения, и что ни один из участников не имел отсечки ниже того, что Cohen (1988) назвал «маленький» эффект (т. е. $d = 0,2$).

Иерархическая логистическая регрессия выполнялась с использованием функции `glmer` из библиотеки `lme4` пакет в R (Бейтс и др., 2015). Бинарный ответ участников нато, являлся ли две истогораммы были значимо разными, была результирующей переменной модели. Предикторы модели были значения $K_{\text{озна}} d$ (между 0,00 и 2,00), возраст, уровень участников с статистического обучения и взаимодействие между значением $K_{\text{озна}} d$ и статистическим обучением.

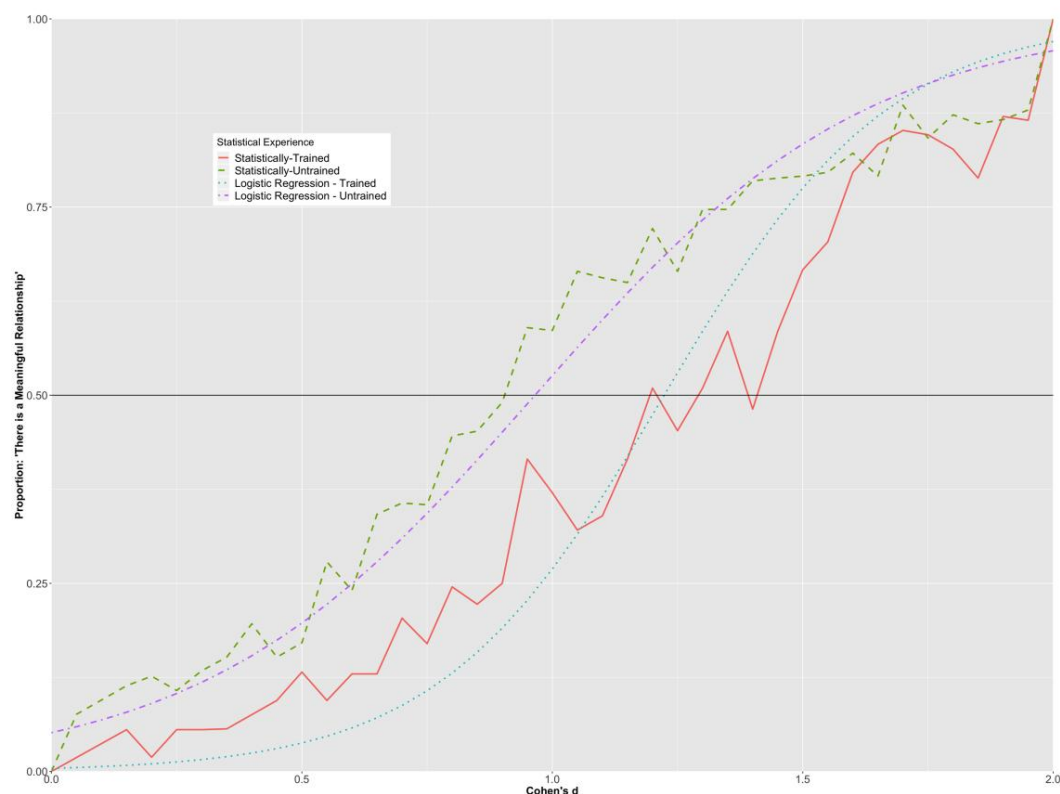


Рисунок 4. Иерархическая смешанная логистическая регрессия для d Коэна. На графике показаны прогнозируемые вероятности статистически обученных и статистически неподготовленных участников, утверждающие, что существует значимая средняя разница между группами, нанесенная рядом с наблюдаемой пропорцией для каждого значения d Коэна.

Полноразмерный DOI: [10.7717/peerj.6853/fig-4](https://doi.org/10.7717/peerj.6853/fig-4)

Общая точность предсказания модели составила 85% с псевдо- R^2 из 0,46 (Накагава и Шильцет, 2013). Статистически значимое взаимодействие наблюдалось между значениями d Коэна и статистической подготовкой участников ($p < 0,001$). Чтобы лучше понять характер взаимодействия, предсказанные вероятности для каждой группы на каждом уровне d Коэна были нанесены на [рис. 4](#) вместе с наблюдаемыми пропорциями для каждого из значений d Коэна в каждой из групп. На [рисунке 4](#) показано, что участникам с более высокой статистической подготовкой требовалось большее значение d Коэна, чем участникам с меньшей статистической подготовкой, чтобы показать, что разница в баллах групп была значимой. Например, хотя вероятности были одинаковыми при малых и больших значениях d Коэна, статистически обученная группа имела более 50% вероятности указания на наличие значимой связи, когда d Коэна достигал 1,00, тогда как статистически необученная группа имела более 50% вероятности констатировать наличие значимой связи, когда d Коэна достигал 1,25. Эти значения были почти идентичны значениям, рассчитанным с использованием не обработанных пропорций для каждого значения d Коэна.

Для дальнейшего изучения взаимодействия между статистической подготовкой участников и различными значениями d Коэна в зависимости от того, расчитывалась ли стандартизированная средняя разница. значимым, мы рассчитываем относительную вероятность того, что обе группы укажут, что

значимое различие между группами при разных значениях d Коэна. Например, при коэффициенте Коэна, равном 0,25, участник статистически обученной группы имел прогнозируемую вероятность 1,3% заявить, что существует значимое различие между группами, в то время как участник статистически необученной группы имел прогнозируемую вероятность 10,4%. Вероятность. Кроме того, учитывая d Коэна 0,90, статистически обученный и статистически нетренированный участник имели 19,2% и 45,1% вероятность утверждения о наличии значимой разницы между группами соответственно. Наконец, при d Коэна, равном 1,5, статистически обученные и статистически нетренированные участники имели 77,5% и 83,4% шанс заявить о наличии значимой разницы между группами соответственно. Как и ожидалось, между двумя группами была очень небольшая разница в крайних значениях d Коэна ($d < 0,10$, $d > 1,60$).

Корреляция Подобно

d Коэна, алгоритм использовался для исследования уникальной точки суждения каждого участника для указания, когда переменные X и Y связаны друг с другом. Поскольку на графиках изображены диаграммы рассеяния как с отрицательной, так и с положительной корреляцией, абсолютное значение корреляции использовалось для расчёта уникального порогового значения для каждого участника.

Те же самые критерии, которые использовались с стандартизированными средними различиями, также применялись в настройке корреляции.

Пороговые значения как для статистически необученных, так и для статистически обученных групп были одинаковыми. Примерно нормально распределены. Когда участники были противоположные ответы на значения корреляции в положительном и отрицательном направлениях (например, отмеченные переменные были значимо связаны для $r = 0,30$, но не были значимо связаны для $r = -0,30$), мы случайным образом назначали значимо связанный или незначительно связанный ответ. Для 140 статистически неподготовленных участников (опять же, пороговые значения невозможно было оценить для семнадцати статистически неподготовленных и девяти статистически подготовленных участников, поскольку их ответы не соответствовали образцу, требуемому алгоритмом), пороговые значения варьировались от $r = 0,08$ до $r = 0,52$, со средним значением $r = 0,32$ ($SD = 0,10$), 95% ДИ [0,13, 0,49]. Для 53 статистически подготовленных участников пороговые значения варьировались от 0,12 до 0,52 со средним значением $r = 0,28$ ($SD = 0,08$), 95% ДИ [0,10, 0,46].

Как и в случае d Коэна, иерархическая логистическая регрессия выполнялась с использованием функции `glmer`. Бинарный ответ участников на то, были ли две переменные значимо связаны, был переменной результата модели. Предикторами модели были значения корреляции (значения от 0 до 0,60, на 0,05, напомнив, что мы взяли абсолютное значение отрицательных корреляций), уровень статистической подготовки участников и возраст, а также взаимодействие между значениями r и уровень статистической подготовки.

Общая точность предсказания модели составила 77%, с $R^2 = 0,55$ (Накагава и Шильдс, 2013 г.). Как и в модели, представляющей d Коэна, существовало статистически значимое взаимодействие между силой корреляции и уровнем статистической подготовки участников ($p < 0,001$). Чтобы дополнительно пролить свет на взаимодействие, предсказанные вероятности для каждой группы на каждом уровне силы корреляции были нанесены на [рис. 5](#) вместе с наблюдаемыми пропорциями для каждого из значений корреляции в каждой из групп.

[На рисунке 5 показано](#), что участники с большей статистической подготовкой имели более низкую прогнозируемую вероятность воспринимать меньшие корреляции как значимые, чем участники с меньшей статистической подготовкой.

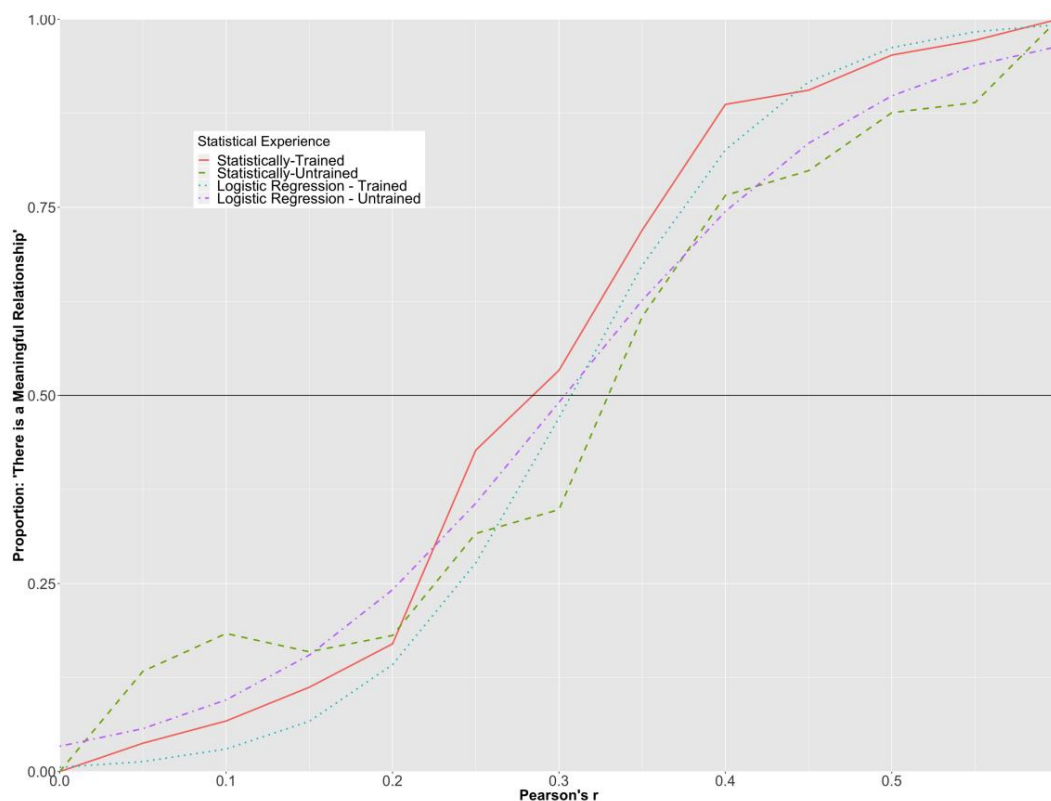


Рисунок 5. Иерархическая смешанная логистическая регрессия для Пирсона. На этом графике показана прогнозируемая вероятность того, что статистически обученные и статистически не обученные участники заявят, что существует значимая связь между двумя переменными, нанесенная рядом с наблюдаемой пропорцией для каждого значения Пирсона.

Полноразмерный DOI: 10.7717/peerj.6853/fig-5

обучения в то время как у них была более высокая прогнозируемая вероятность восприятия более значительных корреляций как значимых, чем у тех, кто прошел менее статистическую подготовку. Однако из графика ясно, что величина взаимодействия была не такой сильной, как для стандартизированных средних различий, и подавляющее большинство объясненной изменчивости можно отнести к различиям в корреляциях популяции. Например, статистически обученная группа с вероятностью более 50 % указала на наличие значимой связи между переменными, когда Пирсона составляло 0,30, в то время как статистически подготовленные участники с вероятностью более 50 % указали, что существует значимая связь между переменными. значимые отношения когда Пирсона был 0,35.

При сравнении относительной вероятности обеих групп указать на наличие значимой связи между переменными было обнаружено, что в то время как человек из статистически неподготовленной группы с большей вероятностью указывал на наличие значимой связи между переменными при $r = 0,15$. (вероятность 15,5 % по сравнению с 6,7 % для статистически подготовленного участника), обе группы имели относительно одинаковую вероятность констатировать наличие значимой связи между переменными при $r = 0,25$ (35,1 % для неопытного участника и 27,7 % для испытуемого). опытный участник). При $r = 0,40$,

статистически обученный участник имел значительно большую вероятность, чем статистически неподготовленный участник, утверждать, что существует связь между переменными (82,6% против 60,1% соответственно).

Когда мы исследовали данные для отрицательных значений корреляции отдельно от данных в положительном направлении мы обнаружили аналогичные результаты как для логистической модели, так и для долей участников, которые указали, что переменные были значимо связаны.

ОБСУЖДЕНИЕ

В этом исследовании изучалось, насколько большим должно быть графическое изображение связи между переменными, чтобы люди могли интерпретировать отношения как значимые. Цель исследования состояла в том, чтобы помочь исследователям, проводящим тестирование эквивалентности, интерпретировать величину величины эффекта, уславливая наименьший значимый эффект в анализе мощности и т. д. в изучении того, когда люди воспринимают минимально важные отношения в условиях без контекста. В прошлом, когда исследователи искали общие (т. е. не зависящие от контекста) рекомендации для интерпретации величины величины эффекта, многие полагались на предложения Коэна для малых, средних и больших эффектов, однако почти полностью игнорировали эмпирические данные. Исследования, изучающие, когда люди начинают обнаруживать значимые отношения

В этом исследовании мы использовали графические изображения стандартизированных средних различий и корреляций, и участники просили указать для нескольких различных степеней взаимосвязи, была ли связь значимой. В частности, участники просили оценить перекрывающиеся ягистогаммы и диаграммы расщепления, отображающие различную степень размера эффекта, и указывали, какие отношения были значимыми.

Когда участники просили оценить перекрывающиеся ягистогаммы, отображающие размеры эффекта в диапазоне от $d = 0,00$ до $d = 2,00$, индивидуальные пороговые значения варьировались от $d = 0,22$ до $d = 1,73$ и имели среднее значение $d = 0,95$ для статистически неподготовленных участников. Для статистически подготовленных участников индивидуальные пороговые значения d Коэна находились между $d = 0,38$ и $d = 1,92$ и имели среднее значение $d = 1,21$. Когда участникам показывали диаграммы расщепления, отображающие взаимосвязь между двумя переменными (в диапазоне от $r = 0$ до $r = 0,60$, включая как положительные, так и отрицательные корреляции), индивидуальные пороговые значения как для статистически неподготовленных участников, так и для статистически обученных участников варьировались от $r = 0,12$ до $r = 0,52$, при этом у статистически обученных участников средняя точка отсечки $r = 0,28$, а у статистически нетренированных участников среднее значение $r = 0,32$.

Соответственно, в условиях, не зависящих от контекста, наше исследование показывает, что большие размеры эффекта, чем ожидаемые, необходимы для того, чтобы участники заявили, что отношения между переменными или различия между группами значимы. Мы говорим «чем ожидалось», поскольку в науках о поведении принято обозначать $r = 0,1$ / $d = 0,2$ небольшим эффектом, $r = 0,3$ /

$d = 0,5$ — эффект средней величины (Cohen, 1988; Nye et al., 2018). Как отметили Rosenthal & Rosnow (1984), r Пирсона, эквивалентный большому эффекту Коэна ($d = 0,80$), равен $r = 0,31$ (стр. 361). Интересно, что в среднем участники нашего исследования начали оценивать ассоциацию как значимые вокруг этих больших отсечений. Интересно, что Szucs & Ioannidis (2017) отметили, что 25-й и 75-й процентиля для величины эффекта в психологии равны $d = (0,29, 0,96)$.

Таким образом, сравнивая результаты нашего исследования с этими прошлыми рекомендациями/результатами,

по-видимому, существует большое расхождение между тем, что участники нашего исследования считают минимально значимой величиной эффекта, и рекомендациями по объявлению эффекта значимой и наблюдаемой величиной эффекта. Учитывая предварительный характер этого исследования и тот факт, что мы исследовали внеконтекстные отношения, мы не утверждаем, что большинство результатов в науках о поведении бесмысленны. Однако эти результаты определенно предполагают, что неохотимы дополнительные исследования того, насколько большим должен быть эффект, прежде чем его можно будет считать значимым; наши результаты, как и всегда, подтверждают, что не рекомендуется простое использование популярных пороговых значений величины эффекта. Таким образом, результаты этого исследования, наряду с многими точками зрения, которые могут быть использованы для понимания того, когда эффект имеет значение, могут быть ценными для исследователей, столкнувшихся с чрезвычайно сложной задачей количественной оценки того, когда связь становится важной.

Взаимодействие между значениями d и уровнем статистической подготовки было интересным найти. Участникам с большей статистической подготовкой требовалось большее стандартизированное среднее различие, прежде чем интерпретировать взаимосвязь как значущую. Одна из возможных причин различия заключается в том, что те, кто посещает больше курсов по статистике (т.е. аспиранты и преподаватели), также имеют больше опыта в оценке того, какой уровень отношений является значимым (или нет) в рамках их собственных исследовательских программ, но не ясно, почему статистически обученной группе потребуется больший эффект, прежде чем объявить взаимосвязь значимой. Взаимодействие между значениями r Пирсона и уровнем статистической подготовки показало, что, хотя между группами с малыми или очень большими корреляциями было очень мало различий с точки зрения того, что представляет собой значимые отношения, больше участников в статистически обученной группе склонны заявлять о корреляциях между $r = 0,35$ и $r = 0,50$ значимо, чем их статистически необученные аналоги. Эффект, как видно из рис. 5, был невелик. Одно из объяснений наших выводов о влиянии уровня статистической подготовки может быть связано с трудностями интерпретации с точки зрения статистически обученной группы (например, Allen, Dorozenko & Roberts, 2016; Lem et al., 2014a). Например, в некоторых случаях статистически подготовленные участники могут с большей вероятностью сосредоточиться на структурных/аналитических, а не на поверхностных компонентах фигур (Lem et al., 2014b). В нашем исследовании внимание к таким структурным особенностям будет включать рассмотрение перекрытия создаваемого двумя расположенными рядом гистограммами. Напротив, те, у кого меньше статистической подготовки, могут иметь тенденцию сосредотачиваться на более поверхностных характеристиках фигур, таких как высота столбцов гистограмм. Стоит отметить, что даже несмотря на то, что в нашем опросе не было ограничения по времени, если какой-либо участник чувствовал себя подавленным или торопливым во время опроса, он мог бы быть более склонен смотреть на поверхностные объекты. Соответственно, требуются дополнительные исследования причин наблюдаемых эффектов взаимодействия.

ВЫВОДЫ

Представленные здесь результаты, наряду с предыдущими выводами, касающимися интерпретации величин взаимосвязей и, что наиболее важно, контекста исследования, могут быть использованы для вынесения суждений о том, когда взаимосвязь становится значимой. Как отмечает Cohen (1988), интерпретация величины эффекта в окончательном анализе зависит от среды, в которой он проявляется. Соответственно, никакие руководства не могут заменить исследователя.

знания об их изучении под рукой. Чтобы представить это в перспективе, в настоящее время многие исследователи используют информацию, такую как пороговые значения Кэна (1988), в сочетании с контекстом исследования, чтобы судить о величине эффекта. Есть надежда, что результаты, полученные в результате этого исследования также могут быть использованы исследователями для лучшего понимания величины эффекта. Эти стратегии уже применяются в таких областях, как онкологические исследования, где графики стандартизированных размеров эффекта, такие как лесные участки, используются для улучшения интерпретации результатов исследования (Bell et al., 2017).

Есть несколько ограничений этого исследования, которые следует подчеркнуть. Во-первых, к различиям между статистически обученными и необученными участниками следует относиться с осторожностью, поскольку в статистически обученной группе было всего 54 участника, и поэтому оценки этих параметров не столь точны, как хотелось бы. Во-вторых, для участников с небольшим опытом работы в статистике или без него обучение в начале опроса может быть недостоверным для того, чтобы они чувствовали себя комфортно при заполнении опроса. Кроме того, было необходимо назначить участникам случайный ответ, когда у них были противоположные ответы между отрицательными и положительными направлениями относительно связи между переменными X и Y для логотипа. Это позволило оценить пороговые точки большего количества участников, но также означало, что иногда было невозможно включить в расчет точную модель реакции каждого участника. Также стоит отметить, что неопытная выборка была получена из пула участников университета и может не быть репрезентативной для более широкого пула неопытных участников, таких как население в целом. Точно так же, хотя мы использовали количество статистических курсов, пройденных участниками, для измерения статистической подготовки, другие факторы, которые не были зарегистрированы, также могли повлиять на статистическую подготовку и опыт, такие как оценки, полученные на курсах, уровень курса и стоимость курса. Наконец, мы исследовали только вопрос о том, что составляет значимую связь в двух групповой стандартизированной средней разнице и двумерной корреляции.

Дальнейшие исследования, как мы надеемся, распространят эти результаты на другие популярные исследовательские проекты.

Подводя итог, наше исследование предоставляет информацию, которую можно использовать вместе с контекстуальными факторами, чтобы помочь исследователям интерпретировать величину эффекта или определить, что представляет собой наименьшую значимую связь (например, для установления границ при тестировании эквивалентности или проведении анализа мощности).

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ И ДЕКЛАРАЦИИ

Финансирование

Это исследование финансировалось Советом социальных и гуманитарных наук Канады (номер гранта: 435-2016-1057). Спонсоры не участвовали в разработке исследования, сборе и анализе данных, принятии решения о публикации или подготовке рукописи.

Раскрытие информации о гранте

Следующая информация о гранте была раскрыта авторами: Совет по социальным и гуманитарным наукам Канады: 435-2016-1057.

Конкурирующие интересы

Авторы заявляют об отсутствии конкурирующих интересов.

Вклад авторов • Натали

Берибиски, Хизер Дэвидсон и Роберт А. Крибби задумали и разработали эксперименты, провели эксперименты, проанализировали данные, подготовили рисунки и/или таблицы, написали или рецензировали черновики статьи, утвердили окончательный вариант.

Человеческая этика

Была представлена ледующая информация касая этических утверждений (т.е. утверждающий орган и/или быссылочные номера):

Комитет по рассмотрению участников-людей Юриског о университета предоставил этическое разрешение на проведение исследования (номер сертификата 2017-001).

Доступность данных В

отношении доступности данных была представлена ледующая информация

Данные доступны в Open Science Framework: https://osf.io/wut42/?view_only=ae5fd48ed18c4868b2b269cdb0045d28.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

- Аллан Т.А., Крибби Р.А. 2013. Оценка эквивалентности или различия между психологические методы лечения исследования недавних интервенционных исследований. *Canadian Journal of Behavioral Science* 45:320–328 DOI [10.1037/a0033357](https://doi.org/10.1037/a0033357).
- Аллен П.Дж., Дорошенко К.П., Робертс Л.Д. 2016. Трудные решения качественное исследование статистического процесса принятия решений с точки зрения студентов-психологов и ученых. *Границы в психологии*: статья 188 DOI [10.3389/fpsyg.2016.00188](https://doi.org/10.3389/fpsyg.2016.00188).
- Баркер Л.Э., Луман Э.Т., Макколи М.М., Чу С.И. 2002. Оценка эквивалентности: альтернатива использованию тестов различия для измерения различий в охвате вакцинацией. *Американский журнал эпидемиологии* 156 (11): 1056–1061.
- Бейтс Д., Маклер М., Болкер Б., Уокер С. 2015. Подгонка линейных моделей с мешанных эффектов с использованием lme4. *Журнал статистического программного обеспечения* 67 (1): 1–48 DOI [10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01).
- Белл М.Л., Фиеро М.Х., Диллон Х.М., Брей В.Дж., Варди Дж.Л. 2017. Статистические противоречия в исследованиях рака: использование стандартизированных графиков размера эффекта для повышения интерпретируемости клинических испытаний, связанных с раком, с результатами, о которых сообщают пациенты. *Анналы онкологии* 28 (8): 1730–1733.
- Коэн Дж. 1988. Статистический анализ мощности для поведенческих наук. 2-е изд. Хилсдейл: Эрльбаум.
- Кук Дж.А., Хислоп Дж.М., Адевуи Т.Э., Харрильд К.А., Альтман Д.Г., Рамзи Ч.Р., Фрейзер С., Бакли Б., Файерс П., Харви И., Бриггс А.Х., Норри Д.Д., Фергюссон Д., Форд И., Вейл Л.Д. 2014. Оценка методов определения элевого различия для рандомизированного контролируемого исследования обзор DELTA (выявление различий в TriAls). *Оценки технологий здравоохранения* 18(28):i-172 DOI [10.3310/hta18280](https://doi.org/10.3310/hta18280).
- Кук Дж.А., Джулиус С.А., Сомес В., Хэммон Л.В., Хьюитт С., Берлин Дж.А., Эшби Д., Эмили Р., Фергюссон Д.А., Уолтерс С.Дж., Уилсон Э.К.Ф., Макленнан Г., Сталлард Н., Ротвелл Дж.К.,

Бланд М., Браун Л., Рамзи Ч.Р., Кук А., Армстронг Д., Альтман Д., Вейл Л.Д. 2018.

Руководство DELTA 2 по выбору целевой разницы, а также по размеру выборки для рандомизированного контролируемого исследования и отчетности. Испытания 19 (1): 1–6 DOI 10.1136/bmj.k3750.

Крибби Р.А., Арпин-Крибби С. 2009. Оценка клинической значимости по редукционному тестированию эквивалентности: расширение подхода к нормативным сравнениям. Psychotherapy Research 19(6):677–686 DOI 10.1080/00220970903224552.

Камминг Г. 2012. От проверки нулевой гипотезы до величины эффекта. Вышли: Тейлор и Фрейс и группа, изд. Понимание новой статистики: размеры эффекта, доверительные интервалы и метаанализ. Нью-Йорк: Академия Рутледж, 21–51.

Дэвидсон Х., Крибби Р.А. 2016. Рекомендации по настройке проверки эквивалентности границ. В: Постер, представленный на ежегодном собрании Ассоциации психологических наук. Вашингтон, округ Колумбия Ассоциация психологических наук.

Дэвидсон Х., Питерс К., Паттон Х., О'Хэган Ф., Крибби Р.А. Статистическое программное обеспечение в Канадских университетских курсах психологии. Преподавание психологии 20 с тр. В печати.

Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinska B, Freedman LS. 2000. Расчет размера выборки для клинических испытаний: влияние убеждений клиницистов. Британский журнал рака 82 (1): 213–219 DOI 10.1054 / bjoc.1999.0902.

Фергюсон Сиджей. 2009. Учебник по величине эффекта: руководство для клиницистов и исследователей. Профессиональная психология исследования и практика 40(5):532–538 DOI 10.1037/a0015808.

Яшке Р., Сингер Дж., Гайятт Г.Х. 1989. Измерение состояния здоровья установление минимальной клинически значимой разницы. Контролируемые клинические испытания 10(4):407–415 DOI 10.1016/0197-2456(89)90005-6.

Келли К., проповедник К.Дж. 2012. О размере эффекта. Психологические методы 17 (2): 137–152. DOI 10.1037/a0028086.

Крушке Дж.К. 2011. Байесовская оценка нулевых значений по редукционному оцениванию параметров и сравнения моделей. Перспективы психологической науки 6 (3): 299–312. DOI 10.1177/1745691612439297.

Крушке Дж. 2013. Байесовская оценка заменяет t-критерий. Журнал экспериментальной психологии: Общие сведения 142: 573–603 DOI 10.1037/a0029146.

Лем С., Онгена П., Вершаффель Л., Друрен В. Ван. 2014а. Интерпретация исторических данных, как кажется? Журнал психологии образования 29 (4): 557–575 DOI 10.1007 / s10212-014-0213-x.

Лем С., Онгена П., Вершаффель Л., Друрен В. Ван. 2014б. Неверная интерпретация экспериментальных данных — подход с двойной обработкой. Psychologica Belgica 54(4):395–405 DOI 10.5334/pb.az.

Мори Р.Д., Роудер Дж.Н. 2011. Подходы фактора Байеса для проверки интервальных нулевых гипотез. Психологические методы 16(4):406–419 DOI 10.1037/a0024377.

Накагава С., Шилзет Х. 2013. Общий и простой метод получения R² из обобщенных линейных моделей с мешанными эффектами. Методы экологии и эволюции 4(2):133–142 DOI 10.1111/j.2041-210x.2012.00261.x.

- Nye CD, Bradburn J, Olenick J, Bialko C, Drasgow F. 2018. Насколько велики мои эффекты? Изучение величины размеров эффектов в исследованиях эквивалентности измерений. Организационные методы исследований DOI 10.1177/1094428118761122.
- Основа команда R. 2016. R: язык и среда для статистических вычислений. Вена: Фонд статистических вычислений R. Доступно на <http://www.R-project.org>.
- Роджерс Дж.Л., Ховард К.И., Весс Дж.Т. 1993. Использование тестов значимости для оценки эквивалентности между двумя экспериментальными группами. Психологический бюллетень 113 (3): 553–565 DOI 10.1037/0033-2909.113.3.553.
- Розенталь Р., Росноу Р.Л. 1984. Основы поведенческих исследований: методы и анализ данных. Нью-Йорк: Макгроу-Хилл.
- Родер Дж. Н., Спекман П.Л., Санд, Мори Р. Д., Айверсон Г. 2009. Байесовские t-тесты для принятия и отклонения нулевой гипотезы. Psychonomic Bulletin & Review 16(2):225–237 DOI 10.3758/PBR.16.2.225.
- Рустикис С.А., Ева К.В. 2016. Определение эквивалентности в оценке и исследованиях медицинского образования работает ли подход, основанный на распределении? Достижения в области медицинского образования 21 (2): 359–373 DOI 10.1007 / s10459-015-9633.
- Рустикис С.А., Ловато С.Ю. 2011. Применение тестов эквивалентности для нескольких групповых сравнений: демонстрация подхода доверительного интервала. Практическая оценка, исследования и оценка 16(7):1–6.
- Щуц Д., Иоаннидис JPA. 2017. Эмпирическая оценка опубликованных размеров и мощности эффектов в недавней литературе по когнитивной нейробиологии и психологии. PLOS Biology 15(3): DOI 10.1371/journal.pbio.2000797.
- Томпсон Б. 2001. Значение, размеры эффекта, пошаговые методы и другие вопросы: веские аргументы двигают поле. Журнал экспериментального образования 70 (1): 80–93 DOI 10.1080/00220970109599499.
- Вебер Р., Попова Л. 2012. Проверка эквивалентности в коммуникативных исследованиях: теория и применение. Методы и меры коммуникации 6(3):190–213 DOI 10.1080/19312458.2012.703834.
- Веллек С. 2010. Проверка статистических гипотез эквивалентности. 2-е изд. Бока-Ратон: Chapman & Hall/CRC.