

**Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет»**

САРАНСКАЯ Ирина Максимовна

Выпускная квалификационная работа

***Анализ воспроизводимости результатов
для различных методов сравнения сетей головного мозга***

Уровень образования: бакалавриат

Направление 50.03.01 «Искусства и гуманитарные науки»

Основная образовательная программа «Свободные искусства и науки»

Научный руководитель:

доцент, кафедра проблем конвергенции естественных
и гуманитарных наук, к.ф. - м.н. Князева Ирина Сергеевна

Рецензент:

PhD, Department of Psychiatry, Psychotherapy and Psychosomatics, Medical
Faculty, RWTH Aachen University, Aachen,
Вотинов Михаил Владимирович

Санкт-Петербург

2023 г.

Содержание

Аннотация	3
Введение	5
Глава 1. Проблема воспроизводимости результатов нейровизуализационных исследований	8
1.1. Исследования сетей головного мозга в состояния покоя . . .	8
1.2. Проверка статистических гипотез	9
1.3. Проблема множественных сравнений	12
Глава 2. Методы и результаты	20
2.1. Описание данных	20
2.2. Методы	21
2.2.1 Оценки воспроизводимости	21
2.2.2 Используемые статистические критерии	22
2.3. Результаты	24
Заключение	29
Список литературы	31
Приложение	39

Аннотация

В данной работе рассматривается проблема воспроизводимости результатов относительно исследования сравнения различных типов состояния покоя – с открытыми и закрытыми глазами, разница между которыми отражается при записи ЭЭГ. Будучи “максимально близкими по поведению и субъективному восприятию” (так как оба являются состояниями покоя и расслабления), состояния покоя с открытыми и закрытыми глазами объективно отличаются паттерном активации нейронов головного мозга. С.Г. Данько был проведён подробный сравнительный анализ этих состояний для формирования набора воспроизводимых различий. При накоплении базы данных ЭЭГ и повторной оценке разницы между этими состояниями результаты отличаются от эксперимента к эксперименту, порождая вопрос об интерпретируемости получаемых результатов относительно уже существующих, о степени их воспроизводимости и достоверности относительно данного эксперимента.

Проблема воспроизводимости остро стоит во многих областях науки, в частности, в нейровизуализации, так как множественные сравнения, возникающие при обработке данных ЭЭГ и в особенности МРТ, приводят к росту ошибки первого рода (к отвержению верной нулевой гипотезы) и размера обнаруживаемых ложных эффектов. Исследователями был выдвинут ряд способов повышения уровня воспроизводимости путём разрешения проблемы множественных сравнений: процедуры контроля групповой вероятности ошибки (FWER, family-wise error rate) и процедуры контроля ожидаемой доли ложных отклонений гипотез (FDR, false discovery rate). Сложность заключается в том, что для каждого отдельного эксперимента набор таких инструментов будет различным в зависимости от дизайна эксперимента, организации данных, нейровизуализационной модальности. Целью данной работы является исследование частного эксперимента и анализ воспроизводимости его результатов для выявления способа обнаружения наиболее устойчивых эффектов.

Проект был выполнен в рамках сотрудничества с Институтом мозга человека им. Н.П. Бехтеревой на основе данных, аккумулированных С.Г. Данько: они представляют собой нормализованные значения межканальной синхронизации (стандартной характеристикой ЭЭГ) для 177 участников экспери-

мента в состояниях покоя с открытыми и закрытыми глазами. Критерием внешней оценки воспроизводимости был выбран коэффициент Сёренсена, или Dice, который позволяет оценивать сходство двух наборов данных, а также статистическая мощность и значение FDR. Для групп участников фиксированного размера, рассматриваемых в данной работе, можно считать уровни значимости каналов, которые можно корректировать с помощью процедур для контроля FWER или FDR. В ходе работы был произведён анализ статистической воспроизводимости, основанный на уровнях значимости с различными способами корректировки на множественность сравнений. В результате получены оценки воспроизводимости эффектов для каждого из методов для различных подгрупп и доверительные интервалы для них. Данные оценки имеют ценность для принятия решения о необходимом количестве испытуемых, а также для оценки надежности результатов использования различных подходов.

Введение

Данная работа посвящена анализу воспроизводимости результатов сравнения сетей головного мозга при использовании различных критериев значимости для выявления эффекта. В основе лежит решение практической задачи сравнения характеристик базовых состояний при записи электроэнцефалограммы (ЭЭГ), а именно состояний покоя с открытыми и закрытыми глазами. Базовое состояние – такое состояние системы, относительно которой производится оценка любого другого, отличного от базового, состояния; в исследованиях головного мозга таким состоянием является состояние покоя, характеризующееся отсутствием какого-либо стимула во время фиксации активности головного мозга[1]. Помимо этого, различные параметры ЭЭГ, характеризующие состояния покоя (rsEEG), "предоставляют значимую информацию об индивидуальных различиях"[2], так как именно в этом состоянии производится "определение свойств эндогенно генерируемой нейронной активности"[3]. Можно выделить два подобных состояния: с открытыми (rsEO) и закрытыми глазами (rsEC); при этом, будучи "максимально близкими по поведению и субъективному восприятию"[4], состояния объективно отличаются паттерном активации нейронов головного мозга, так как каждое состояние характеризуется различными движениями глаз и нейронными сетями[2], которые представляют объединения областей мозга, участвующих в реализации того или иного поведения[5]. Интерпретация результатов фиксации такой активности осложняется из-за того, что "когнитивное содержание независимых от стимулов мыслей нелегко связать с объективно измеримыми показаниями"[3]. Широкий спектр использования rsEEG, а также сложности в его интерпретации, приводят исследователей к важному вопросу о воспроизводимости результатов такого исследования, поскольку их неустойчивость на протяжении n -ного количества экспериментов снижает статистическую мощность (способность обнаружить действительные эффекты)[2]. В более широком контексте, неспособность выявить истинные эффекты является препятствием для дальнейшего развития научной области[6] и углубления представлений об активности головного мозга.

Проблема воспроизводимости научных исследований остро стоит во

многих областях науки, в частности, социологии и теоретической медицине[7], вычислительной науке[8], психологии[9], нейровизуализации[10][11]. Более того, согласно проведённому в 2016 году опросе “более 70% исследователей были неуспешны в воспроизведении результатов чужих экспериментов”[12]. К тому же, в настоящее время в области нейровизуализации растёт интерес к распространению открытых данных[13][14], на основе которых любой может попытаться воспроизвести полученные в оригинальном эксперименте результаты. Такая открытость позволяет большему количеству людей и на различных платформах (от чего может зависеть полученный результат[15]) повторить анализ, что может служить дополнительной мотивацией искать более надёжные способы оценки полученных результатов, так как невозможность воспроизвести заявленные результаты ставит под сомнение значимость исследования. Всё это заставляет идентифицировать источники возникновения этой проблемы и искать подходы для её решения.

Проблема множественности сравнений идентифицируется как одна из причин неудовлетворительных уровней воспроизводимости результатов, в том числе и в области нейровизуализации [16][17][18][19][20][11][21], так как испытания множественных гипотез, возникающих при обработке нейрофизиологических данных ЭЭГ и, в особенности, магнитно-резонансной томографии (МРТ), приводят к росту ошибки первого рода (к отвержению верной нулевой гипотезы об отсутствии эффекта) и размера обнаруживаемых ложных эффектов. Исследователями был выдвинут ряд способов повышения уровня воспроизводимости путём разрешения этой проблемы с помощью применения различных процедур корректировки "в зависимости от того, скоррелированы ли данные и является ли анализ исследовательским или направленным на проверку определённой гипотезы"[22].

Целью данной работы является оценка воспроизводимости эффектов в подгруппах разных размеров при различных способах корректировки на множественные сравнения. Проект был выполнен в рамках сотрудничества с Институтом мозга человека им. Н.П. Бехтеревой на основе данных, аккумулированных в течении многих лет С.Г. Данько: они представляют собой нормализованные значения межканальной синхронизации, стандартной для ЭЭГ метрики, для 177 участников эксперимента в состояниях покоя с от-

крытыми и закрытыми глазами. Уникальность заключается в том, что в данном случае эксперимент реплицировался с использованием одного и того же оборудования, организация всех экспериментов и последующая обработка данных проводилась одной и той же группой учёных, что исключает различия в технике проведения эксперимента и методах обработки данных и даёт одинаковую ошибку измерения для всех повторных экспериментов. В данной работе были протестированы разные подходы к коррекции на множественные сравнения для выделения значимых отличий и проведен сравнительный анализ воспроизводимости в зависимости от метода и от размера выборки. В качестве критерия воспроизводимости для оценки межканальных отличий использовалась частота воспроизводимости, а для совокупной оценки воспроизводимости паттерна – коэффициент Сёренсена, или Dice, который показывает, насколько воспроизводятся найденные значимые характеристики ЭЭГ в совокупности (в данном случае, каналов, или отведений). Помимо коэффициента Сёренсена, в качестве оценок воспроизводимости были использованы статистические характеристики тестирования гипотеза – мощность теста и доля ложно отвергнутых нулевых гипотез (FDR).

В первой главе данной работы будет рассмотрена предметная область, а именно: причины, обуславливающие низкую воспроизводимость, и предложенные способы разрешения этой проблемы, которые затрагивают как структуру самого эксперимента (количество участников), так и методов его анализа (контроль различных статистических значений). Вторая глава будет посвящена описанию проведённого эксперимента и оценке значимых различий с разными уровнями p -значений без корректировки и с поправками на множественные сравнения.

Глава 1. Проблема воспроизводимости результатов нейровизуализационных исследований

1.1 Исследования сетей головного мозга в состоянии покоя

Нейровизуализационные исследования преследуют целью описание структурной и функциональной организации головного мозга – "от нейронного до системного уровня, в масштабе от миллисекунды до декады"[23] – в целях как и "лучшего понимания природы познания"[24] (которое включает в себя такие когнитивные функции, как внимание, речь, память), так и повышения качества неврологической оценки состояния человека, что в дальнейшем может привести к выдающимся достижениям в исследовании патологий человеческого мозга[25]. Так, благодаря исследованиям в этой области сейчас известно, что нейродегенеративные заболевания (медленно прогрессирующие заболевания нервной системы, характеризующиеся гибелью нервных клеток – болезнь Альцгеймера, Паркинсона, Пика, деменция[26]) характеризуются структурными и функциональными изменениями в головном мозге, которые фиксируются методами нейровизуализации: например, болезнь Альцгеймера характеризуется повреждением сети мозга по умолчанию (Default mode network)[27]. Нейронную сеть необходимо понимать с точки зрения взаимодействия нескольких областей мозга по мере её развёртывания во времени[5]. Для каждой сети характерна функциональная и анатомическая (структурная) связанность. Функциональную связанность можно определить как пространственную созависимость паттернов нейронной активации анатомически отдельных областей головного мозга[28], в то время, как структурная связанность описывает анатомические связи, соединяющие различные нейронные структуры (отдельные нейроны, группы нейронов, различные корковые и подкорковые структуры головного мозга)[29]. Данные о паттернах нейронной активации возможно извлечь с помощью современных способов нейровизуализации, таких как фМРТ и ЭЭГ, которые являются одними из самых популярных способов для оценки функциональной связанности ввиду своей доступности, относительно более дорогостоящих модальностей (например, как МЭГ), и при этом достаточно для анализа извлекаемых данных про-

странственного и временного разрешения. Нейронные сети – динамические образования, которые активируются при выполнении какого-то действия (в том числе, и при отсутствии целенаправленного действия) и нахождении в каком-либо эмоциональном состоянии. В XX веке был принят такой взгляд на активность головного мозга, согласно которому "большинство областей мозга остаются неактивными до тех пор, пока их не вызовут для выполнения какой-то конкретной задачи"[30]. По этой причине большое количество исследований того времени было посвящено определению областей мозга, ответственных за формирование определённого поведения[30]. По мере развития нейровизуализационных техник, исследованиям мозга в состоянии покоя стало уделяться больше внимания[30], и сейчас такой тип экспериментального дизайна "имеет всё большее значение при характеристике нормальных и аномальных характеристик функциональной связанности"[31].

1.2 Проверка статистических гипотез

При проверке статистических гипотез возможно несколько исходов, отображённых в таблице 1. Используя классический подход, предполагающий наличие нулевой гипотезы об отсутствии эффекта или связи между явлениями или переменными и альтернативной гипотезы об их существовании, возможно либо принятие, либо отвержение нулевой гипотезы, которая может оказаться как верной, так и ложной. Решение о принятии или отвержении гипотезы принимается исходя из установленного уровня значимости (α в таблице 1).

Значение $1 - \beta$ в таблице 1 соответствует мощности, которая возрастает пропорционально количеству верно отвергнутых нулевых гипотез. Проблема множественности сравнений, с которой сталкиваются исследователи, приводит к росту ошибки первого рода, которая характеризует ситуацию, когда отвергается верная нулевая гипотеза и обнаруживаются ложные эффекты

Нулевая гипотеза H_0	Верна	Ложна
Принята	true positive $1 - \alpha$	false negative β
Отвергнута	false positive α	true negative $1 - \beta$

Таблица 1: Исходы проверки статистической гипотезы

(или преувеличивается их размер). Использование слишком либерального статистического порога повышает долю ошибок первого рода[22]. Ошибка второго рода характеризует неспособность идентифицировать истинные эффекты[22]. Это понятие связано с мощностью, поскольку оно соответствует значению β в таблице 1. Такие ошибки могут возникнуть из-за, наоборот, слишком консервативных порогов, а также малого размера обнаруживаемых эффектов[22]. Эти ошибки скоррелированы: так, пытаясь снизить долю ошибок первого типа, исследователи часто повышают порог уровня значимости, но при этом слишком высокие пороги приводят к постоянному принятию нулевой гипотезы, так как значимые ранее эффекты не достигают новых пороговых значений. Таким образом, повышение воспроизводимости осложнено тем, что при снижении вероятности ошибки первого рода (доли ложноположительных результатов) для разрешения проблемы множественных сравнений в то же время повышается вероятность ошибки второго рода (доли ложноотрицательных результатов), если её оставить без контроля. Множество исследований[32][33][10][34][35][36][37][13] посвящено тому, как найти баланс между этими двумя ошибками и повысить статистическую мощность.

Спонтанная активность головного мозга, проявляющаяся в состояниях покоя, подвергается исследованию с использованием различных нейровизуализационных техник. При анализе данных ЭЭГ и фМРТ выявляется значительный объём исследуемых характеристик. В случае ЭЭГ, нейронная активность регистрируется при помощи специальных электродов, размещённых на поверхности головы в соответствии с различными международными системами[38]. Каждый из этих электродов, количество которых может достигать 256 штук[38], регистрирует электрическую активность в каждый момент времени с высокой временной разрешимостью, в нескольких частотных диапазонах, или ритмах[39]. В результате, исследователь сталкивается с "тысячами или десятками тысяч статистических сравнений"[22]. При работе с фМРТ этот объём становится ещё больше, поскольку "высокоразрешающая съёмка предоставляет сотни тысяч сигналов для каждого индивида, соответствующих измерениям активности мозга в каждом вокселе на протяжении всего эксперимента"[40]. Отмечается, что существует "сотни или тысячи альтернативных подходов к анализу одних и тех же данных"[6], особенно в случае

высокоразмерных наборов данных. В обзорной статье, посвященной различным исследованиям фМРТ, было показано, что для анализа, состоящего из 10 этапов, для каждого из которых было определено от 2 до 4 аналитических стратегий, распространенных в литературе, существует около 7000 уникальных аналитических цепочек, или пайплайнов[41]. Это приводит к тому, что некоторые методики оказывают значительное влияние на результаты, увеличивая ложноположительные результаты (false positives) и снижая уровень воспроизводимости.

В данной работе рассматривается проблема воспроизводимости результатов конкретного исследования С.Г. Данько: в 2006 г. он выдвинул предположение, что "детальное сравнение состояний покоя (rsEO и rsEC) должно стать решением проблемы отсутствия общепринятой концепции значимости характеристики ЭЭГ, подходящих для применения к современным исследованиям ментальных активностей человека"[4]. В то время, как Данько подтверждает значимую разницу между этими состояниями, выделяя набор их усреднённых характеристик и показывая возможность использовать их в качестве базовых (или референтных) состояний, при повторной оценке разницы между этими состояниями в идентичных условиях результаты не воспроизводятся, и возникает вопрос об интерпретируемости получаемых результатов относительно уже существующих, о степени их воспроизводимости и достоверности. Для формирования рекомендаций по использованию различных методов корректировок на множественные сравнения, которые могут повысить уровень воспроизводимости, предполагается выполнение следующих задач:

1. создать серии экспериментов, представляющих собой непересекающиеся подгруппы фиксированного размера,
2. для каждой подгруппы получить оценки значимых различий с разными уровнями p -значений без корректировки и с поправками на множественные сравнения,
3. получить оценки воспроизводимости внутри групп заданного размера для каждого подхода,
4. используя полную выборку в качестве "ground truth" (то есть предла-

гая её в качестве эталона, относительно которого оценивается точность корректировок), получить оценки воспроизводимости для всех рассматриваемых подходов,

5. провести анализ полученных результатов.

1.3 Проблема множественных сравнений

На уровень воспроизводимости исследований влияет статистическая мощность, о чем свидетельствует ряд исследований[16][42][19][43][11][21]. Они подчеркивают, что проблема множественных сравнений является главной причиной низкой мощности, преувеличения значимости эффектов или их полного игнорирования, а также невозможности воспроизвести полученные результаты. Часто используемыми методами для решения этой проблемы являются контроль групповой вероятности ошибки (FWER, family-wise error rate) и ожидаемая доля ложных отклонений гипотез (FDR, false discovery rate)[17][18][43].

Кроме того, применение непараметрического аналога теста на статистическую значимость[16][44][11][21][22] способствует повышению чувствительности, так как он не основывается на предположении о распределении данных с определёнными параметрами. Для повышения статистической мощности и воспроизводимости результатов исследований также рекомендуется использование нескольких подходов. Во-первых, рекомендуется проводить предварительный мощностный анализ для определения оптимального размера выборки с учётом ожидаемых эффектов и допустимого уровня статистической значимости. Во-вторых, предлагается рассмотрение возможности использования менее строгих порогов для проверки гипотез с целью обнаружения большего числа значимых эффектов[42]. В-третьих, рекомендуется рассмотреть анализ группы измерений или объединение данных, а не ограничиваться значимостью отдельных отведений в ЭЭГ или вокселей в МРТ, чтобы учесть возможные взаимосвязи[42][45]. Кроме того, в некоторых случаях рекомендуется увеличение размера выборки на основе применения центральной предельной теоремы[42]. Важно отметить, что каждый отдельный эксперимент требует индивидуального подхода к выбору соответствующих

инструментов и методов в зависимости от его дизайна, структуры данных и используемой нейровизуализационной модальности.

В [16] приводится описание процедуры проведения непараметрического кластерного теста на перемешивание, позволяющего контролировать меру ложных тревог (false alarm rate), которая представляет собой пропорцию ложноположительных событий к общему числу событий и связана с ошибкой первого рода. Использование непараметрического теста в общем случае предоставляет бóльшую исследовательскую свободу, так как такие тесты не основаны на предположении о виде распределения тестовой статистики при верной нулевой гипотезе и не зависят от выбора статистики[16]. Согласно авторам, использование таких тестов позволяет разрешить проблему множественных сравнений и включить предварительные знания о типе эффекта, который, как ожидается, может резко повысить чувствительность (то есть вероятность верно определить положительные эффекты) статистического теста: так, при сравнении матриц пространственно-временных данных в двух экспериментальных условиях можно использовать тот факт, что смежные значения пар электрод-время могут иметь схожий эффект, вследствие чего исследователь может выбрать соответствующую тестовую статистику, которая бы принимала во внимание кластеризацию смежных пар электрод-время: например, размер наибольшего связанного кластера, который превышает некоторый порог, или сумма t -значений данного кластера. Авторы статьи указывают, что данный тест на перемешивание может быть применён как к исследованию данных одного участника (single-subject study), так и к данным группы людей (multiple-subject study). Исследования групп людей делятся на две группы: between-subjects исследование, в котором каждый человек наблюдается в одном единственном условии, и within-subject исследование, в котором каждый человек наблюдается во всех экспериментальных условиях[16]. Однако при этом непараметрические кластерные тесты имеют ряд ограничений: кластеризация зависит только от порога кластеризации, который устанавливается лично исследователем, из-за чего возникает проблема выбора этого значения для максимизации чувствительности теста. В зависимости от типа эффекта (слабый, но распространённый, или сильный, но локальный) порог должен быть разным: более низкий для слабого и более высокий для сильного.

Большое количество данных, полученных во время эксперимента, помогает улучшить пространственное и временное разрешение, вместе с тем создавая проблему множественного тестирования[17], которая вкупе с не откорректированными порогами в зависимости от объёма данных приводит к недостоверным результатам. В [17] исследователи указывают на множественное тестирование как на источник большого количества ложноположительных результатов, подчёркивая необходимость найти баланс между ошибками первого и второго рода: “если наши критерии слишком консервативные, то мы не сможем обнаружить значимые результаты. Если наши пороги слишком либеральны, наши результаты будут загрязнены избытком ложноположительных результатов”[17]. Авторами предлагается ряд корректировок, которые контролируют две меры – FWER и FDR. FWER – вероятность совершения хотя бы одной ошибки первого рода при множественном тестировании гипотез, а FDR – ожидаемая доля правдивых нулевых гипотез, которые были отвергнуты; и эти значения должны быть включены в итоговый обзор результатов работы для демонстрации того, насколько исследование подвержено ошибкам первого рода. Корректировка этих значений включает в себя либо ограничение значения FWER: так, ограничение FWER до 0.05 будет означать, что существует 5%-ная вероятность обнаружения одного или больше ложноположительных результатов среди всех тестов (применяются следующие методы: поправка Бонферрони, Гауссовская теория случайных полей, непараметрическая корректировка на перемешивание); либо ограничение значения FDR: установка значения в 0.05 означает, что в среднем 5% полученных результатов являются ложноположительными (поправка Беньямини-Хохберга (LSU), поправка Дженовезе). Контроль FDR – это более слабая мера контроля ложноположительных результатов, так как он не стремится полностью исключить такие результаты, а скорее дать оценку их присутствию и контролировать их “настойчивость”, то есть присутствие среди результатов[17]. При таком более слабом контроле, однако, удаётся повысить мощность исследования и обнаружить значимые эффекты.

В [42] на основе анализа фМРТ данных проведён анализ воспроизводимости слабых, но пространственно распространённых, эффектов, и сильных, но локализованных, эффектов. Так как авторами было показано, что

слабые эффекты воспроизводились в меньшей степени, на основании этого предложено несколько способов для увеличения мощности и, следовательно, повышения воспроизводимости. Первым они называют увеличение размера выборки, которое приводит к увеличению мощности благодаря центральной предельной теореме (ЦПТ). Исследователи также предлагают использовать неклассические пороги, позволяющие найти баланс между ошибкой первого и второго рода, учитывая специфические характеристики исследования. В работе упоминается мощностный анализ, который направлен на определение минимального размера выборки в зависимости от желаемого уровня значимости, размера эффекта и мощности. Проведение мета-анализа как части преэкспериментального анализа позволяет выявить некие паттерны активации отдельных частей сетей головного мозга, которые появляются среди большого количества исследований[42].

Более поздние исследования подтверждают эффективность кластерных тестов на перемешивание для разрешения проблемы множественных сравнений. В [43] авторы представляют собственную библиотеку на языке Python, которая позволяет использовать различные техники для контроля FWER и FDR. Для контроля FWER предлагаются следующие методы: основанные на теории случайных полей, тесты на перемешивание, поправка Бонферрони и Холма-Бонферрони, поправка Шидака и поправка Хохберга. Для контроля FDR в работе предложены поправка Беньямини-Хохберга, ряд адаптационных FDR методов. После испытания вышеперечисленных методов исследователи пришли к выводу, что лучшие результаты демонстрирует тестирование перемешиванием, которое при этом является самым вычислительно затратным способом среди всех предложенных, демонстрируя таким образом прямую зависимость между вычислительными затратами и мощностью теста. Это исследование также подтвердило, что будучи менее строгим по сравнению с FWER, контроль FDR приводит к большей мощности исследования, несмотря на увеличение доли ложноположительных результатов. Помимо этого, в работе было исследовано положительное влияние предварительных знаний и их использования при проверки гипотез, что подчёркивает необходимость уделять больше внимания тому, как следует объединять данные для статистического анализа[43]. Общий вывод работы акцентирует внимание на том,

что использование чрезмерно строгих порогов для корректировки ошибок не приведёт к желаемому результату, а наоборот – увеличит количество ложно-отрицательных результатов, снижая долю верноположительных, а использование слишком либеральных значений приведёт к возрастанию ложноположительных результатов. Выбор между контролем FWER и FDR, по мнению авторов, зависит от "стоимости" ложноположительных результатов в рассматриваемом исследовании. Поскольку FWER можно контролировать с помощью тестов на перемешивание, одновременно сохраняя высокую мощность, авторы предлагают использовать этот вариант, при условии, что связанная с этим вычислительная нагрузка не является серьезным ограничением[43].

В [11] в качестве способа разрешения проблемы множественных сравнений указан непараметрический кластерный тест на перемешивание с акцентом на то, что такой тест “кластеризует” (объединяет, или группирует) эффекты по временно-, пространственно- и спектрально-смежным характеристикам. При выборе статистики исследователь так же обладает свободой, позволяющей ему делать выбор, основываясь на архитектуре самого эксперимента и биофизиологических данных, так как использование непараметрического варианта кластерного теста даёт такую возможность, в отличие от его параметрических аналогов. При этом авторы отмечают осторожность, которую должны проявить исследователи при выборе порога кластеризации, так как в некоторых случаях специфическое значение может привести к более низкому уровню чувствительности: так, “при использовании высокого порога эффекты, которые имеют небольшую амплитуду, но распространяются на несколько каналов и более длительный период, могут оставаться незамеченными”[11]. Для оценки размеров эффекта используется оценка Cohen’s d, которая представляет из себя стандартизированную разницу между двумя средними. Полученное значение этой оценки показывает разницу между двумя группами в единицах среднего отклонения. Для кластерного анализа, однако, эта оценка не является столь очевидной и прямолинейной, поэтому авторы работы предлагают три подхода к её вычислению. В первом эффект рассчитывается на основе среднего самого крупного кластера, но при этом делается примечание о том, что кластеры могут не иметь чёткую, легко определяемую форму, из-за чего эффекты, основанные на каком-то

определённом кластере не будут воспроизводиться, потому что такой же кластер будет воспроизводиться с малой вероятностью. Второй подход основан на вычислении максимального эффекта внутри каждого кластера: Cohen's d считается для каждого канала и времени в кластере, после чего выбирается самый большой. Недостатком такого подхода является то, что размер эффекта будет колебаться из-за случайной дисперсии, и такая оценка может отражать верхнюю границу размеров эффекта и, собственно, переоценивать его. В третьем случае авторы предлагают лишь приблизительно обозначить кластер какой-то фиксированной геометрической фигурой (например, прямоугольником), после чего посчитать Cohen's d для усреднённых данных, попавших в эту фигуру. Недостатком этого подхода может быть, наоборот, некое преуменьшение размеров эффектов, поскольку при попадании в эту фигуру каналов и временных точек, которые находятся за пределами кластера, значение оценки может отражать нижнюю границу размеров эффектов[11]. При этом, пока все эти оценки транслируются с достаточной прозрачностью, они служат хорошей приблизительной оценкой размера эффектов. При использовании непараметрического теста на перемешивание авторы отмечают, что он не позволяет сделать вывод относительно пространственного, временного или частотного расположения эффекта, а только даёт возможность количественно описать полученные результаты относительно трёх вышеперечисленных измерений. При этом этот тест позволяет смягчить отрицательное влияние множественности сравнений, минимизируя ошибку первого рода, благодаря тому, что “возможно использовать любую статистику для исследования разницы между различными условиями или отношения между данными и независимыми переменными”[11]. Цитируя [16], авторы дают краткое описание методики непараметрического кластерного теста на перемешивание в 4 шага: “1. случайное переназначение порядковых номеров для различных условий; 2. вычисление непараметрической тестовой статистики; 3. повторение этих вычислений для создания перестановочного распределения; 4. определение, куда в распределении попадает статистика для наблюдаемых данных при исходных условиях”. Это стандартная процедура для проведения такого теста, которая встречается во всех перечисленных работах, где он используется.

Аналогичная процедура также рассматривается в [21], где подчёрки-

важется важность применения кластерного теста на перемешивание при анализе данных ЭЭГ. Авторы указывают на необходимость использования его непараметрической версии в связи с тем, что распределение данных ЭЭГ может не соответствовать стандартным описательным моделям. Таким образом, требования параметрического статистического анализа могут противоречить особенностям данных ЭЭГ. Выбор между параметрическим и непараметрическим тестами должен строго зависеть от характеристик конкретного эксперимента и соответствовать их требованиям. Например, непараметрические тесты могут быть предпочтительны, "когда случайные величины имеют равные медианы и различаются по асимметрии и/или эксцессу"[21].

Таким образом, в современной нейровизуализации активно разрабатываются методы и подходы, направленные на повышение уровня воспроизводимости результатов. Исследователи предлагают использовать различные методы контроля ошибок первого и второго рода, включая поправки при множественных сравнениях и скорректированные пороговые значения. Кроме того, внедрение нейрофизиологических характеристик и мощностного анализа для определения оптимального размера выборки являются важными дополнительными шагами в пре- и пост-анализе данных. Непараметрический кластерный тест на перемешивание также предлагается в качестве одного из способов статистической проверки гипотез[21].

Особенности анализа данных ЭЭГ

Исследования на основе электроэнцефалографии (ЭЭГ) и функциональной магнитно-резонансной томографии (фМРТ) характеризуются небольшим количеством участников. В анализируемых статьях, охватывающих период с 2016 по 2022 годы, отмечаются медианные значения выборок, варьирующиеся от 12 до 36 участников[46][47][48][49][50]. Однако выводы этих исследований указывают на необходимость увеличения размера выборки и применения дополнительных процедур для повышения воспроизводимости результатов.

В сравнении с исследованиями, где размер выборки может исчисляться тысячами участников, как в случае социологических опросов, небольшое ко-

личество участников в нейровизуализационных экспериментах объясняется несколькими факторами. Во-первых, сами люди могут испытывать нежелание принимать участие в таких экспериментах из-за требования участия в нескольких сеансах, продолжительность которых может составлять несколько часов и более[51]. Во-вторых, проведение исследований может быть связано с высокими затратами, связанными с использованием специализированного оборудования[52]. Кроме того, возникающие помехи при записи могут привести к непригодности части данных для последующего анализа. Размер выборки напрямую связан с достоверностью и надежностью полученных результатов[50]. Недостаточное количество участников эксперимента приводит к низкой мощности исследования, что снижает его статистическую силу.

Таким образом, одним из ключевых понятий, влияющих на мощность и воспроизводимость исследования, является размер выборки, на который исследователь может в некоторой степени влиять. Вместе с тем, исследователи также подчеркивают важность количества проводимых испытаний[53], особенно при условии, что дисперсия внутри-субъектных исследований превышает дисперсию между-субъектных исследований, что соответствует современным парадигмам нейровизуализационного исследования. Несмотря на появление крупных нейровизуализационных исследований в последние годы, таких как проекты "Human Connectome" и "UK Biobank где размеры выборок достигают сотен тысяч участников[54], большинство экспериментов все же ограничивается скромными размерами (менее 50 участников).

Глава 2. Методы и результаты

2.1 Описание данных

Для реализации цели данной работы в основу практической части были положены данные, агрегированные С.Г. Данько в течение многих лет. В оригинальной работе 2006 г. были рассмотрены состояния покоя с закрытыми и открытыми глазами. Данько взял за основу своего исследования отсутствие общепринятой концепции функциональной значимости характеристик ЭЭГ, указывающих на активацию определённых областей мозга; он говорит о том, что "функциональная значимость индивидуальных частотных диапазонов несомненно является полезной": так, комбинация подавления α -ритма и усиления γ -ритма – один признаков активации коры головного мозга, однако на тот момент отсутствовала единая концепция функциональности значимости характеристик ЭЭГ из-за недостатка внимания к изучению воспроизводимых различий между ними в отличающихся условиях[4].

Так как состояния покоя с открытыми и закрытыми глазами схожи, но характеристики ЭЭГ для них различаются, именно эти состояния были выбраны в качестве предмета исследования. В ходе каждого эксперимента были зафиксированы сигналы ЭЭГ в состоянии покоя с открытыми и закрытыми глазами. В результате проверки статистических гипотез о значимости различий между состояниями были показаны "значимые различия в средней когерентности в высокочастотных диапазонах ($\alpha_2, \beta_1, \beta_2, \gamma$) при переходе от одного состояния к другому"[4]. Обсуждая результаты проведённой работы, Данько говорит о том, что при повторном проведении эксперимента для тех же участников результаты были воспроизведены в частотных диапазонах $\theta, \alpha, \beta_1, \beta_2$ для обоих состояний. Однако при оценке значимых различий по мере накопления данных ЭЭГ новых участников не всегда удаётся воспроизвести эти результаты. В настоящей работе оценка воспроизводимости значимых пар каналов была произведена на данных ЭЭГ, которые представляют значения межканальной синхронизации. Датасет организован в виде таблицы размером 177 на 2394, где столбцы содержат характеристики ЭЭГ для 177 участников эксперимента. На схеме 1 отображён принцип формирования столбцов. В

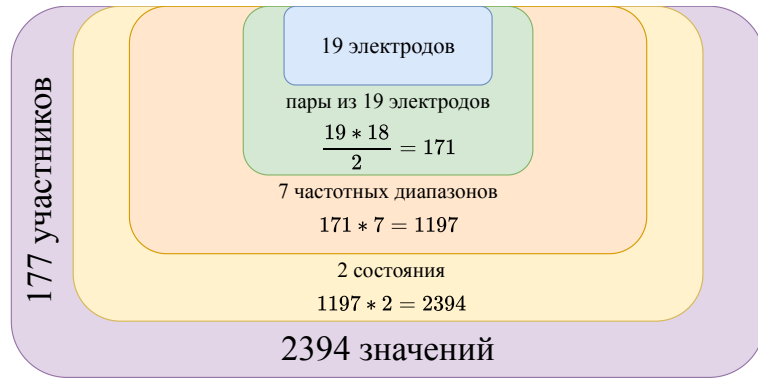


Схема 1: Организация столбцов датасета

датасете представлены значения для 171 пары каналов, образованных из 19 электродов, расположенных по схеме "10-20%". Значения фиксируются для 7 частотных диапазонов: δ (1,5–4 Гц), θ (4–7 Гц), α_1 (7–10 Гц) и α_2 (10–13 Гц), β_1 (13–18 Гц) и β_2 (18–30 Гц), γ (30–40 Гц)[4][55]. Эти данные были получены для обоих состояний – rsEO и rsEC.

2.2 Методы

2.2.1 Оценки воспроизводимости

Для оценки воспроизводимости значимых каналов в настоящем исследовании был выбран коэффициент Сёренсена (или Dice coefficient, DC) – мера сходства, предложенная датским учёным Т. Сёренсеном в 1948 г, который определял коэффициент подобия (Quotient of Similarity, QS) следующим образом:

$$QS = \frac{2c}{a + b} 100,$$

где a – размер одной популяции, b – размер другой популяции, c – количество элементов, принадлежащих обоим популяциям[56]; в математических терминах c можно выразить как пересечение двух множеств. В контексте данной работы, DC позволяет оценить степень воспроизводимости паттерна путём сравнения сходства двух множеств значимых пар каналов. Чем выше значение DC, тем больше сходство между двумя множествами, то есть тем больше пар каналов сохраняется при повторении эксперимента. Выбор данной оценки обусловлен, во-первых, её простотой и интуитивной интерпретацией, а во-вторых, её широким применением как метрики сходства для различ-

ных наборов данных (например, при оценке сегментации и классификации изображений[57][58][59], а также при конструировании и оценке картирования нейронных сетей с использованием фМРТ и ЭЭГ[60][61]). DC представляет собой внешнюю оценку, не требующую дополнительных критериев или параметров для применения. Кроме DC, воспроизводимость оценивается с использованием мощности (доли верно отвергнутых нулевых гипотез, то есть верно обнаруженных эффектов) и FDR (доли ложно отвергнутых нулевых гипотез среди всех найденных гипотез, то есть доли ложных эффектов). Эти оценки предлагается применить для достижения поставленных задач, включая оценку стабильности паттерна при применении различных корректировок для множественных сравнений.

2.2.2 Используемые статистические критерии

Для получения групп значимых каналов в данном исследовании был использован непараметрический тест на перемешивание, подробно описанный в статье [44]. Цель этого тестирования заключается в определении вероятности того, что наблюдаемая тестовая статистика могла быть получена при верной нулевой гипотезе. Для этого процедура тестирования включает следующие этапы: во-первых, необходимо определить тестовую статистику, подходящую для теста в соответствии с предметом исследования, а затем

1. сформировать случайным образом две группы из единого множества результатов для обоих условий с размерами, соответствующими размерам выборок в каждом из экспериментальных условий — нулевая гипотеза заключается в том, что при перемешивании значений для двух условий значимой разницы между ними нет,
2. вычислить тестовую статистику ts_{emp} для этого случайного разделения
3. повторить шаги 1 и 2 множество раз для формирования распределения тестовой статистики,
4. вычислить так называемое "Monte Carlo p -value" то есть пропорцию тестовых статистик, которые больше ts_{emp} , относительно всего количе-

ства тестовых статистик: если это значение меньше критического уровня значимости (определяемого исследователем исходя из необходимой строгости теста, зачастую равного 0.05), то можно прийти к выводу, что экспериментальные условия значимо различаются.

В исследовании, рассматриваемом в данной работе, тестовая статистика представляет собой разницу в значениях межканальной синхронизации для состояний rsEO и rsEC. Для каждого участника случайным образом меняются метки экспериментальных условий, после чего вычисляется тестовая статистика $t_{s_{emp}}$ и усредняется относительно всех испытуемых. Затем процедура перемешивания меток и вычисления среднего между состояниями проводится многократно. Для каждого участника вычисляется p -значение, которое определяется как пропорция значений статистики (среднего после перемешивания), превышающих $t_{s_{emp}}$. Значимые пары каналов определяются как те, чьё p -значение меньше заданного порога значимости. Таким образом, процедура тестирования на перемешивание используется для извлечения значимых пар каналов, которые затем используются для дальнейшей оценки воспроизводимости. Для этого процедура перемешивания и вычисления p -значений повторяется многократно для групп фиксированного размера.

Контроль FWER и FDR Корректировка уровней значимости пар каналов на множественность сравнений реализуется с помощью функциональной библиотеки `multipy`, предложенной в работе [43]. С её помощью проводится контроль FWER и FDR путём корректировок p -значений: методами Бонферрони, Холма-Бонферрони[62], Шидака[63] и Беньямини-Хохберга[64], соответственно. FWER-корректировки предусматривает более строгий контроль, поскольку они направлены на снижение общей вероятности совершить хотя бы одну ошибку первого рода[65]; важно отметить, что они подразумевают, что p -значения отсортированы по мере возрастания. FDR-корректировки допускают ложноположительные результаты и направлены на снижение их пропорции относительно верноположительных исходов. Процедуры корректировок представлены в Табл. 2, где используются следующие обозначения: α – уровень значимости, α^* – полученный уровень значимости, m – общее

количество тестов, k – ранг гипотезы и соответствующего p_k -значения, α_1 – уровень значимости для каждого независимого теста.

Корректировка	Основная процедура
Бонферрони[65]	гипотеза отвергается, если $p < \alpha^* = \frac{\alpha}{m}$
Шидака[63][43]	гипотеза отвергается, если $p < \alpha^* = 1 - (1 - \alpha_1)^{1/m}$
Холма-Бонферрони[62]	гипотеза отвергается, если $p_k < \alpha^* = \frac{\alpha}{m + 1 - k}$; продолжается до первой принятой гипотезы, после которой все последующие будут приняты
Беньямини-Хохберга[64][43]	отвергаются все гипотезы рангом от 1 до самого большого значения k , которое удовлетворяет условию, что $p_k \leq \frac{k}{m}\alpha$

Таблица 2: Описание методов, направленных на контроль FWER (Бонферрони, Холма-Бонферрони, Шидака) и FDR (Беньямини-Хохберга)

Размер эффектов Помимо качественной оценки разницы между состояниями, можно также давать количественную оценку – вычислять размеры эффектов, которые можно в общем определить как "количественное отражение величины явления"[66]. В рамках нейровизуализационных исследований эффекты могут быть слабые и сильные[16], пространственно распространённые и локализованные[42]. Отбирать значимые пары каналов можно в зависимости от размера эффекта, то есть количественной характеристики различий между состояниями, что было реализовано с помощью техники бутстрэпа. Значимыми пары каналов считаются, если размер эффекта (то есть увеличение или уменьшение значения межканальной синхронизации) превысил критическое значение (0.05, 0.1 или 0.15).

2.3 Результаты

Для неоткорректированных p -значений была получена оценка воспроизводимости с помощью DC, демонстрирующая общие тенденции, отражающие прямую зависимость воспроизводимости от размера группы (см. Граф. 1). Также на данном этапе анализа выявлена зависимость воспроизводимости

от частотного диапазона (см. Граф. 2): ритм α_2 показывает бóльшую устойчивость паттерна, что можно объяснить важными изменениями в α -диапазоне при активации коры головного мозга, и долгое время изменения именно на этой частоте считали "главным показателем активации"[4]. α -ритм – самый хорошо изученный ритм активности головного мозга, и именно он фиксируется при "расслабленном состоянии бодрствования с закрытыми глазами"[67]. Меньшие показатели DC характерны для γ -активности, которая отражает работу масштабных сетей головного мозга, вовлечённых в мультимодальную комплексную обработку[68], и в меньшей мере присутствует в состояниях покоя.

Для непересекающихся подгрупп фиксированного размера (от 20 до 90) с помощью DC была оценена воспроизводимость результатов в случае:

1. использования неоткорректированных p -значений при различных уровнях значимости ($\alpha = 0.01$ или 0.05),
2. применения корректировок на множественные сравнения,
3. отбора значимых каналов по размеру эффекта.

Значения DC демонстрируют более высокий уровень при использовании некорректированных p -значений практически для всех размеров подгрупп (см. Граф. 3): в случае использования более лояльного уровня значимости (0.05) воспроизводится больше пар каналов, и в среднем соответствующий DC практически во всех случаях больше (см. Табл. 3, где **отмечено** наилучшее значение для каждого размера), а для строгого уровня значимости (0.01) воспроизводимость в среднем больше, чем при любом способе корректировки на множественность сравнений для групп всех размеров (кроме размеров групп в 60–90 человек). При этом все способы контроля FWER (Бонферрони, Шидак, Холма-Бонферрони) показывают минимальные значения DC, так как они устанавливают строгие уровни значимости из-за большого количества тестов.

Используя в качестве "ground truth" полную выборку, представляющая собой значимые пары каналов, для которых FDR контролируется на уровне 0.05 , были получены оценки воспроизводимости для рассматриваемых подходов. Так как оценка воспроизводимости только с помощью DC не позволяет

	LSU	Uncor 0.01	Uncor 0.05	Eff 0.05	Eff 0.1	Eff 0.15
20	0.19	0.328	0.412	0.415	0.391	0.366
25	0.31	0.386	0.459	0.451	0.428	0.404
30	0.411	0.448	0.511	0.495	0.47	0.452
35	0.447	0.477	0.535	0.519	0.492	0.473
40	0.491	0.509	0.57	0.545	0.52	0.495
45	0.522	0.546	0.595	0.573	0.547	0.523
50	0.569	0.574	0.62	0.598	0.571	0.55
60	0.619	0.619	0.658	0.632	0.609	0.591
70	0.651	0.647	0.684	0.658	0.633	0.602
80	0.679	0.67	0.705	0.682	0.657	0.639
90	0.711	0.702	0.733	0.711	0.687	0.67

Таблица 3: Средние значения DC для групп размерами от 20 до 90 для разных способов корректировки

сделать вывод о том, истинно ли значимые каналы оказываются устойчивыми, значения мощности и FDR призваны восполнить этот пробел. По сравнению с полной выборкой, DC (см. Граф. 4) остаётся устойчиво высоким для некорректированных p -значений, а также для тех пар каналов, размер эффекта которых (разницы между состояниями) превышает критическое значение.

При оценке мощности максимальные значения демонстрируют неоткорректированные p -значения с уровнем значимости 0.05 (см. Граф. 5). Вероятность верно отвергать нулевые гипотезы, то есть не признавать ложные эффекты истинными, для данного эксперимента более высокая при менее консервативном пороге значимости. Для групп малого размера использование неоткорректированных p -значений с уровнем значимости 0.01 позволяет достичь более высокого уровня мощности, чем поправка Беньямини-Хохберга, однако по мере увеличения размера группы ситуация меняется на противоположную (см. Табл. 4): для групп больших размеров относительно ground truth использование строгих уровней значимости обеспечивает меньшую мощность, чем контроль FDR на уровне 0.05. Поправка Беньямини-Хохберга, будучи направленной на контроль FDR, обеспечивает большую мощность, но при этом она рискует принимать больше ложных альтернативных гипотез, повышая вероятность ошибки первого рода. При отборе значимых каналов с опорой на размер эффекта так же получается достичь достаточно высокого уровня

мощности, что, однако, не позволяет оценить истинность принятых гипотез.

	LSU	Uncor 0.01	Uncor 0.05	Eff 0.05	Eff 0.1	Eff 0.15
20	0.125	0.252	0.487	0.504	0.442	0.383
25	0.221	0.33	0.567	0.556	0.486	0.416
30	0.323	0.398	0.636	0.605	0.526	0.446
35	0.409	0.461	0.691	0.649	0.562	0.472
40	0.497	0.521	0.75	0.695	0.6	0.503
45	0.572	0.58	0.787	0.728	0.63	0.532
50	0.644	0.632	0.832	0.768	0.666	0.557
60	0.74	0.709	0.884	0.815	0.709	0.598
70	0.831	0.791	0.93	0.868	0.768	0.642
80	0.892	0.848	0.96	0.906	0.807	0.678
90	0.929	0.89	0.977	0.933	0.838	0.706

Таблица 4: Средние значения мощности для групп размерами от 20 до 90 для разных способов корректировки

Граф. 6 наглядно демонстрирует точно такую же картину: значения без корректировок дают самое большое значение FDR, более строгий уровень значимости при этом приводит к более низким значениям FDR, чем направленная конкретно на контроль этой ошибки поправка Беньямини-Хохберга. Однако при этом величина FDR, отражающая долю ложно отвергнутых нулевых гипотез, тем предпочтительнее, чем она меньше. Строгие поправки (Бонферрони, Шидака и Холма-Бонферрони) демонстрируют наименьшие значения, так как по своей сути они действуют иначе: ложноположительные результаты не допускаются при использовании этих поправок, которые направлены на контроль общей вероятности совершения хотя бы одной ошибки первого рода. В Табл. 5 отражено, как меняются значения FDR с поправкой Беньямини-Хохберга и с помощью использования критического значения 0.15 для отбора эффектов: по мере роста размера выборок эффективность использования корректировки немного падает, так как по мере увеличения тестируемых гипотез возрастает количество ошибок первого рода. Худший результат (отмеченный в Табл. 5) демонстрирует подход, при котором p -значения с уровнем значимости 0.05 не корректируются, однако использование более строгого порога (0.01) позволяет снизить значение FDR лучше, чем поправка Беньямини-Хохберга.

	LSU	Uncor 0.01	Uncor 0.05	Eff 0.05	Eff 0.1	Eff 0.15
20	0.074	0.174	0.309	0.325	0.283	0.25
25	0.106	0.167	0.305	0.304	0.257	0.218
30	0.121	0.159	0.301	0.285	0.236	0.189
35	0.138	0.164	0.312	0.285	0.229	0.176
40	0.153	0.165	0.311	0.276	0.211	0.164
45	0.162	0.162	0.314	0.269	0.202	0.146
50	0.171	0.162	0.319	0.263	0.188	0.128
60	0.201	0.177	0.333	0.268	0.185	0.118
70	0.225	0.187	0.348	0.27	0.176	0.104
80	0.251	0.204	0.368	0.275	0.176	0.097
90	0.275	0.218	0.385	0.282	0.173	0.089

Таблица 5: Средние значения *FDR* для групп размерами от 20 до 90 для разных способов корректировки

При оценке воспроизводимости эффектов (см. Граф. 7) методы контроля FWER показывают ожидаемо наименьшее количество воспроизведённых эффектов, так как строгость поправок не позволяет принять больше; неоткорректированные *p*-значения приводят к обнаружению наибольшего количества эффектов, что вкупе с высоким уровнем мощности позволяет сделать вывод о высоком уровне воспроизводимости, однако, обладая при этом наибольшим значением *FDR*, этот подход склонен выявлять ложные эффекты. Поправка Беньямини-Хохберга позволяет обнаруживать сравнительно меньше эффектов, однако при высоком уровне мощности она лучше контролирует *FDR*, таким образом, являясь надёжным инструментом для корректировки *p*-значений при работе с нейровизуализационными данными.

Заключение

В настоящей работе были представлены различные методы корректировок p -значений на множественность сравнений при работе с нейровизуализационными данными как способ повысить воспроизводительность результатов исследований фМРТ и ЭЭГ. Анализ производился на основе такой характеристики ЭЭГ как межканальная синхронизация, полученной для различных состояний покоя – с открытыми и закрытыми глазами. На основе этих данных была оценена воспроизводительность результатов – паттерна значимых каналов – с помощью коэффициента Сёренсена, или DC, значений FDR и мощности. Четыре распространённые в области метода корректировки (Бонферрони, Шидака, Холма-Бонферрони, Беньямини-Хохберга) были применены к p -значениям, соответствующим гипотезам о значимости пары каналов, с целью контроля FWER и FDR – уровней, отражающих количество ложных решений, принятых по итогам проверки статистических гипотез о наличии эффекта.

DC Оценка воспроизводимости с помощью коэффициента Сёренсена показала, во-первых, прямую зависимость устойчивости паттерна от размера выборки вне зависимости от способа корректировки; во-вторых, было показано влияние нейрофизиологических свойств на воспроизводимость; в-третьих, использование менее строгого (0.05 по сравнению с 0.01) уровня значимости в большей мере способствует устойчивости паттерна даже при отсутствии корректировок на множественность сравнений. При сравнении результатов с полной выборкой DC "лидирующие" способы не потеряли своего положения, а методы, контролирующие FWER, ухудшили результат.

Мощность и FDR При оценке воспроизводимости с помощью мощности (доли верно отвергнутых нулевых гипотез) неоткорректированные p -значения демонстрируют значения, близкие к 1, однако не способны в должной мере контролировать FDR, из-за чего отсутствие поправок на множественность сравнений приводит к обнаружению ложных эффектов. Из всех рассмотренных методов именно поправка Беньямини-Хохберга является оптимальным

вариантом корректировки на множественность сравнений.

Таким образом, проведённый анализ показывает эффективность использования менее строгих уровней значимости и, в особенности, способов корректировки, направленных на контроль уровня FDR, нежели FWER, при работе с нейровизуализационными данными. Эти выводы имеют ценность при принятии решения о требуемом числе испытуемых и оценке надёжности результатов, полученных с использованием различных подходов. Проблема воспроизводительности и сопутствующая ей проблема множественности сравнений представляют огромный пласт вопросов, на которые исследователям необходимо отвечать перед, во время и после проведения эксперимента. Различные методы корректировки статистических значений, а также способы влияния на само устройство эксперимента, должны быть прозрачно отражены в исследовании для накопления практической базы, на основе которой возможно улучшение существующих техник анализа данных нейровизуализационной модальности.

Список литературы

- [1] Resting State fMRI: A Review on Methods in Resting State Connectivity Analysis and Resting State Networks / KA Smith [et al.]. — 2017. — Vol. 30, no. 4. — P. 305–317. — pmid : 28353416.
- [2] Reproducibility of Power Spectrum, Functional Connectivity and Network Construction in Resting-State EEG / Wei Duan [et al.]. — 2021-01-15. — Vol. 348. — P. 108985. DOI: 10.1016/j.jneumeth.2020.108985.
- [3] Snyder Abraham Z., Raichle Marcus E. A Brief History of the Resting State: The Washington University Perspective. — 2012-08-15. — Vol. 62, no. 2. — P. 902–910. DOI: 10.1016/j.neuroimage.2012.01.044.
- [4] Danko S. G. The Reflection of Different Aspects of Brain Activation in the Electroencephalogram: Quantitative Electroencephalography of the States of Rest with the Eyes Open and Closed. — 2006-07-01. — Vol. 32, no. 4. — P. 377–388. DOI: 10.1134/S0362119706040013.
- [5] Pessoa Luiz. Understanding Brain Networks and Brain Organization. — 2014-09. — Vol. 11, no. 3. — P. 400–435. — pmid : 24819881.
- [6] A Manifesto for Reproducible Science / Marcus R. Munafò [et al.]. — 2017-01-10. — Vol. 1, no. 1. — P. 1–9. DOI: 10.1038/s41562-016-0021.
- [7] An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis / Sean Laraway [et al.]. — 2019-03-01. — Vol. 42, no. 1. — P. 33–57. DOI: 10.1007/s40614-019-00193-3.
- [8] Ivie Peter, Thain Douglas. Reproducibility in Scientific Computing. — 2018-07-16. — Vol. 51, no. 3. — P. 63:1–63:36. DOI: 10.1145/3186266.
- [9] Replicability, Robustness, and Reproducibility in Psychological Science / Brian A. Nosek [et al.]. — 2022. — Vol. 73, no. 1. — P. 719–748. — <https://doi.org/10.1146/annurev-psych-020821-114157>.

- [10] Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research / Russell A. Poldrack [et al.]. — 2017-02. — Vol. 18, no. 2. — P. 115–126. DOI: 10.1038/nrn.2016.167.
- [11] Enhancing Reproducibility in Developmental EEG Research: BIDS, Cluster-Based Permutation Tests, and Effect Sizes / Marlene Meyer [et al.]. — 2021-12-01. — Vol. 52. — P. 101036. DOI: 10.1016/j.dcn.2021.101036.
- [12] Baker Monya. 1,500 Scientists Lift the Lid on Reproducibility. — 2016-05-01. — Vol. 533, no. 7604. — P. 452–454. DOI: 10.1038/533452a.
- [13] Open and Reproducible Neuroimaging: From Study Inception to Publication / Guiomar Niso [et al.]. — 2022-11-01. — Vol. 263. — P. 119623. DOI: 10.1016/j.neuroimage.2022.119623.
- [14] The Neuro Bureau Preprocessing Initiative: Open Sharing of Preprocessed Neuroimaging Data and Derivatives / Cameron Craddock [et al.]. — Vol. Conference Abstract: Neuroinformatics 2013. — Front. Neuroinform, 2013-08. DOI: 10.3389/conf.fninf.2013.09.00041.
- [15] Reproducibility of Neuroimaging Analyses across Operating Systems / Tristan Glatard [et al.]. — 2015. — Vol. 9. — URL: <https://www.frontiersin.org/articles/10.3389/fninf.2015.00012> (online; accessed: 2023-04-04).
- [16] Maris Eric, Oostenveld Robert. Nonparametric Statistical Testing of EEG- and MEG-data. — 2007-08-15. — Vol. 164, no. 1. — P. 177–190. DOI: 10.1016/j.jneumeth.2007.03.024.
- [17] Bennett Craig M., Wolford George L., Miller Michael B. The Principled Control of False Positives in Neuroimaging. — 2009-12-01. — Vol. 4, no. 4. — P. 417–422. DOI: 10.1093/scan/nsp053.
- [18] Lindquist Martin A., Mejia Amanda. Zen and the Art of Multiple Comparisons. — 2015. — Vol. 77, no. 2. — P. 114–125. — pmid : 25647751.

- [19] Sassenhagen Jona, Draschkow Dejan. Cluster-Based Permutation Tests of MEG/EEG Data Do Not Establish Significance of Effect Latency or Location. — 2019. — Vol. 56, no. 6. — P. e13335. DOI: 10.1111/psyp.13335.
- [20] A Difference Degree Test for Comparing Brain Networks / Ixavier A. Higgins [et al.]. — 2019. — Vol. 40, no. 15. — P. 4518–4536. DOI: 10.1002/hbm.24718.
- [21] Candia-Rivera Diego, Valenza Gaetano. Cluster Permutation Analysis for EEG Series Based on Non-Parametric Wilcoxon–Mann–Whitney Statistical Tests. — 2022-07-01. — Vol. 19. — P. 101170. DOI: 10.1016/j.softx.2022.101170.
- [22] Cohen Mike X. Analyzing Neural Time Series Data: Theory and Practice. — MIT Press, 2014-01-17. DOI: 10.7551/mitpress/9609.001.0001.
- [23] Bandettini Peter A. What's New in Neuroimaging Methods? — 2009. — Vol. 1156, no. 1. — P. 260–293. DOI: 10.1111/j.1749-6632.2009.04420.x.
- [24] Poldrack Russell A. Can Cognitive Processes Be Inferred from Neuroimaging Data? — 2006-02-01. — Vol. 10, no. 2. — P. 59–63. DOI: 10.1016/j.tics.2005.12.004.
- [25] Neural Network Imaging to Characterize Brain Injury in Cardiac Procedures: The Emerging Utility of Connectomics / B. Indja [et al.]. — 2017-05. — Vol. 118, no. 5. — P. 680–688. DOI: 10.1093/bja/aex088.
- [26] Dugger Brittany N., Dickson Dennis W. Pathology of Neurodegenerative Diseases. — 2017-07-01. — Vol. 9, no. 7. — P. a028035. — pmid : 28062563.
- [27] Hohenfeld Christian, Werner Cornelius J., Reetz Kathrin. Resting-State Connectivity in Neurodegenerative Disorders: Is There Potential for an Imaging Biomarker? — 2018. — Vol. 18. — P. 849–870. DOI: 10.1016/j.nicl.2018.03.013.

- [28] A Review of Structural and Functional Brain Networks: Small World and Atlas / Zhijun Yao [et al.]. — 2015-03-01. — Vol. 2, no. 1. — P. 45–52. DOI: 10.1007/s40708-015-0009-z.
- [29] Sporns Olaf. Structure and Function of Complex Brain Networks. — 2013-09. — Vol. 15, no. 3. — P. 247–262. — pmid : 24174898.
- [30] Raichle Marcus E. The Brain's Dark Energy. — 2010-03. — Vol. 302, no. 3. — P. 44–49. — pmid : 20184182.
- [31] Resting-State Functional MRI: Everything That Nonexperts Have Always Wanted to Know / H. Lv [et al.]. — 2018-08. — Vol. 39, no. 8. — P. 1390–1399. — pmid : 29348136.
- [32] Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience / Katherine S. Button [et al.]. — 2013-05. — Vol. 14, no. 5. — P. 365–376. DOI: 10.1038/nrn3475.
- [33] Gorgolewski Krzysztof J., Poldrack Russell A. A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research — PLOS Biology. — 2016. — Vol. 14, no. 7. DOI: 10.1371/journal.pbio.1002506.
- [34] Szucs Denes, Ioannidis John P. A. Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature. — 2017-03-02. — Vol. 15, no. 3. — P. e2000797. DOI: 10.1371/journal.pbio.2000797.
- [35] Small Sample Sizes Reduce the Replicability of Task-Based fMRI Studies / Benjamin O. Turner [et al.]. — 2018-06-07. — Vol. 1, no. 1. — P. 1–10. DOI: 10.1038/s42003-018-0073-z.
- [36] Poldrack Russell A., Whitaker Kirstie, Kennedy David N. Introduction to the Special Issue on Reproducibility in Neuroimaging. — 2019. DOI: 10.1016/j.neuroimage.2019.116357.

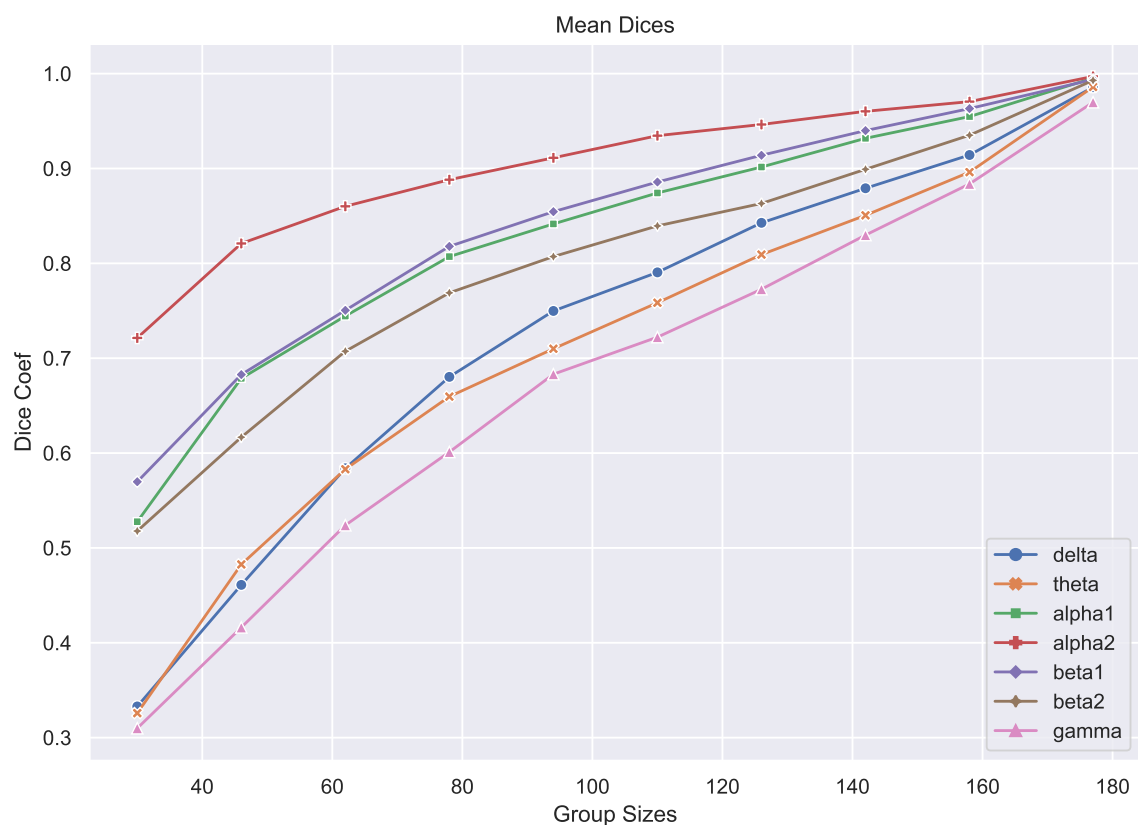
- [37] Opportunities for Increased Reproducibility and Replicability of Developmental Neuroimaging / Eduard T. Klapwijk [et al.]. — 2021-02-01. — Vol. 47. — P. 100902. DOI: 10.1016/j.dcn.2020.100902.
- [38] The Standardized EEG Electrode Array of the IFCN / Margitta Seeck [et al.]. — 2017-10-01. — Vol. 128, no. 10. — P. 2070–2077. DOI: 10.1016/j.clinph.2017.06.254.
- [39] Buzsáki György. Rhythms of the Brain. — Illustrated edition edition. — Oxford University Press, 2006. — ISBN: 978-0-19-982823-4.
- [40] Depthgram: Visualizing Outliers in High-Dimensional Functional Data with Application to fMRI Data Exploration / Yasser Alemán-Gómez [et al.]. — 2022. — Vol. 41, no. 11. — P. 2005–2024. — pmid : 35118686.
- [41] Carp Joshua. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. — 2012. — Vol. 6. — URL: <https://www.frontiersin.org/articles/10.3389/fnins.2012.00149> (online; accessed: 2023-05-14).
- [42] Cremers Henk R., Wager Tor D., Yarkoni Tal. The Relation between Statistical Power and Inference in fMRI. — 2017-11-20. — Vol. 12, no. 11. — P. e0184923. DOI: 10.1371/journal.pone.0184923.
- [43] Puoliväli Tuomas, Palva Satu, Palva J. Matias. Influence of Multiple Hypothesis Testing on Reproducibility in Neuroimaging Research: A Simulation Study and Python-based Software. — 2020-05-01. — Vol. 337. — P. 108654. DOI: 10.1016/j.jneumeth.2020.108654.
- [44] Maris Eric, Schoffelen Jan-Mathijs, Fries Pascal. Nonparametric Statistical Testing of Coherence Differences. — 2007-06-15. — Vol. 163, no. 1. — P. 161–175. DOI: 10.1016/j.jneumeth.2007.02.011.
- [45] Improving Power in Functional Magnetic Resonance Imaging by Moving beyond Cluster-Level Inference / Stephanie Noble [et al.]. — 2022-08-09. — Vol. 119, no. 32. — P. e2203020119. DOI: 10.1073/pnas.2203020119.

- [46] Hyperbolic Trade-off: The Importance of Balancing Trial and Subject Sample Sizes in Neuroimaging / Gang Chen [et al.]. — 2022-02-15. — Vol. 247. — P. 118786. DOI: 10.1016/j.neuroimage.2021.118786.
- [47] Szucs Denes, Ioannidis John PA. Sample Size Evolution in Neuroimaging Research: An Evaluation of Highly-Cited Studies (1990–2012) and of Latest Practices (2017–2018) in High-Impact Journals. — 2020-11-01. — Vol. 221. — P. 117164. DOI: 10.1016/j.neuroimage.2020.117164.
- [48] Reproducible Brain-Wide Association Studies Require Thousands of Individuals / Scott Marek [et al.]. — 2022-03. — Vol. 603, no. 7902. — P. 654–660. DOI: 10.1038/s41586-022-04492-9.
- [49] Pajula Juha, Tohka Jussi. How Many Is Enough? Effect of Sample Size in Inter-Subject Correlation Analysis of fMRI. — 2016. — Vol. 2016. — P. 2094601. — pmid : 26884746.
- [50] The Sample Size Matters: To What Extent the Participant Reduction Affects the Outcomes of a Neuroscientific Research. A Case-Study in Neuromarketing Field / Alessia Vozzi [et al.]. — 2021-01. — Vol. 21, no. 18. — P. 6088. DOI: 10.3390/s21186088.
- [51] A Priori Sample Size Determination for the Number of Subjects in an EEG Experiment / E. Guttmann-Flury [et al.] // 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). — 2019-07. — P. 5180–5183. DOI: 10.1109/EMBC.2019.8857482.
- [52] Distilling Neural Representations of Data Structure Manipulation Using fMRI and fNIRS / Yu Huang [et al.] // 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). — 2019-05. — P. 396–407. DOI: 10.1109/ICSE.2019.00053.
- [53] Power Contours: Optimising Sample Size and Precision in Experimental Psychology and Human Neuroscience / Daniel H. Baker [et al.]. — 2021. — Vol. 26. — P. 295–314. DOI: 10.1037/met0000337.

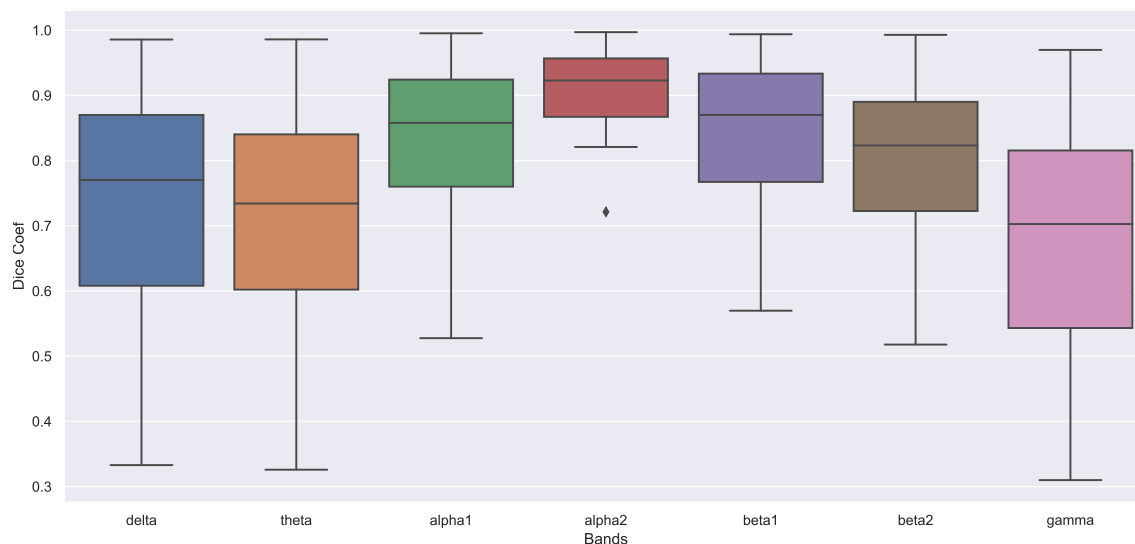
- [54] Smith Stephen M., Nichols Thomas E. Statistical Challenges in “Big Data” Human Neuroimaging. — 2018-01-17. — Vol. 97, no. 2. — P. 263–268. DOI: 10.1016/j.neuron.2017.12.018.
- [55] Boytsova Yu. A., Danko S. G. EEG Differences between Resting States with Eyes Open and Closed in Darkness. — 2010-05-01. — Vol. 36, no. 3. — P. 367–369. DOI: 10.1134/S0362119710030199.
- [56] Sörensen T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content. — 1948. — Vol. V, no. 4. — P. 1–34.
- [57] Guindon Bert, Zhang Ying. Application of the Dice Coefficient to Accuracy Assessment of Object-Based Image Classification. — 2017-01-02. — Vol. 43, no. 1. — P. 48–61. DOI: 10.1080/07038992.2017.1259557.
- [58] Setiawan Agung W. Image Segmentation Metrics in Skin Lesion: Accuracy, Sensitivity, Specificity, Dice Coefficient, Jaccard Index, and Matthews Correlation Coefficient // 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). — 2020-11. — P. 97–102. DOI: 10.1109/CENIM51130.2020.9297970.
- [59] Automated Cartilage and Meniscus Segmentation of Knee MRI with Conditional Generative Adversarial Networks / Sibaji Gaj [et al.]. — 2020. — Vol. 84, no. 1. — P. 437–449. DOI: 10.1002/mrm.28111.
- [60] Presurgical Brain Mapping of the Language Network in Patients with Brain Tumors Using Resting-State fMRI: Comparison with Task fMRI / Haris I. Sair [et al.]. — 2016. — Vol. 37, no. 3. — P. 913–923. DOI: 10.1002/hbm.23075.
- [61] Study of Resting-State Functional Connectivity Networks Using EEG Electrodes Position As Seed / Gonzalo M. Rojas [et al.]. — 2018. — Vol. 12. — URL: <https://www.frontiersin.org/articles/10.3389/fnins.2018.00235> (online; accessed: 2023-03-24).

- [62] Holm Sture. A Simple Sequentially Rejective Multiple Test Procedure. — 1979. — Vol. 6, no. 2. — P. 65–70. — jstor : 4615733.
- [63] Šidák Zbyněk. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. — 1967-06-01. — Vol. 62, no. 318. — P. 626–633. DOI: 10.1080/01621459.1967.10482935.
- [64] Benjamini Yoav, Hochberg Yosef. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. — 1995. — Vol. 57, no. 1. — P. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.
- [65] Austin Stefanie R, Dialsingh Isaac, Altman Naomi S. Multiple Hypothesis Testing: A Review. — 2014. — Vol. 68, no. 2. — P. 303–314.
- [66] Kelley Ken, Preacher Kristopher J. On Effect Size. — 2012. — Vol. 17. — P. 137–152. DOI: 10.1037/a0028086.
- [67] Berger Hans. Über das Elektrenkephalogramm des Menschen. — 1929-12-01. — Vol. 87, no. 1. — P. 527–570. DOI: 10.1007/BF01797193.
- [68] Jia Xiaoxuan, Kohn Adam. Gamma Rhythms in the Brain. — 2011-04-12. — Vol. 9, no. 4. — P. e1001045. DOI: 10.1371/journal.pbio.1001045.

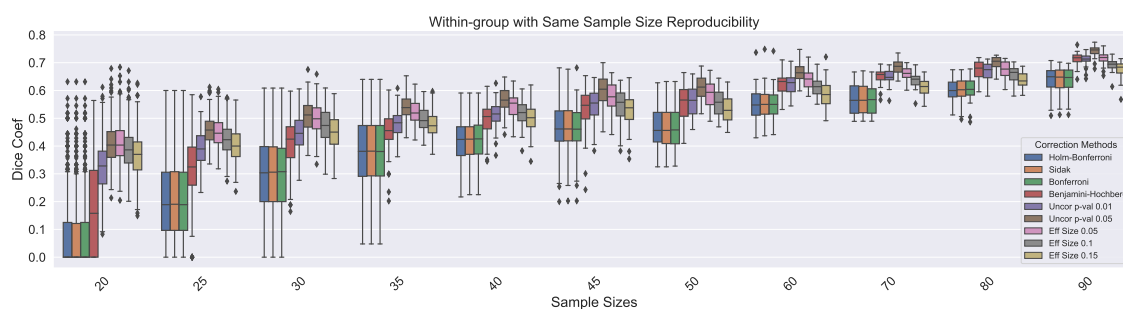
Приложение



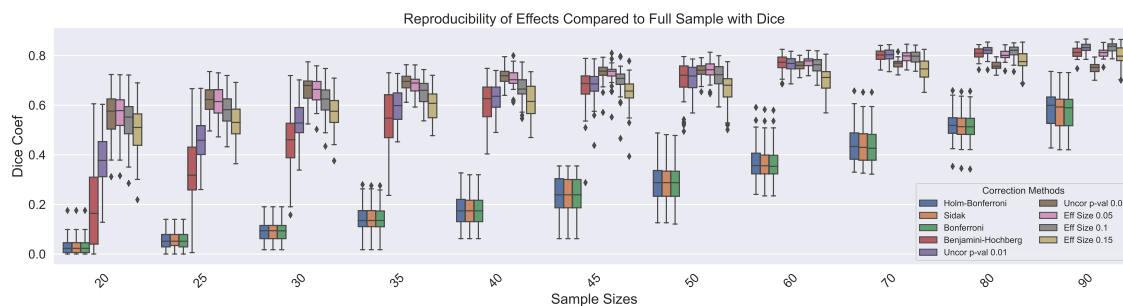
Граф. 1: Среднее значения DC для групп фиксированного размера для каждого частотного диапазона



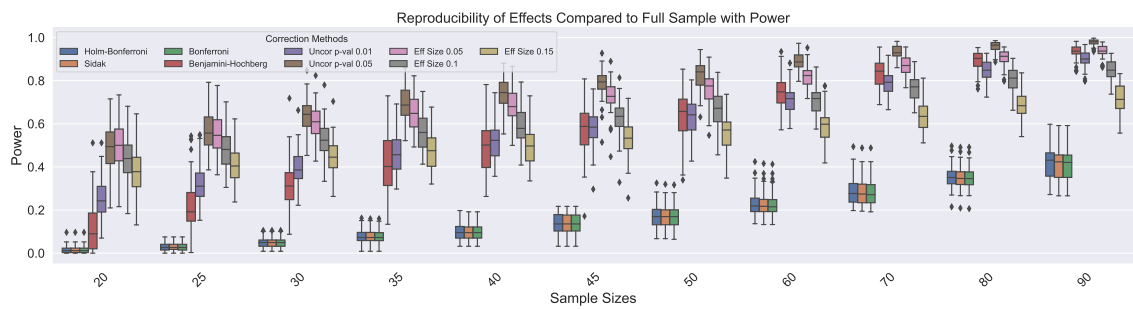
Граф. 2: Значение DC , усреднённого для групп всех размеров для каждого частотного диапазона



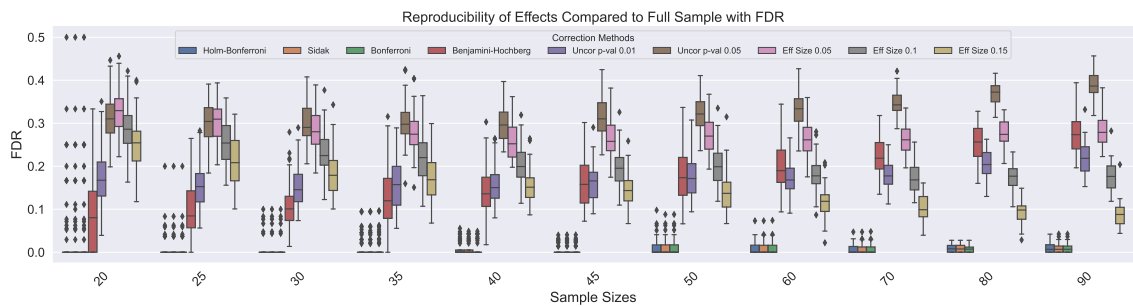
Граф. 3: Распределение значений DC для непересекающихся групп без корректировки p -значений и с поправками на множественность сравнений



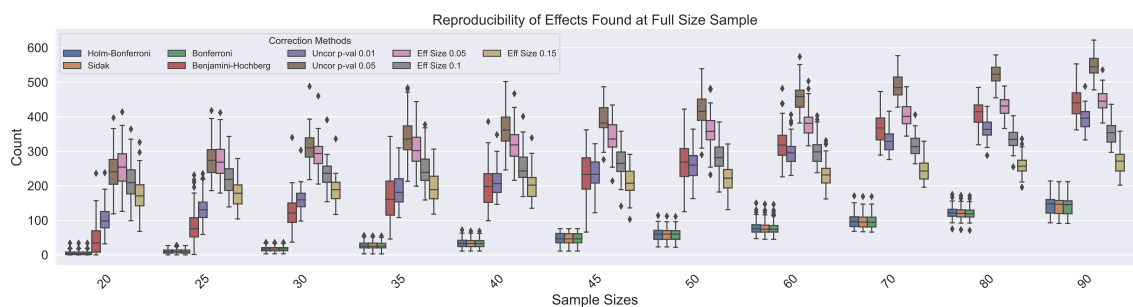
Граф. 4: Распределение значений DC для непересекающихся групп без корректировки p -значений и с поправками на множественность сравнений относительно полной выборки



Граф. 5: Распределение значений мощности для непересекающихся групп без корректировки p -значений и с поправками на множественность сравнений относительно полной выборки



Граф. 6: Распределение значений FDR для непересекающихся групп без корректировки p -значений и с поправками на множественность сравнений относительно полной выборки



Граф. 7: Оценка воспроизводимости эффектов для непересекающихся групп без корректировки p -значений и с поправками на множественность сравнений относительно полной выборки