**К вопросу о воспроизводимость классического эффекта Бергера открывания глаз на изменение когерентности ЭЭГ человека**

**ВВЕДЕНИЕ (постановка проблемы).**

По мере нарастания темпов накопления нейрофизиологических данных, все больше отмечается плохая воспроизводимость ранее наблюдаемых эффектов не только в изучении «тонких» когнитивных функций, но классических ЭЭГ реакций. Одним из таких ЭЭГ-эффектов безусловно является эффект Бергера – при открывании глаз, по сравнению с закрытыми глазами, отмечается падение мощности альфа ритма. Говорят, о так называемой десинхронизации альфа-ритма. Классическим проявлением эффекта Бергера при анализе когерентности является \_\_\_\_\_\_.

Вместе с тем, даже при рассмотрении данного классического эффекта, отражающего весьма выраженную разницу в физиологическом состоянии оперативного покоя, отмечается значительная вариабельность проявлений этого эффекта. При анализе групповых данных, позволяющем экстраполировать выявляемые эффекты при анализе ограниченной выборки испытуемых на всю популяцию, такая вариабельность проявляется как зависимость статистически значимых эффектов от размера выборки.

И далее - Данные Сергея Георгиевича, показывающие вариабельность выявляемых эффектов – как это сделано в отчете.

Следует отметить, что проиллюстрированная проблема варьирования эффекта весьма значима, поскольку в электрофизиологических исследования не так часто выборки испытуемых превышают 100-150 человек, а изучаемые физиологические и психологические состояния гораздо более тоньше, чем разница между состояниями открытых и закрытых глаз ().

Исходя из вышесказанного, возникает методическая проблема оценки надежности выявляемых, на относительно малочисленных объемах выборок, эффектов.

Прежде всего каковы причины вариабельности: 1) при относительно малых выборках (20-30-40 человек) возникают отклонения от нормальности распределения эффектов по группе испытуемых, что также вносит вклад с «стабильность» статистических тестов, 2) в силу множественности сравнений неизбежно увеличивается количество ложно-положительных значений, 3) мы используем стат. тесты, которые не являются оценкой вероятности нахождения эффекта, хотя зачастую таковыми воспринимаются.

Для решения второй проблемы, множественных сравнений, стандартной практикой является использование различных подходов корректировки получаемых уровней значимости, как правило контролируется две характеристики: величина групповой вероятности ошибки FWER или долю ложных отклонений (FDR). Но оба этих подхода используют только информацию о нескорректированных уровнях значимости и никак не учитывают структуру данных. Более того, в рамках этих подходов, предполагается независимости всех тестируемых гипотез, что нарушается при работе с распределенными данными. В представляемой работе, в качестве критерия значимости предлагается использовать пороговое значение доверительного интервала размера эффекта, полученного с помощью непараметрический подхода, основанный на построении бутстрап подвыборок исходной выборки.

Данная работа, является исследованием методического характера, используя базу данных записей ЭЭГ открытые – закрытые глаза, собираемую на протяжении длительного периода времени ставит перед собой следующие цель: получить оценку воспроизводимости значимых различий между состояниями открытые и закрытые глаза в непересекающихся подгруппах разного размера при использования наиболее популярных подходов корректировки на множественные сравнения и предложенного в работе подхода, основанного на интервальной оценке размера эффекта.

1. **ЛИТ ОБЗОР, АКТУАЛЬНОСТЬ ИССЛЕДОВАНИЯ**

**Про открытые закрытые глаза**

Кто-то но не я)

**Проблема воспроизводимости**

Особенностью нейровизуализационных исследований, в том числее ЭЭГ - исследований является небольшое количество участников: согласно работам (Chen et al., 2022) (Marek et al., 2022) указываются медианные размеры выборок от 12 до 36. При этом, выводы, к которым приходят в данных исследованиях, заключаются в необходимости увеличить этот размер, а также провести ряд других процедур для повышения воспроизводимости результатов. Относительно исследований, где размер выборки может исчисляться тысячами участников (например, в социологических опросах), небольшой размер групп участников нейровизуализационных экспериментов объясняется, как сложностью в наборе испытуемых, так как исследования могут идти много часов, так и стоимостью проведения эксперимента. Кроме того, чем сложнее процедура, тем больше вероятность каких-то накладок, часть данных может стать непригодной из-за помех различной природы, возникающих при записи. Размер выборки напрямую связан с достоверностью и надёжностью полученных результатов так как недостаточное количество участников эксперимента приводит к низкой мощности исследования (Vozzi et al., 2021).

Несмотря на то, что в последние годы появляются крупные нейровизуализационные исследования, где размеры выборок доходят до сотен тысяч (например, Human Connectome Project или UK Biobank), однако большинство экспериментов всё же придерживаются более скромных размеров (менее 50 участников) (Smith & Nichols, 2018)

Большое количество данных, полученных во время эксперимента, помогает улучшить пространственное и временное разрешение, вместе с тем создавая проблему множественных сравнений, которая вкупе с не откорректированными порогами в зависимости от объёма данных приводит к недостоверным результатам. В работе (Bennett et al., 2009) авторы 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 (= исследованию будет не хватать мощности). Если наши пороги слишком либеральны, наши результаты будут загрязнены избытком ложноположительных результатов. Авторы рассматривают наиболее популярные корректировки на множественные сравнения: 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. FWER равное 0.05 будет означать, что существует 5%-ная вероятность обнаружения одного или больше ложноположительных результатов среди всех тестов; либо ограничение значения FDR, при котором установка значения в 0.05 означает, что в среднем 5% полученных результатов являются ложноположительными. Контроль FDR - это более слабая мера контроля наличия ложноположительных результатов, так как он не стремится полностью исключить такие результаты, а скорее дать оценку их присутствию и контролировать их “устойчивость”, то есть присутствие среди результатов. При таком более слабом контроле, однако, удаётся повысить мощность исследования и обнаружить значимые эффекты, однако эти значения должны быть включены в итоговый обзор результатов работы для демонстрации того, насколько исследование подвержено ошибкам первого рода. Именно проблема множественных сравнений приводит к низкой мощности исследования и преувеличению значимости эффектов либо к их полному игнорированию, а также к невозможности воспроизвести полученные результаты (Bennett et al., 2009; Cremers et al., 2017; Maris et al., 2007; Meyer et al., 2021; Puoliväli et al., 2020)

В работе Cremers et al. на основе анализа фМРТ данных проведён анализ воспроизводимости слабых, но пространственного 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). Авторы также предлагают использовать неклассические пороги, позволяющие найти баланс между ошибкой первого и второго рода, учитывая специфические характеристики исследования. В работе так же упоминается мощностный анализ, который направлен на определение минимального размера выборки в зависимости от желаемого уровня значимости, размера эффекта и мощности. Также предлагается проведение мета-анализа как части преэкспериментального анализа, для выявления неких паттернов активации отдельных частей сетей головного мозга, которые появляются среди большого количества исследований (Cremers et al., 2017).

Оба перечисленные подхода , и FWER и FDR, основаны на коррекции значимостей индивидуальных тестов на сравнение, то есть используют только набор p\_value полученных для всех тестируемых эффектов, при этом они никак не учитывают структуру данных, неоднородности в распределении параметров. Для учета последних достаточно широко используются непараметрические подходы. Главное преимущество последних заключается в том, что они не подразумевают, что данные должны подчиняться определённому закону распределения. Более того, статистики, которые получаются в процессе зависят исключительно от структуры данных. Среди непараметрических подходов различаются две основные техники: основанные на перемешивании и на бустрапе. При этом, в отличии от именованных параметрических тестов, в каждом конкретном эксперименте можно организовывать процесс уникальным образом, главное определить как именно делать симуляцию и конструировать нулевую гипотезу. В данной работе мы предлагаем использовать метод бутстрап как альтернативу множественным сравнениям. В литературе достаточно много примеров того, каким образом можно делать бутстрап, например (Westfall, 2011). Одним из подходов для тестирования нулевого эффекта является использование доверительных интервалов, как правило если доверительный интервал с заданным уровнем значимости не содержит 0, то гипотеза отвергается, либо можно потребовать чтобы наблюдалось какое-то заданное расстояние до 0 от ближайшей границы интервала. В случае сравнения зависимых выборок самым очевидным решением было бы построение бутстрап интервала для разницы средних, но это практически ничем не отличается от результатов параметрического теста и не возможно предложить какого-то универсального порога расстояния до границы. Вместо этого, мы предлагаем использовать бутстрап интервал размера эффекта. Данная характеристика учитывает кроме разницы в средних, еще и дисперсию подвыборок, кроме того является универсальной характеристикой, не зависимой от масштаба данных.

Оценки мощности поправок и контролируемой доле ложно-отрицательных гипотез получены для случая, когда все гипотезы независимы, данное предположение существенно нарушается при использовании их для анализа межканальной синхронизации ЭЭГ данных в наборе частотных диапазонов. Данные теряют независимость как при переходе от одного частотного диапазона к другому, так что и пространственно: электроды меряют сигналы на поверхности, поступающие от источников внутри головного мозга, информация распределяется по поверхности и в разные отведения может поступать импульс от одной и той же зоны, кроме того, области головного мозга физически связаны между собой, поэтому нельзя считать каждую связь между отведениями отдельной независимой единицей. В связи с этим необходима верификация оценок наиболее часто встречающихся подходов к коррекции на множественные сравнения.

1. **МЕТОДЫ И РЕЗУЛЬТАТЫ**

Основная цель данной работы заключается в оценке уровня неопределенности при построении выводов о найденных тем или иным методов значимых различий между двумя состояниями покоя: открытые и закрытые глаза. А именно получении оценок воспроизводимости значимых различий матриц синхронизаций для двух состояний покоя ЭЭГ для разных размеров выборки и статистических критериях значимости. В качестве критерия воспроизводимости было предложено использовать индивидуальную частоту воспроизводимости значимых межканальных синхронизаций для каждой пары отведений, и коэффициентСёренсена или Dice коэффициент для оценки воспроизводимости паттерна целиком в подвыборках одинакового размера. Кроме того, были проведены оценки воспроизводимости эффектов полученных в полной выборке на подбвыборках меньшего размера. Для этой цели также использовался Dice коэффициент, а также были вычислены коэффициенты мощности и доли ложно отвергнутых гипотез.

**Данные**

Данные представляют собой ЭЭГ записи состояния покоя агрегированные в течение разных физиологических экспериментов, проведённых на базе Института мозга человека им Н.П. Бехтеровой. Всего база данных насчитывает 177 человек в возрасте от 17 до 30 лет . Для каждого испытуемого записывались ЭЭГ сигналы в состоянии покоя с открытыми и закрытыми глазами. Участники сидели в затемнённой комнате в комфортном кресле. Точка для фиксации взгляда (при rsEO) была расположена на мониторе, расположенном на расстояния 1.2 метра. Испытуемым было запрещено впадать в состояния дремоты; если такое происходило (отмечалось на ЭЭГ), участнику вновь проводили инструктаж и возобновляли эксперимент. Если такое происходило во второй раз, то данные ЭЭГ, полученные для этого человека, не использовались. Было использовано 19 электродов, соответствующих международной системе расположения электродов "10-20\%". После удаления артефактов был проведён количественный анализ: для каждого испытуемого в заданном состоянии (rsEO или rsEC) рассчитывали средние значения функции когерентности по каждой паре отведений. Эти оценки производились для спектральных компонентов, усреднённых по частотным диапазонам (7 диапазонов общей частотой от 1,5 Гц до 40 Гц). Матрицы оценок функции когерентности (фазовой синхронизации) были нормализованы. Далее проводился статистический анализ этих матриц синхронизации для обнаружение значимых различий между параметрами ЭЭГ в сравниваемых состояниях.

**Статистические критерии для выявления эффектов**

В данной работе используется только так-называемый mass-univariate подход, то есть когда для каждой пары отведений проводится параметрический t-test с повторным измерениями или что тоже самое t-test на равенство разницы между двумя состояниями нулю. В данной работе были в качестве базы для сравнения использовались неоткорректированные уровни значимости 0.05 и 0.01.

Далее были использованы несколько различных корректировок на множественность сравнений,а именно: методы Бонферрони, Холма-Бонферрони и Шидака для контроля FWER и Беньямини-Хохберга ( Benjamini–Hochberg Linear Step Up , далее LSU) (Benjamini & Hochberg, 1995) соответственно. FWER-корректировки предусматривает более строгий контроль, поскольку они направлены на снижение общей вероятности совершить хотя бы одну ошибку первого рода (Austin et al., 2014). FDR-корректировки допускают ложноположительные результаты и направлены на снижение их пропорции относительно верноположительных исходов. Процедуры корректировок представлены в Табл.1 где используются следующие обозначения: α - уровень значимости: полученный уровень значимости, m - общее количество тестов, k - ранг гипотезы и соответствующего p\_k-значения, α\_1 - уровень значимости для каждого независимого теста.

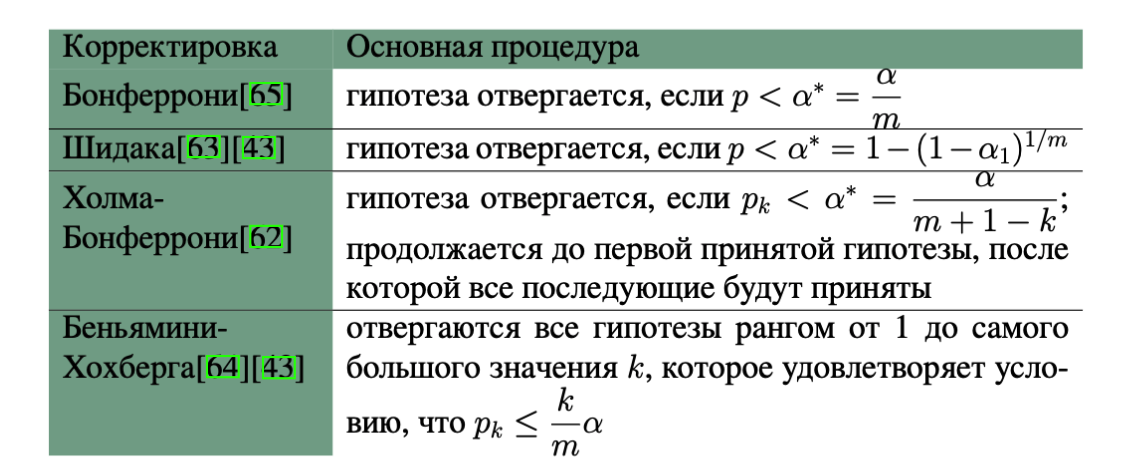


Таблица 1. Способы контроля множественных сравнений

Также была использована корректировка с помощью доверительного интервала размера эффекта. Идея использования бустрап интервалов различных характеристик для тестирования нулевых гипотез не нова (Martin, 2007), однако так как это непараметрический подход, может быть выбрана произвольная статистика для построения доверительного интервала, то универсальных рекомендаций и критерией использования подхода нет. В данной работе в качестве характеристики была выбрана величина размера эффекта. Она является универсальной характеристикой «масштаба» обнаруженного эффекта и позволяет оперировать не только статистической значимостью, но и «практической» ценностью. Для построения распределения размеров эффекта использовалась техника бутстрап с повторением, где для каждого эксперимента симулировалось 1000 экспериментов со случайным добавлением или удалением испытуемых. Далее, из полученного распределения размера эффектов вычислялся симметричный относительно среднего эффекта доверительный 95 процентный интервал и задавался критерий значимости как величина на которую отстоит от нуля ближайшая граница к нулю. Например 0.1 будет соответствовать тому что левая граница доверительного интервала больше или равна 0.1, а -0.1 – правая граница интервала меньше или равна 0.1.

**Метрики воспроизводимости**

В качестве метрик воспроизводимости в работе были использованы следующие характеристики: частота воспроизводимости, коэффициентСёренсена или Dice коэффициент, доля правильно выявленных различий из имеющихся в полной выборке (мощность, power), доля правдивых нулевых гипотез, которые были отвергнуты (fdr)

**Описание использованных характеристик:**

***Коэффициент Сёренсена***

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:

где a- размер одной популяции, b - размер другой популяции, c - количество элементов, принадлежащих обоим популяциям 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 различных наборов данных (например, при оценке сегментации и классификации изображений (Guindon & Zhang, 2017)}, а также при конструировании и оценке картирования нейронных сетей с использованием фМРТ и ЭЭГ (Rojas et al., 2018; Sair et al., 2016) DC представляет собой внешнюю оценку, не требующую дополнительных критериев или параметров для своего использования.

***Мощность и доля ложных эффектов***

Также воспроизводимость можно оценивать с использованием мощности (доли верно отвергнутых нулевых гипотез, то есть верно обнаруженных эффектов) и FDR (доли ложно отвергнутых нулевых гипотез среди всех найденных гипотез). Для использования этих оценок необходимо иметь информацию об истинных эффектах. В данной работе, в качестве истины (ground true) были использованы эффекты, полученные на полной выборке с умеренно консервативной корректировкой на множественные сравнения, основанной на использовании доверительного интервала размера эффекта.

**Описание экспериментов**

Для каждой пары отведений в каждом ритме был проведен параметрический тест на равенство среднего значения нулю для разницы состояний Открытые – Закрытые глаза. Далее были использованы различные поправки на множественное сравнение, а именно для FWER коррекции: поправки Бонферрони и Холма-Бонферрони и поправка Шидака, а для FDR классическая поправка Беньямини-Хохберга. Для поправок была использована библиотека на языке Python multipy (Puoliväli et al., 2020), которая позволяет использовать различные техники для контроля FWER и FDR. Для получения значимых эффектов с помощью техники бутстрап с повторениями был получен доверительный интервал для размера эффекта для каждой пары отведений, а далее были рассмотрены варианты с тремя различными порогами для принятия решения о значимости: [0.05, 0.1, 0.15], отличающихся степенью консервативности.

***Тестирование воспроизводимости.***

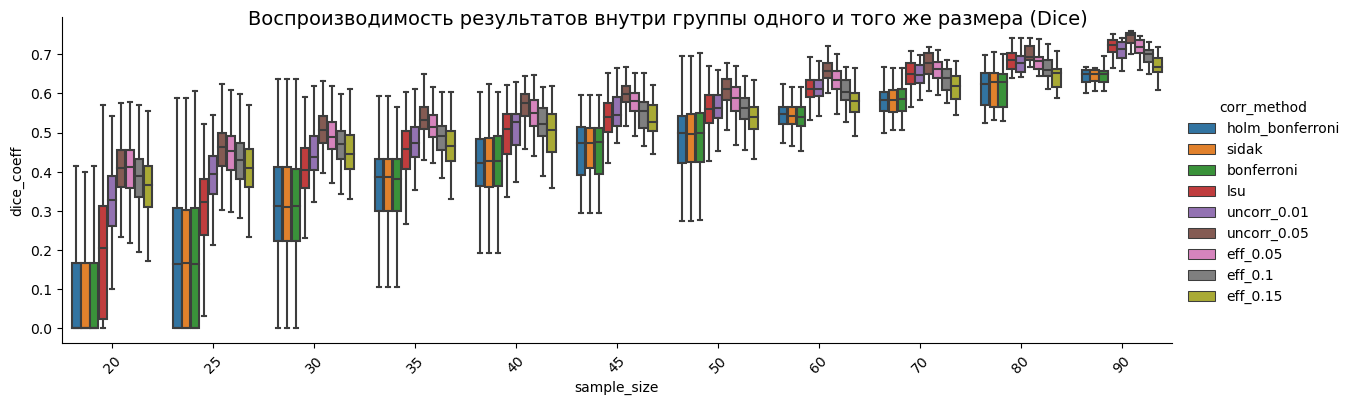
Для тестирования воспроизводимости эффектов была реализована следующая процедура: для каждого исследуемого размера выборки, вся выборка случайным образом разбивалась на непересекающиеся группы этого размера. Далее, было проведено два эксперимента.

В первом, для каждой подгруппы вычислялись значимые эффекты с использованием перечисленных в методах корректировках. Для оценки воспроизводимости результатов мы использовали два параметра: частота воспроизводимости эффекта (число раз когда пара отведений оказывалась значимой) и коэффициент перекрытия или Dice коэффициент, показывающий на сколько паттерн воспроизводится целиком. Затем процедура повторялась: проводилось переразбиение на непересекающиеся подгруппы с вычислением коэффициентов воспроизводимости. С помощью данного эксперимента можно получить ответ на вопрос «Насколько похожи эффекты найденные в подгруппах одного размера между собой».

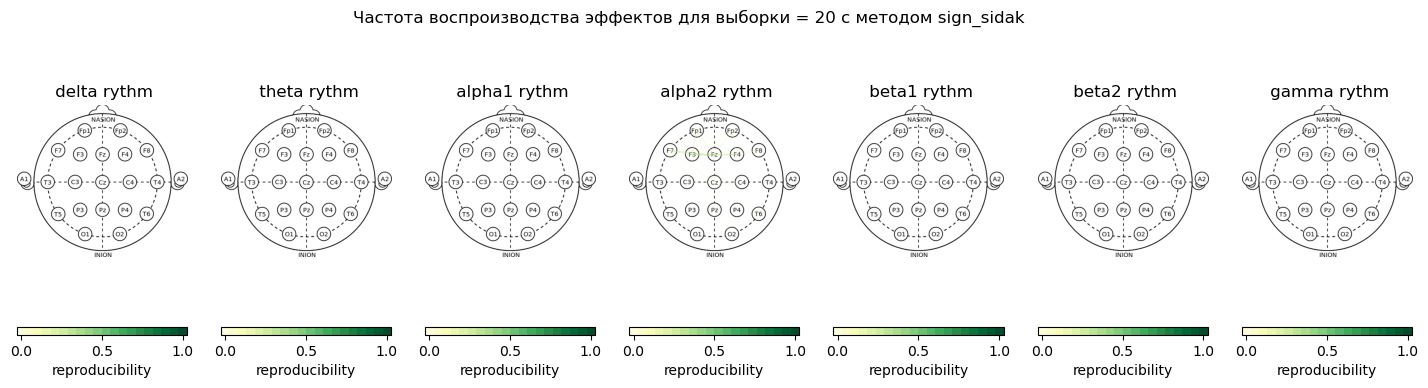
Второй эксперимент заключался в том, что все найденные эффекты во всех подгруппах разного размера сравнивались с эталоном. В качестве эталона были выбраны те эффекты, которые найдены в полной выборке с использованием коррекции с границей размера эффекта равной 0.1. Для каждой подгруппы вычислялись метрики dice, power и fdr

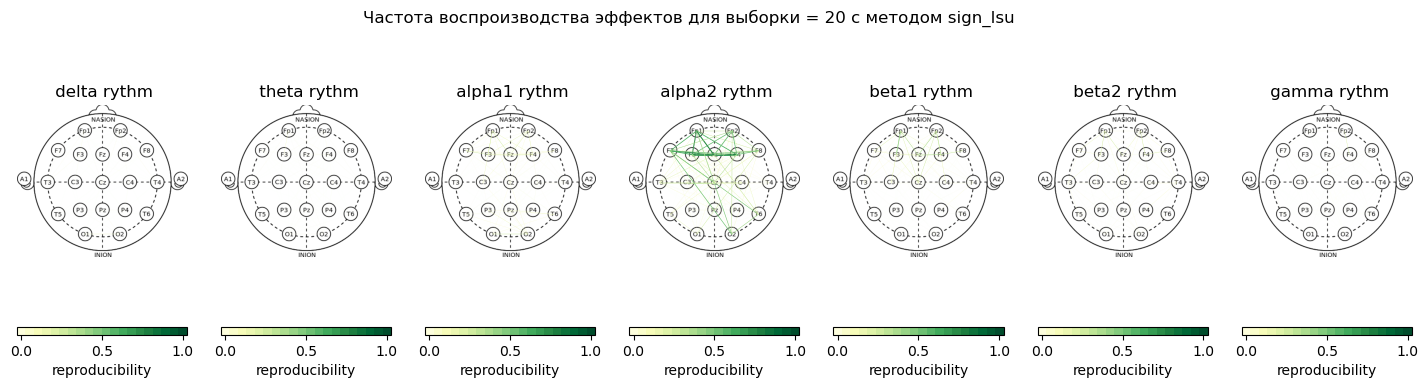
**Результаты**

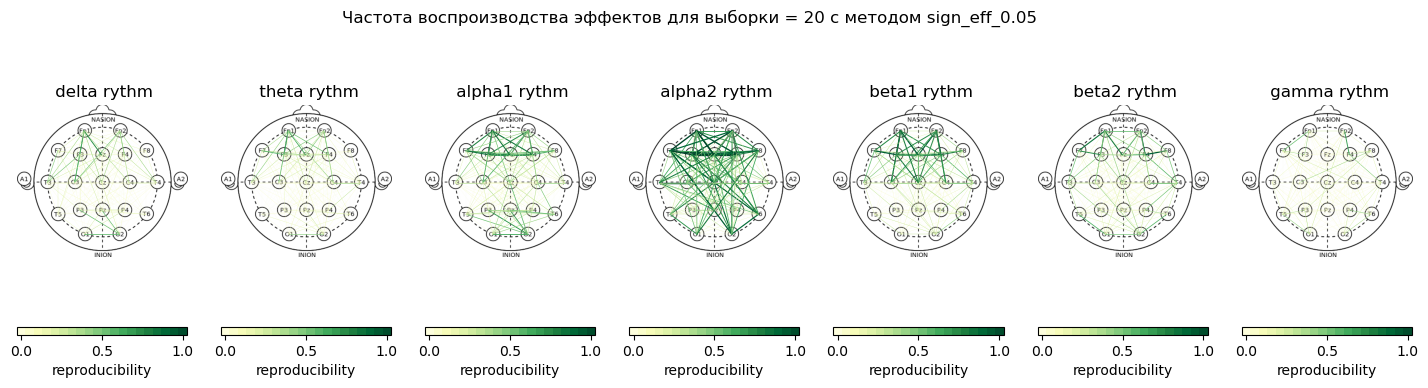
Первый эксперимент: воспроизводимость результатов между непересекающимися подгруппами. Чем выше коэффициент – тем лучше, видно, что на малых выборках коэффициент для FWER, FDR поправок очень низкий. Самые большие для случая, когда коррекция не проводится вообще и для коррекции с размером эффекта.

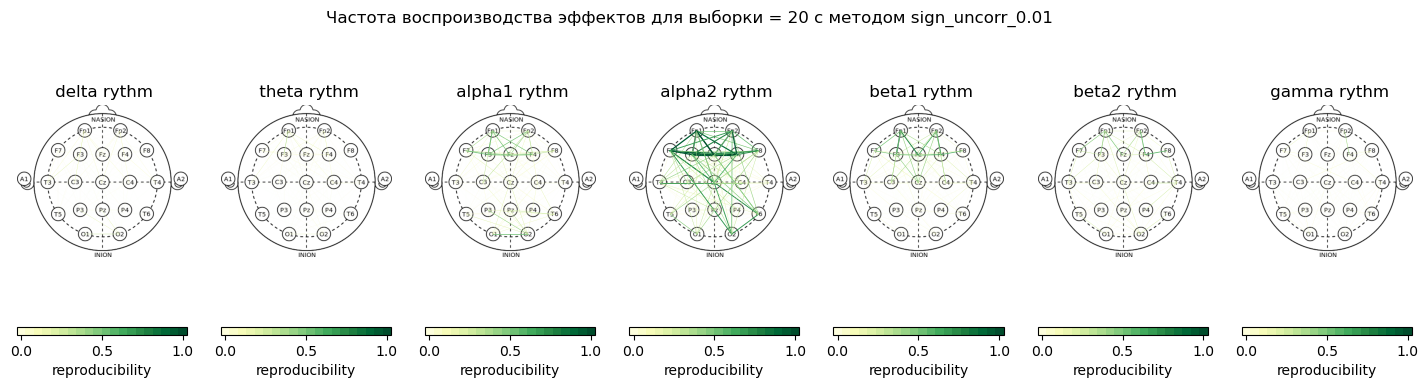
****

С разбивкой по каждому каналу это видно на графиках с частотой воспроизводства. Приведем результаты с самым маленьким размером выборки 20 и с достаточно большим для обычного эксперимента – 45.

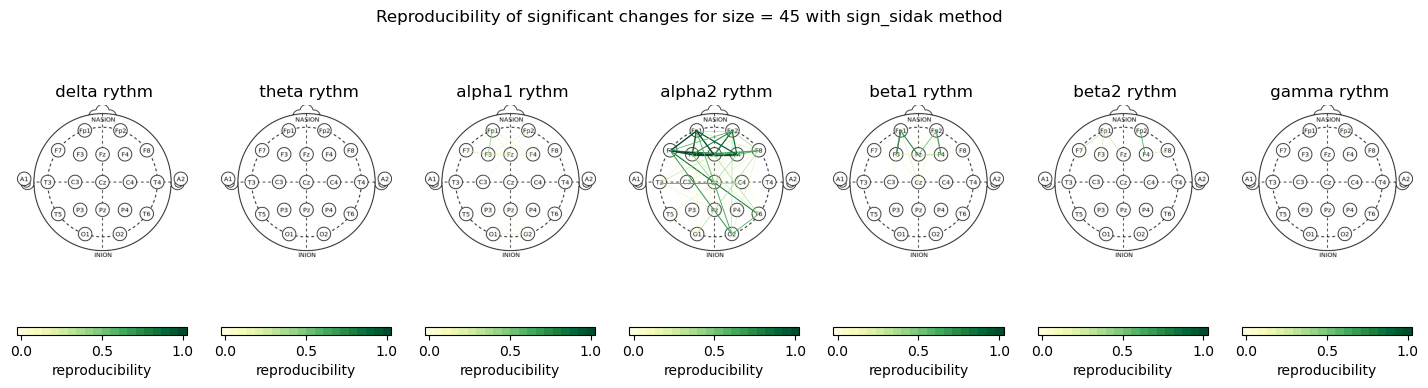


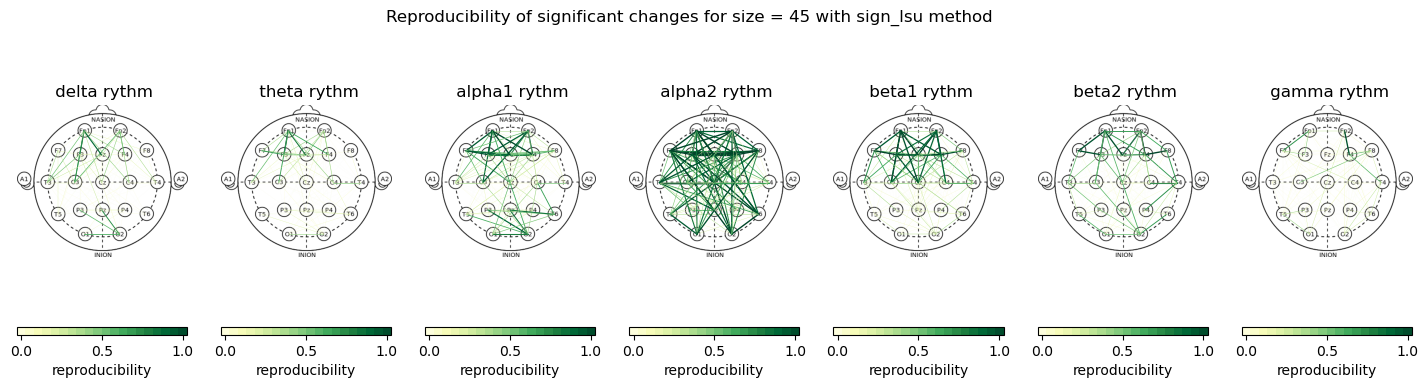


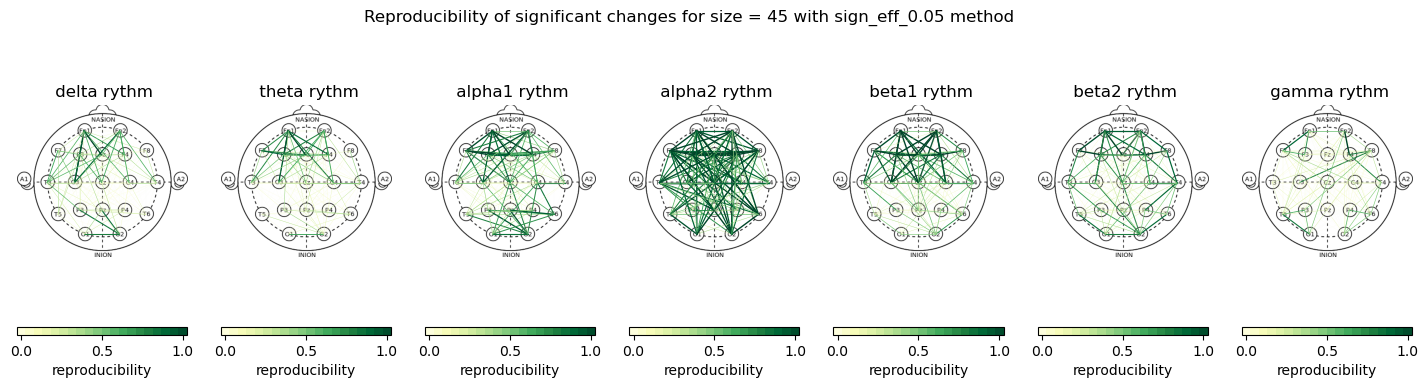


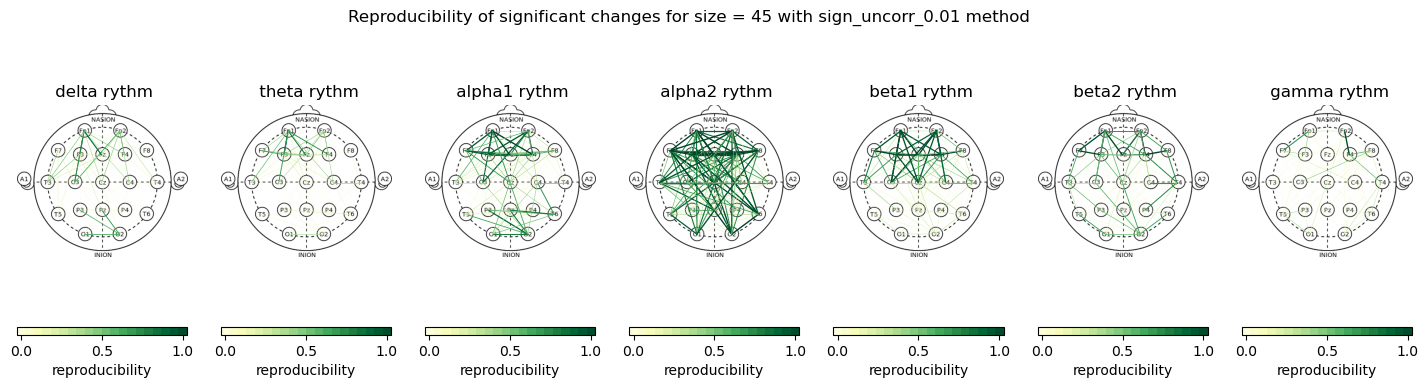


Размер выборки 45



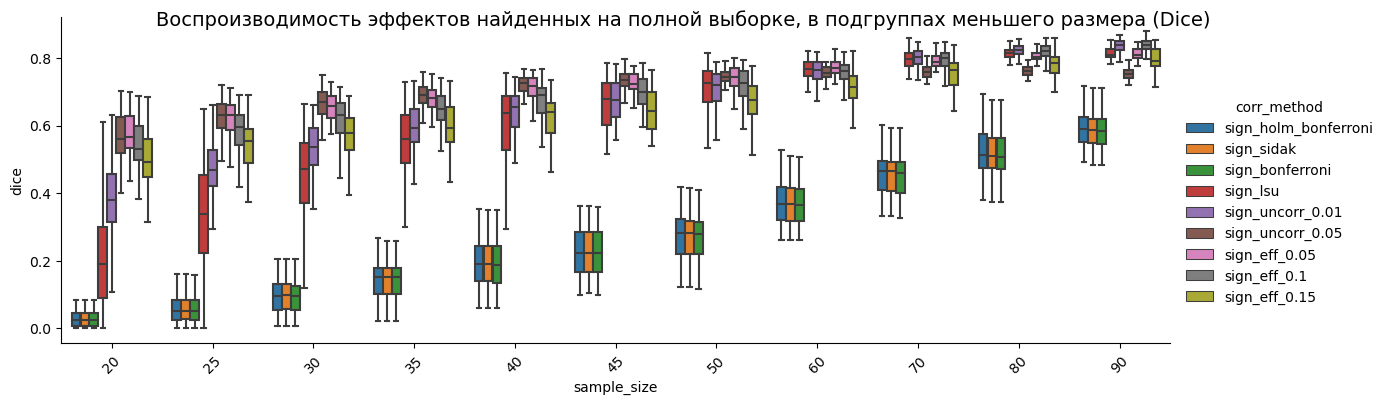


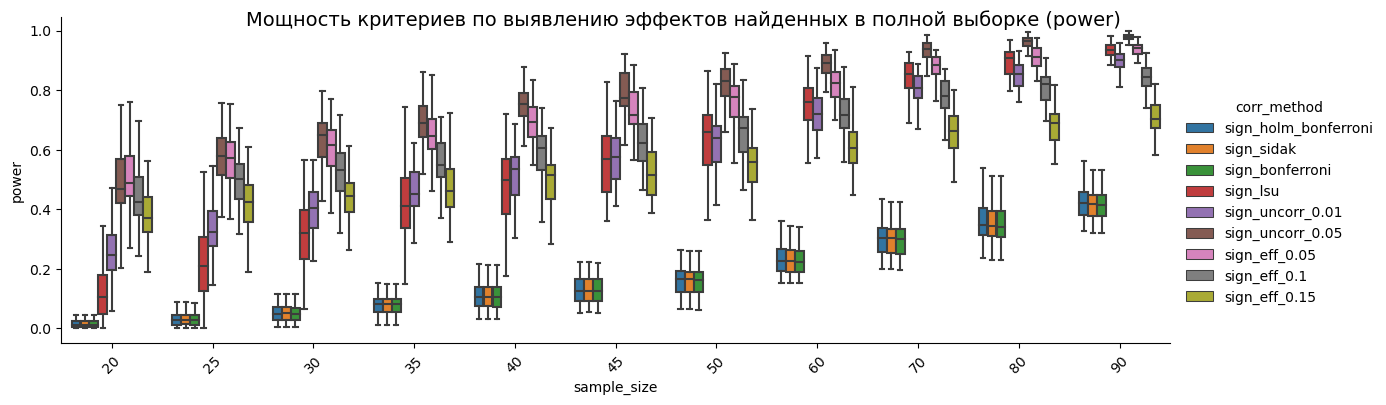


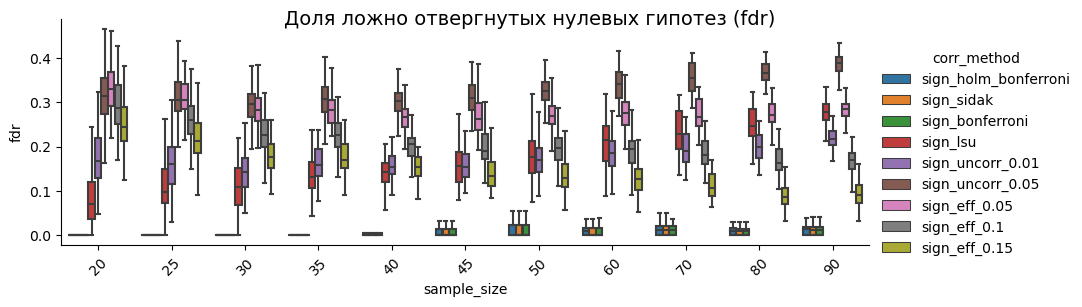


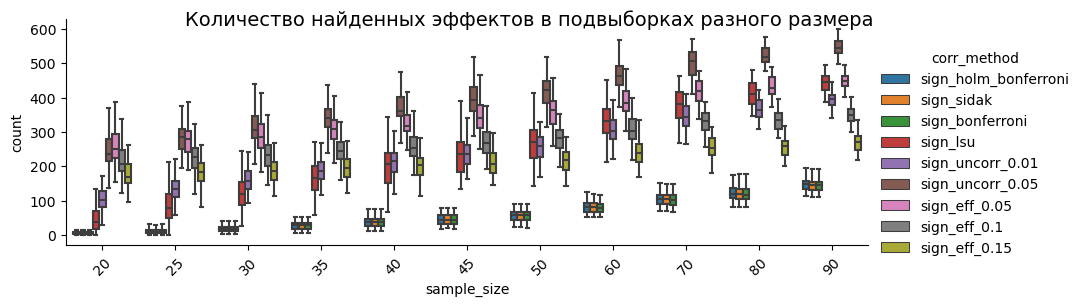
Второй эксперимент:

Поправки FWER обеспечивают очень низкую мощность и с их использованием теряем почти все эффекты. FDR также не эффективен на совсем маленьких выборках, но мощность быстро растет с размером выборки и приближается при этом к неоткорректированным вариантам и критерию размера эффекта. Но при этом растет и число ложно отвергнутых гипотез. Неким компромисом является использование размера эффекта, при этом можно варьировать уровень строгости.

****

****





Что показано:

1. Bспользование бутстрап доверительных интервалов размера эффекта для выявления значимых эффектов показывает лучшие результаты воспроизводимости на всех размерах выборки, как по частоте воспроизводимости отдельных ребер, так и паттерна целиком.
2. Количество выявляемых эффектов на малых выборках сравнимо с числом, получаемым при использовании неоткорректированных значений p\_value, и существенно превышает полученные как с помощью fwer, так и fdr коррекцией.
3. Воспроизводимость результатов (dice коэффициент) оказывается лучше для критерия размера эффекта лучше для всех размеров выборки.
4. Оценка воспроизводимости эффекта различий когерентности между состояниями открытых и закрытых глаз для конкретной пары отведений показала, что даже для малых выборок сильные эффекты (с точки зрения размера эффекта) воспроизводятся достаточно уверенно, если руководствоваться критерием расстояния до нижней границы бутстреп интервала, при этом использование корректировок на множественные сравнения не позволяет обнаружить эффекты. Использование неоткорректированного порога оказывается также менее эффективным
5. Установлены диапазоны и группы пар отведений, в которых изменения когерентности в условиях эффекта Бергера наиболее воспроизводимы – альфа 1-2 и бета-1 диапазоны. Наиболее вариабельны показатели когерентности в гамма диапазоне.

Полученные данные могут быть использованы в качестве опорных для дальнейших исследований по анализу когерентности в состоянии оперативного покоя – где можно ожидать эффект (например, влияния лекарственных средств, индивидуальных характеристик и т.д.), а также для работ методической направленности по развитию методов оценки когерентности (например, оценивать вклад тех или иных методических инноваций

**Пункты дискуссии.**

**Выводы по пунктам дискуссии.**