



СТАТИСТИЧЕСКИЙ АНАЛИЗ МАЛЫХ ВЫБОРОК В МОРФОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ: РЕЗУЛЬТАТЫ СИМУЛЯЦИОННОГО МОДЕЛИРОВАНИЯ И ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ

1, ⁵Гржибовский А.М., ²Гвоздецкий А.Н., ¹Полиданов М.А., ^{1, 4}Хайруллин Р.М.

¹Университет РЕАВИЗ, ²Северо-Западный государственный медицинский университет им. И.И. Мечникова,

³Городская психиатрическая больница № 3 им. И.И. Скворцова-Степанова, ⁴Первый Санкт-Петербургский государственный медицинский университет им. акад. ИП. Павлова, Санкт-Петербург; ⁵Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Россия; ⁶Ошский государственный университет, Ош, Кыргызстан;

e-mail: maksim.polidanoff@yandex.ru

Резюме

ВВЕДЕНИЕ. В морфологических исследованиях объем выборки часто крайне ограничен из-за редкой встречаемости некоторых патологических процессов, этических ограничений и высокой стоимости методов. При этом корректный статистический анализ малых выборок остается мало освещенной проблемой в отечественной литературе, а исследователи нередко применяют неподходящие критерии (тесты), что уменьшает достоверность результатов исследований.

ЦЕЛЬ. Представить результаты симуляционного исследования перечня статистических методов для сравнения независимых групп в условиях малых выборок и сформулировать практические рекомендации по выбору критериев с учетом специфики морфологических данных.

МАТЕРИАЛЫ И МЕТОДЫ. Проведено симуляционное исследование с генерацией 10 000 выборок для каждого сценария (объем от 2 до 12 наблюдений в каждой группе, включая несбалансированные сочетания; результаты представлены по общему числу наблюдений). Оценивались категориальные, порядковые и метрические данные. Критерии оценки тестов: способность удерживать номинальную ошибку первого рода, устойчивость (критерий Брэдли), статистическая мощность и минимальное количество наблюдений для отклонения нулевой гипотезы.

РЕЗУЛЬТАТЫ. Для таблиц 2×2 наилучшие показатели продемонстрировали G-тест с поправкой Уильямса и тест Бошлу (минимальное N = 4–5). Для таблиц 3×2 предпочтительным оказался тест Фишера–Фримена–Гальтона с mid-p коррекцией. Среди критериев для порядковых данных лучшие результаты показал асимптотический тест Бруннера–Мюнцеля (от 5 наблюдений суммарно). Для метрических данных наиболее состоятельными оказались бета-регрессия с коррекцией смещения и критерий Уэлча (от 4–5 наблюдений суммарно).

ЗАКЛЮЧЕНИЕ. При работе со малыми выборками в морфологии предпочтение следует отдавать точным и пермутационным тестам. Не рекомендуется понижать уровень шкалы измерения без веской причины. Предложенные алгоритмы выбора критериев учитывают специфику морфологических данных и могут обеспечивать приемлемый контроль ошибки I рода в рассмотренных сценариях при ограниченном объеме материала.

Ключевые слова

морфометрия, малая выборка, статистическая мощность, статистические критерии, морфологические исследования.

Для цитирования

Гржибовский А.М., Гвоздецкий А.Н., Полиданов М.А., Хайруллин Р.М. Статистический анализ малых выборок в морфологических исследованиях: результаты симуляционного моделирования и практические рекомендации. Морфологические ведомости. 2026;34(3):1065. [https://doi.org/10.20340/mv-mn.2026.34\(3\).1065](https://doi.org/10.20340/mv-mn.2026.34(3).1065)

Статья поступила в редакцию 01 сентября 2026

Статья принята к публикации 30 сентября 2026



THE SMALL SAMPLES STATISTICAL ANALYSIS IN MORPHOLOGICAL STUDIES: RESULTS OF SIMULATION MODELING AND PRACTICAL RECOMMENDATIONS

1, 5Grzhibovskiy AM, 2, 3Gvozdetskiy AN, 1Polidanov MA, 1, 4, 6Khayrullin RM

¹REAVIZ Private University; ²Mechnikov North-West State Medical University; ³Skvortsov-Stepanov City Psychiatric Hospital No. 3; ⁴Academician Pavlov First St. Petersburg State Medical University, St. Petersburg; ⁵Ammosov North-Eastern Federal University, Yakutsk, Russia; ⁶Osh State University, Osh, Kyrgyzstan; e-mail: maksim.polidanoff@yandex.ru

Abstract

INTRODUCTION. In morphological studies, sample size is often extremely limited due to rarity of pathological processes, ethical constraints, and high cost of methods. Correct statistical analysis of small samples remains an understudied problem, and researchers often apply inappropriate tests.

OBJECTIVE. To present the results of a simulation study of statistical methods for comparing independent groups under small sample sizes and to formulate practical recommendations for test selection considering the specifics of morphological data.

MATERIALS AND METHODS. A simulation study was performed with 10,000 replications per scenario (sample size from 2 to 12 per group). Categorical, ordinal, and metric data were evaluated. Test performance criteria: type I error control, robustness (Bradley's criterion), statistical power, and minimum sample size to reject the null hypothesis.

RESULTS. For 2×2 tables, the Williams-corrected G-test and Boschloo's test showed the best performance (minimum

$n = 4 - 5$). For 3×2 tables, the Fisher-Freeman-Halton test with mid-p correction was preferred. Among ordinal tests, the asymptotic Brunner-Munzel test performed best (from 5 observations in total). For metric data, beta regression (with bias reduction) and Welch's test were the most robust (from 4–5 observations in total).

CONCLUSION. For small samples in morphology, exact and permutation tests should be preferred. Reducing the measurement scale without a strong biological reason is strongly discouraged. The proposed algorithms for selecting criteria take into account the specifics of morphological data and can provide acceptable control of type I errors in the considered scenarios with a limited amount of material.

Keywords

morphometry, small samples, statistical power, statistical tests, morphological study.

For the citation

Grzhibovskiy AM, Gvozdetskiy AN, Polidanov MA, Khayrullin RM. The small samples statistical analysis in morphological studies: results of simulation modeling and practical recommendations. *Morfologicheskie Vedomosti – Morphological newsletter*. 2026;34(3):1065. [https://doi.org/10.20340/mv-mn.2026.34\(3\).1065](https://doi.org/10.20340/mv-mn.2026.34(3).1065)

Article received 01 September 2026

Article accepted 30 September 2026

ВВЕДЕНИЕ

Настоящей публикацией мы открываем методологическую рубрику, посвященную статистической обработке данных с учетом особенностей морфологических исследований, к числу которых относятся: редкость изучаемых нозологий и морфологических феноменов; уникальность и неповторимость коллекций препаратов, музейных и ископаемых артефактов, высокая стоимость и трудоемкость современных методов (электронная микроскопия, иммуногистохимия с широкими панелями антител); строгие этические ограничения при работе с людьми и

лабораторными животными, человеческим и животным биоматериалом, требующие минимизации их числа (принцип 3R Рассела – Берча); а также преобладание порядковых (полуколичественных) шкал оценки гистологических препаратов с результатами иммуногистохимии.

На сегодняшний день значительный пласт учебной литературы по анализу биомедицинских данных имеет классическую структуру, в которой основное внимание уделяется тестам, требующим больших выборок. Не вызывает сомнений, что у исследователей есть понимание важности

«достаточной» выборки. Однако на практике ситуация, при которой возможность легко набрать требуемый объем наблюдений отсутствует и для морфологических исследований является скорее правилом, чем исключением. С малым объемом наблюдений связывают недостоверность доказательств и выводов [1], хотя само понятие малой выборки строго не определено. В зависимости от задачи и источника, малый объем наблюдений может равняться 10 [2], 8 [3] и даже 4 (т.е. по два наблюдения в двух группах) [4].

Особое значение проблема малых выборок имеет для медицинской анатомии, гистологии, эмбриологии и патологической анатомии. Накопление даже 10–15 наблюдений, к примеру, по редким формам опухолей или дисплазиям может занимать годы. В публикациях профильных журналов также регулярно встречаются исследования с ограниченным объемом выборки, что подчеркивает актуальность методологической поддержки в этой области [5–6]. Между тем, в доступной литературе практически отсутствуют руководства, которые были бы одновременно строгими с методологической точки зрения и доступными для понимания специалистами-медиками без глубокой математической подготовки.

ЦЕЛЬ ИССЛЕДОВАНИЯ

Проведение серии симуляционных экспериментов для оценки характеристик различных статистических критериев при малых выборках и формулирование практических рекомендаций по их применению в контексте указанной выше специфики морфологических исследований.

МАТЕРИАЛЫ И МЕТОДЫ

Для оценки характеристик статистических тестов было проведено симуляционное исследование. Для каждого сценария генерировалось 10 000 выборок. Перебирались все сочетания объемов групп от 2 до 12 наблюдений в каждой (2 и 2, 2 и 3, 3 и 3 и т.д.), включая несбалансированные; при сравнении трех групп – аналогично для всех троек. Результаты представлены в зависимости

от общего числа наблюдений во всех сравниваемых группах. Рассматривались три типа данных: категориальные (таблицы сопряженности), порядковые (ранговые шкалы) и метрические (непрерывные ряды количественных данных).

Оценка тестов проводилась по четырем строгим критериям:

1) способность удерживать вероятность ошибки первого рода, оценивалась на данных с одинаковыми параметрами между группами; ожидаемая частота ложноположительных результатов – 5% ($\alpha = 0,05$);

2) соответствие эмпирического уровня ошибки I рода либеральному критерию Брэдли (Bradley) [7]: эмпирический уровень значимости должен был укладываться в диапазон 0,025–0,075 ($\alpha \pm 0,5\alpha$);

3) статистическая мощность; оценивалась на распределениях с разными параметрами для сравниваемых групп; чем выше доля корректно отклоненных нулевых гипотез – тем выше мощность;

4) минимальное количество наблюдений, определялось как наименьший объем выборки, при котором тест хотя бы однократно был способен отклонить нулевую гипотезу.

По каждому из четырех критериев тесты ранжировались внутри своей группы сравнения (тип шкалы и число сравниваемых групп) методом плотного ранжирования, наилучшему значению присваивался максимальный ранг, наихудшему – ранг 1. Так, при девяти тестах в группе лучший результат получал 9 баллов. Соппадающим значениям присваивался одинаковый ранг, а максимально возможный ранг в группе, соответственно, уменьшался. Итоговая рейтинговая оценка теста рассчитывалась как сумма его рангов по всем четырем критериям и использовалась для определения лидеров в каждом сценарии сравнения.

Для генерации данных использовались специализированные алгоритмы, реализованные в среде статистического программирования R (версия 4.3.x). Для таблиц сопряженности применялась библиотека rTableICC [8]; для порядковых

данных – пакет LikertMakeR. Пакет LikertMakeR (автор Н. Winzar) доступен в репозитории CRAN: <https://CRAN.R-project.org/package=LikertMakeR>.

Для метрических данных генерировались значения из бета-распределения с заданными параметрами с помощью базовой функции `rbeta`. Параметры генерации данных в каждом сценарии приведены в табл. 1.

Таблица 1. Параметры генерации распределений
Table 1. Parameters of the generated distributions

	Две группы		Три группы	
	Номинальная ошибка	Мощность	Номинальная ошибка	Мощность
Таблица сопряженности	1/4 1/4 1/4 1/4	0,15 0,35 0,35 0,15	1/6 1/6 1/6 1/6 1/6 1/6	0,35 0,05 0,10 0,10 0,05 0,35
Порядковые данные	Min - 1 Max - 5 $\mu_1 = 3$ $\mu_2 = 3$ $\sigma = 1$	Min - 1 Max - 5 $\mu_1 = 2,5$ $\mu_2 = 3,5$ $\sigma = 1$	Min - 1 Max - 5 $\mu_1 = 3$ $\mu_2 = 3$ $\mu_3 = 3$ $\sigma = 1$	Min - 1 Max - 5 $\mu_1 = 2,5$ $\mu_2 = 3,0$ $\mu_3 = 3,5$ $\sigma = 1$
Метрические данные	$\alpha_1 = 10$, $\beta_1 = 10$ $\alpha_2 = 10$, $\beta_2 = 10$	$\alpha_1 = 8$, $\beta_1 = 12$ $\alpha_2 = 10$, $\beta_2 = 10$	$\alpha_1 = 10$, $\beta_1 = 10$ $\alpha_2 = 10$, $\beta_2 = 10$ $\alpha_3 = 10$, $\beta_3 = 10$	$\alpha_1 = 8$, $\beta_1 = 10$ $\alpha_2 = 10$, $\beta_2 = 10$ $\alpha_3 = 12$, $\beta_3 = 10$

В анализ были включены следующие тесты и их программные реализации (все расчеты выполнены с использованием пакетов `stats`, `exact2x2`, `coin` [9], `lawstat`, `brunner.munzel`, `betareg` и др.). Для таблиц 2×2: точный тест Фишера (варианты `minlike`, `central`, `midp`), тесты Барнарда и Бошлу, χ^2 Пирсона (с поправкой Йетса и без), G-тест (с поправкой Уильямса и без).

Для таблиц 3×2: χ^2 , G-тест, тест Фишера–Фримена–Гальтона (все в вариантах `exact`, `midp`, `asymptotic`). Для порядковых данных: Вилкоксона–Манна–Уитни, Флигнера–Полиселло, Бруннера–Мюнцеля, Брауна–Муда, Краскела–Уоллиса, Ван дер Вардена, Сэведжа, Джонкхира–Терпстры. Для метрических данных: тест Стьюдента, тест Уэлча, дисперсионный анализ, тест Фишера–Питмана, обобщенные линейные модели

(Пуассона), бета-биномиальный тест [10], бета-регрессия [11] с коррекцией смещения (`bias-reduced`, BR).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результаты симуляции показали, что почти все классические тесты демонстрируют критически низкую мощность при малом количестве наблюдений, а для многих точных критериев существует жесткий лимит по минимальному объему выборки, ниже которого они математически не работают.

Категориальные данные. Для таблиц сопряженности 2×2 минимальное количество наблюдений ($N = 4$) зафиксировано у G-теста с поправкой Уильямса, G-теста и χ^2 . По совокупности критериев устойчивости и контроля ошибки первого рода (в рамках критерия Брэдли [7], т.е. эмпирический уровень значимости в пределах 0,025–0,075) наилучшие результаты показали G-тест с поправкой Уильямса и тест Бошлу (минимальное $N = 4$ –5). Классический точный тест Фишера и χ^2 с поправкой Йетса оказались чрезмерно консервативными и маломощными. Для таблиц 3×2 наилучшим выбором стал тест Фишера–Фримена–Гальтона с `mid-p` коррекцией, который превзошел асимптотические аналоги по мощности при адекватном контроле ошибки первого рода. Результат симуляции приведен на рис. 1.

Порядковые данные. При сравнении двух групп лидером стал асимптотический тест Бруннера–Мюнцеля [12–13], следом за ним идет критерий Вилкоксона–Манна–Уитни с коррекцией. Тест Бруннера–Мюнцеля формально применим уже при 5-ти наблюдениях суммарно (например, 2 и 3) и, что критически важно, не требует допущения о равенстве дисперсий. Точные тесты для порядковых данных требуют не менее 7-ми наблюдений суммарно. При сравнении трех и более групп лучшие показатели продемонстрировали тесты Ван дер Вардена и Сэведжа, а также не асимптотический вариант теста Краскела–Уоллиса. Результат симуляции приведен на рис. 2.

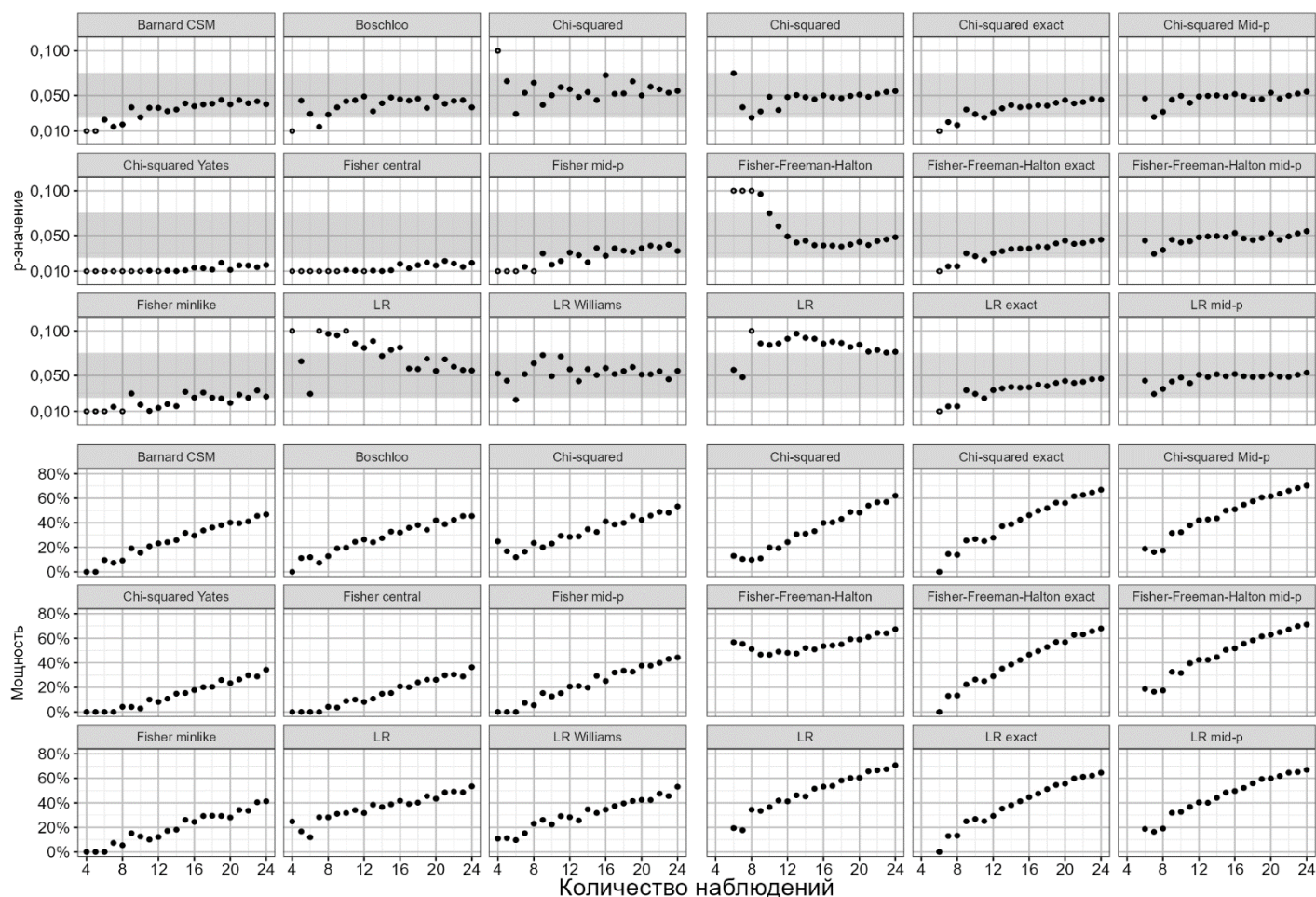


Рис. 1. Динамика p-значений и мощности тестов для таблиц сопряженности. По оси абсцисс – общее число наблюдений во всех сравниваемых группах

Fig. 1. Changes in p-values and test power for contingency tables. The x-axis shows the total number of observations across the compared groups

Метрические данные. Максимальную результативность и способность удерживать номинальную ошибку при малых выборках (от 4–5 наблюдений суммарно) показала бета-регрессия [11] (в варианте с коррекцией смещения, bias-reduced). Тест Стьюдента и его расширение закономерно оказался на втором месте за счет мощности и удержания номинальной ошибки на должном уровне при минимальном количестве наблюдений. Следует подчеркнуть, что параметры бета-распределения, использованные при генерации данных (см. табл. 1), задавали близкие дисперсии сравниваемых групп, поэтому преимущество поправки Уэлча в наших сценариях проявиться не могло. Результат симуляции приведен на рис. 3.

В табл. 2 приведены результаты по сформулированным критериям сравнения и общая рейтинговая оценка внутри каждого типа теста.

Проведенное симуляционное моделирование подтвердило гипотезу о том, что классические асимптотические критерии (χ^2 Пирсона, t-тест Стьюдента, стандартный тест Манна–Уитни) теряют свою мощность при объеме выборок менее 10 наблюдений в группе. Наиболее драматично это проявляется в контроле ошибки первого рода: реальная частота ложноположительных выводов в условиях малых выборок может в 1,5–2 раза превышать номинальный уровень 5%, что напрямую ведет к публикации случайных артефактов.

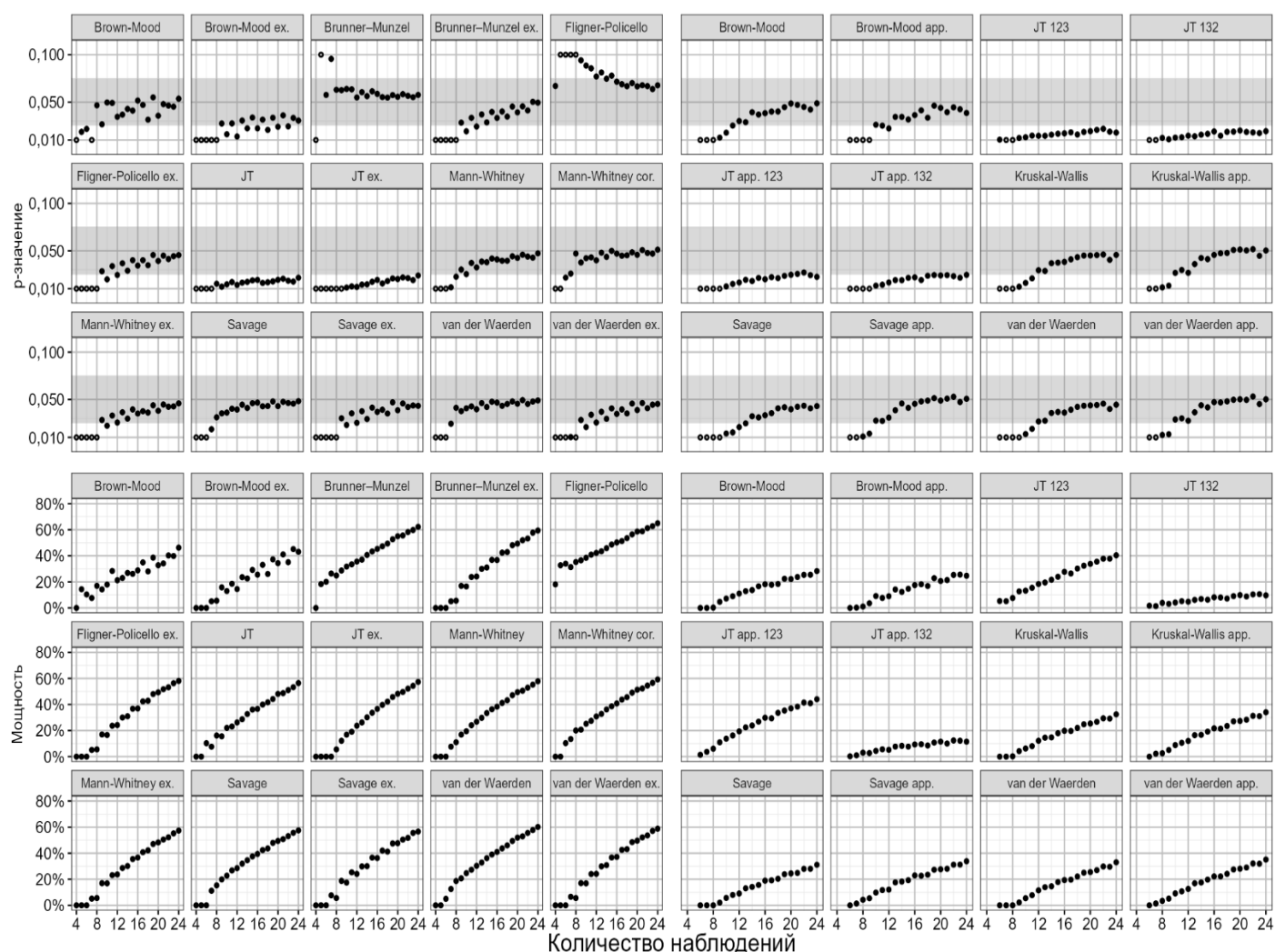


Рис. 2. Динамика р-значений и мощности тестов для порядковых данных. По оси абсцисс – общее число наблюдений во всех сравниваемых группах

Fig. 2. Changes in p-values and test power for ordinal data. The x-axis shows the total number of observations across the compared groups

Полученные данные согласуются с классическими работами по робастности [7], однако в настоящей работе представлены количественные пороговые значения минимального N для каждого класса тестов именно применительно к трем типам морфологических шкал.

Биологическая и методологическая интерпретация для таблиц сопряженности. Превосходство G-теста с поправкой Уильямса над точным тестом Фишера для таблиц 2×2 объясняется природой этих критериев: точный тест Фишера фиксирует оба маргинальных итога (суммы по строкам

и столбцам), что при малых частотах создает избыточно жесткое дискретное распределение и приводит к консервативности (р-значения систематически завышены). Напротив, G-тест и тест Бошлу используют более гибкие подходы (аппроксимация отношения правдоподобия и безусловный точный тест соответственно), что позволяет «улавливать» реальные различия во встречаемости гистологических феноменов даже при N = 4. В патологоанатомической практике это означает, что сравнение групп крайне редко встречающихся опухолевых вариантов (3–5 наб-

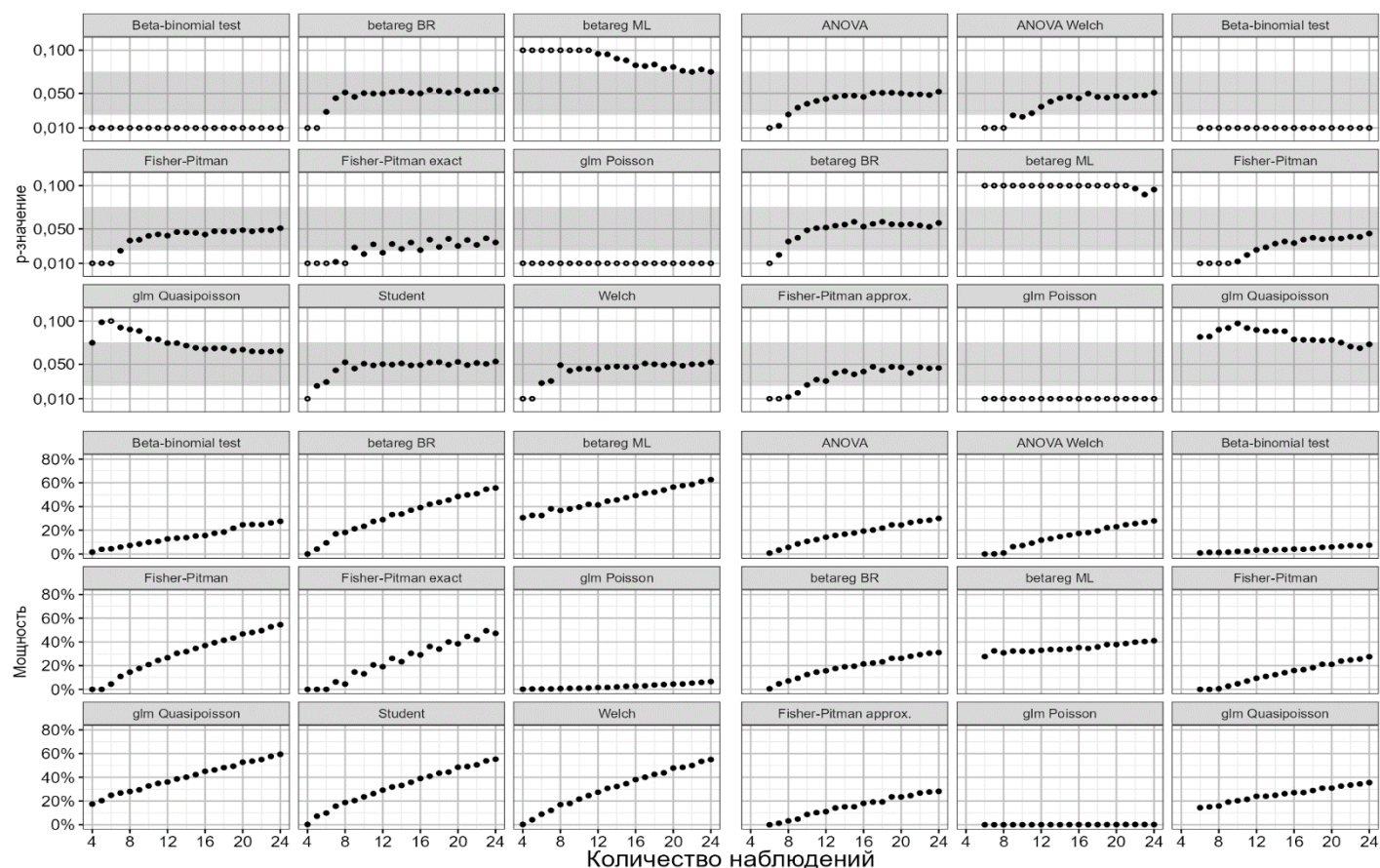


Рис. 3. Динамика р-значений и мощности тестов для метрических данных. По оси абсцисс – общее число наблюдений во всех сравниваемых группах

Fig. 3. Changes in p-values and test power for metric data. The x-axis shows the total number of observations across the compared groups

людений) все же может дать статистически обоснованный результат при правильном выборе теста, что ранее считалось практически невозможным.

Природа преимуществ теста Бруннера – Мюнцеля для порядковых шкал. В морфологии полуколичественные шкалы (например, IRS-баллы в иммуногистохимии, степени фиброза) обладают двумя принципиальными особенностями: отсутствием равных интервалов и, что еще более важно, закономерным увеличением дисперсии при возрастании выраженности патологического признака. Гетероскедастичность в таких данных – не статистическая помеха, а биологический феномен (патологический процесс делает ткань более вариабельной). Классический тест Манна–Уитни не учитывает различий в дисперсиях и

при малых N показывает неприемлемо высокий уровень ложных отклонений нулевой гипотезы. Тест Бруннера – Мюнцеля, решая непараметрическую проблему Беренса – Фишера, разделяет эффекты сдвига центра распределения и изменения дисперсии, что делает его предпочтительным тестом среди рассмотренных при сравнении контрольной и патологической групп в иммуногистохимических исследованиях [12–13].

Обоснование выбора бета-регрессии для метрических данных. Многие морфометрические показатели являются долями или процентами (например, площадь ядра относительно цитоплазмы, процент клеток с позитивной реакцией). Эти данные ограничены интервалом (0, 1) и часто имеют бимодальное или скошенное распределение. Традиционная арксинус-трансформация,

Таблица 2. Результат симуляционного исследования
Table 2. Results of the simulation study

Тип теста	Статистический тест	Среднее отклонение от α	Средняя мощность	Устойчивость по Брэдли	Минимальная выборка	Рейтинг (балл)
2x2	Fisher minlike	0,031 (3)	0,198 (3)	9 (2)	7 (2)	10
2x2	Fisher central	0,039 (2)	0,152 (2)	0 (1)	8 (1)	6
2x2	Fisher mid-p	0,026 (4)	0,225 (4)	13 (4)	7 (2)	14
2x2	Barnard CSM	0,018 (6)	0,256 (5)	16 (5)	6 (3)	19
2x2	Boschloo	0,012 (7)	0,272 (6)	19 (6)	5 (4)	23
2x2	Chi-squared	0,007 (8)	0,33 (8)	20 (7)	4 (5)	28
2x2	Chi-squared Yates	0,041 (1)	0,142 (1)	0 (1)	8 (1)	4
2x2	LR	0,025 (5)	0,363 (9)	11 (3)	4 (5)	22
2x2	LR Williams	0,003 (9)	0,311 (7)	20 (7)	4 (5)	28
3x2	Fisher-Freeman-Halton exact	0,018 (3)	0,404 (3)	15 (2)	7 (1)	9
3x2	Fisher-Freeman-Halton mid-p	0,004 (6)	0,474 (7)	19 (4)	6 (2)	19
3x2	Fisher-Freeman-Halton	0,021 (2)	0,549 (9)	15 (2)	6 (2)	15
3x2	Chi-squared exact	0,016 (5)	0,404 (4)	16 (3)	7 (1)	13
3x2	Chi-squared Mid-p	0,003 (7)	0,466 (6)	19 (4)	6 (2)	19
3x2	Chi-squared	0,003 (9)	0,344 (1)	19 (4)	6 (2)	16
3x2	LR exact	0,016 (4)	0,394 (2)	15 (2)	7 (1)	9
3x2	LR mid-p	0,003 (8)	0,455 (5)	19 (4)	6 (2)	19
3x2	LR	0,033 (1)	0,486 (8)	2 (1)	6 (2)	12
OrdTS	Mann-Whitney	0,018 (9)	0,305 (10)	16 (6)	7 (2)	27
OrdTS	Mann-Whitney correct	0,011 (13)	0,338 (13)	18 (8)	6 (3)	37
OrdTS	Mann-Whitney exact	0,022 (5)	0,293 (4)	15 (5)	7 (2)	16
OrdTS	Jonckheere Terpstra	0,035 (2)	0,304 (9)	0 (1)	6 (3)	15
OrdTS	Jonckheere Terpstra exact	0,037 (1)	0,282 (3)	0 (1)	8 (1)	6
OrdTS	Fligner-Policello	0,034 (3)	0,46 (15)	11 (3)	4 (5)	26
OrdTS	Fligner-Policello exact	0,022 (6)	0,299 (6)	14 (4)	7 (2)	18
OrdTS	van der Waerden exact	0,021 (7)	0,302 (8)	15 (5)	7 (2)	22
OrdTS	van der Waerden	0,012 (12)	0,336 (12)	17 (7)	6 (3)	34
OrdTS	Savage exact	0,021 (8)	0,297 (5)	15 (5)	7 (2)	20
OrdTS	Savage	0,015 (10)	0,318 (11)	17 (7)	7 (2)	30
OrdTS	Brown-Mood	0,013 (11)	0,253 (2)	17 (7)	5 (4)	24
OrdTS	Brown-Mood exact	0,029 (4)	0,223 (1)	9 (2)	7 (2)	9
OrdTS	Brunner–Munzel	0,01 (14)	0,393 (14)	18 (8)	5 (4)	40
OrdTS	Brunner–Munzel exact	0,021 (7)	0,301 (7)	14 (4)	7 (2)	20
OrdKS	Kruskal-Wallis	0,02 (8)	0,162 (7)	13 (4)	8 (2)	21
OrdKS	Kruskal-Wallis approx.	0,015 (10)	0,179 (8)	15 (6)	7 (3)	27
OrdKS	Jonckheere-Terpstra 123	0,034 (2)	0,234 (11)	0 (1)	6 (4)	18
OrdKS	Jonckheere-Terpstra 132	0,035 (1)	0,066 (1)	0 (1)	6 (4)	7
OrdKS	JT approx. 123	0,032 (4)	0,25 (12)	2 (2)	6 (4)	22
OrdKS	JT approx. 132	0,033 (3)	0,075 (2)	0 (1)	6 (4)	10
OrdKS	van der Waerden approx.	0,014 (11)	0,184 (10)	15 (6)	7 (3)	30
OrdKS	van der Waerden	0,021 (6)	0,16 (6)	13 (4)	9 (1)	17

Тип теста	Статистический тест	Среднее отклонение от α	Средняя мощность	Устойчивость по Брэдли	Минимальная выборка	Рейтинг (балл)
OrdKS	Savage approx.	0,014 (12)	0,183 (9)	15 (6)	7 (3)	30
OrdKS	Savage	0,024 (5)	0,151 (5)	11 (3)	9 (1)	14
OrdKS	Brown-Mood approx.	0,021 (7)	0,14 (3)	14 (5)	7 (3)	18
OrdKS	Brown-Mood	0,019 (9)	0,146 (4)	14 (5)	8 (2)	20
ContTS	Student	0,005 (8)	0,322 (6)	19 (5)	4 (4)	23
ContTS	Welch	0,008 (7)	0,31 (5)	19 (5)	4 (4)	21
ContTS	Fisher-Pitman exact	0,025 (5)	0,247 (3)	14 (3)	7 (1)	12
ContTS	Fisher-Pitman	0,012 (6)	0,299 (4)	17 (4)	6 (2)	16
ContTS	glm Poisson	0,05 (2)	0,025 (1)	0 (1)	4 (4)	8
ContTS	glm Quasipoisson	0,026 (4)	0,399 (8)	14 (3)	4 (4)	19
ContTS	Beta-binomial test	0,044 (3)	0,147 (2)	0 (1)	4 (4)	10
ContTS	betareg ML	0,057 (1)	0,463 (9)	2 (2)	4 (4)	16
ContTS	betareg BR	0,005 (9)	0,325 (7)	19 (5)	5 (3)	24
ContKS	ANOVA	0,009 (8)	0,173 (6)	17 (6)	6 (3)	23
ContKS	ANOVA Welch	0,015 (7)	0,149 (4)	14 (4)	7 (2)	17
ContKS	Fisher-Pitman approx.	0,016 (6)	0,155 (5)	15 (5)	7 (2)	18
ContKS	Fisher-Pitman	0,023 (5)	0,135 (3)	13 (3)	8 (1)	12
ContKS	glm Poisson	0,05 (2)	0,001 (1)	0 (1)	6 (3)	7
ContKS	glm Quasipoisson	0,032 (4)	0,256 (8)	4 (2)	6 (3)	17
ContKS	Beta-binomial test	0,048 (3)	0,04 (2)	0 (1)	6 (3)	9
ContKS	betareg ML	0,086 (1)	0,349 (9)	0 (1)	6 (3)	14
ContKS	betareg BR	0,002 (9)	0,189 (7)	17 (6)	6 (3)	25

Примечание: α – вероятность ошибки первого рода 0,05, критерий Брэдли – $\alpha \in [0,025; 0,075]$; 2x2 – таблица сопряженности 2 на 2; 3x2 – таблица сопряженности 3 на 2; OrdTS – две выборки с порядковыми данными; OrdKS – три выборки с порядковыми данными; ContTS – две выборки с метрическими данными; ContKS – три выборки с метрическими данными; для каждого критерия в скобках указан рейтинг среди группы тестов (чем больше, тем лучше), итоговый балл указывает на общий рейтинг теста среди группы тестов. Минимальная выборка – наименьшее общее число наблюдений во всех сравниваемых группах, при котором тест способен отклонить нулевую гипотезу; значение 0 означает, что тест не отклонил нулевую гипотезу ни при одном из рассмотренных объемов. Устойчивость по Брэдли – число значений общего объема выборки, при которых эмпирический уровень ошибки I рода остался в пределах 0,025–0,075

Note: α – is the probability of a type I error of 0.05, Bradley's criterion is – $\alpha \in [0.025; 0.075]$; 2x2 – is a 2-by-2 contingency table; 3x2 – is a 3-by-2 contingency table; OrdTS – are two samples with ordinal data; OrdKS – are three samples with ordinal data; ContTS – are two samples with metric data; ContKS – are three samples with metric data; for each criterion, the rating among the group of tests is given in parentheses (the higher the better), the final score indicates the overall rating of the test among the group of tests. The minimum sample is the smallest total number of observations in all compared groups for which the test is capable of rejecting the null hypothesis; a value of 0 means that the test did not reject the null hypothesis at any of the sizes considered. Bradley robustness is the number of values of the total sample size for which the empirical level of type I error remains within the range of 0.025–0.075

применяемая для «нормализации» таких данных, крайне нестабильна при малых выборках и искажает реальные различия между группами. Бета-регрессия [11] моделирует данные на их исходной шкале с помощью гибкого параметрического подхода (два параметра – среднее и параметр точности распределения), что позволяет достичь максимальной мощности даже при $N = 3 - 4$. Для

сравнения: тест Уэлча, хотя и является робастным к неравенству дисперсий, все же опирается на предположение о непрерывности и симметричности данных, поэтому в случае строгих долей (0–100%) уступает бета-регрессии.

Прежде чем перейти к практическим рекомендациям, необходимо четко обозначить гра-

ницы применимости наших результатов. Во-первых, симуляционное моделирование всегда опирается на заданные параметры распределений (в нашем случае – бета-распределение для метрических данных и определенные маргинальные частоты для таблиц сопряженности). В реальных морфологических данных могут встречаться более сложные структуры: выбросы, многомодальность, выраженная асимметрия, а также корреляции между наблюдениями (например, при использовании нескольких срезов от одного пациента), которые не были полностью воспроизведены в наших сценариях. Кроме того, метрические данные генерировались исключительно из бета-распределения, то есть были ограничены интервалом (0, 1); поведение тестов на неограниченных непрерывных величинах (линейных размерах, площадях) в наших сценариях не проверялось. Также в метрических данных мы не использовали генерацию счетных данных, которая является естественным выбором для регрессии (квази) Пуассона и другие семейства распределений, которые могут подходить лучше к моделированию конкретного явления.

Во-вторых, мы не рассматривали смешанные модели и методы множественных сравнений. При сравнении трех и более групп отклонение нулевой гипотезы омнибус-тестом не указывает, какие именно группы различаются, а процедуры апостериорных сравнений при малых выборках теряют мощность быстрее самих омнибус-тестов; выбор такой процедуры требует отдельного исследования.

В-третьих, наши выводы справедливы для независимых выборок; в морфологии иногда встречаются парные сравнения (до и после воздействия), которые имеют свои собственные критерии (например, знаковый тест Вилкоксона), не анализировавшиеся в данной работе.

В-четвертых, минимальное количество наблюдений, которое мы определили как порог для возможности отклонения нулевой гипотезы, не следует путать с достаточным объемом для обеспечения высокой мощности (> 80%). Для до-

стижения последнего могут потребоваться большие выборки. Кроме того, приведенные оценки мощности и эмпирического уровня ошибки I рода усреднены по всем сочетаниям объемов групп, включая существенно несбалансированные (например, 2 и 10), поведение тестов при которых может заметно отличаться от сбалансированного случая.

В-пятых, в каждой группе тестов, строго говоря, проверяются неодинаковые статистические гипотезы. Объединение тестов выполнено по шкале измерения, чтобы дать краткий обзор по возможностям статистического инструментария. Для решения реальной задачи необходимо выбирать тест, который проверяет необходимую гипотезу. Таким образом, полученные результаты следует рассматривать как общие ориентиры для эксплораторных исследований, а не как универсальные правила для конфирматорных испытаний.

На основании результатов симуляционного моделирования и литературных данных можно сформулировать специфические рекомендации для исследователей, сталкивающихся с проблемой малых выборок.

1. Анализ полуколичественных шкал (иммуногистохимия, гистологический скоринг).

Часто встречающейся ошибкой анализа данных, полученных по полуколичественным шкалам (например, оценка интенсивности окрашивания от 0 до 3+ или степени фиброза по 4-балльной шкале), является использование критерия Стьюдента. Порядковые данные не имеют равных интервалов между категориями (разница между 1+ и 2+ не обязательно равна разнице между 2+ и 3+) [14], поэтому вычисление среднего арифметического и применение параметрических тестов методологически некорректно и ведет к ложноположительным результатам [15–16].

Рекомендация. Используйте тест Бруннера–Мюнцеля [12–13]. Наши симуляции показали, что он обладает максимальной мощностью при малых выборках и, в отличие от классического теста Манна–Уитни, не требует допущения о равенстве дисперсий. Это критически важно при сравнении

группы контроля и группы с выраженной патологией, где разброс значений в патологической группе зачастую многократно выше.

2. Анализ частот редких морфологических феноменов.

При изучении архивного материала патологоанатомы часто сравнивают частоту встречаемости редких гистологических вариантов опухолей или аномалий развития. Объемы групп могут составлять 3–5 наблюдений. Применение χ^2 Пирсона здесь недопустимо из-за нарушения требований к ожидаемым частотам, а классический точный тест Фишера часто оказывается слишком консервативным, снижая вероятность выявления заданного в симуляции эффекта из-за своей математической природы [17].

Рекомендация: Применяйте G-тест с поправкой Уильямса или критерий Бошлу. Они в рассмотренных сценариях обеспечивали приемлемый контроль ошибки I рода и обладают большей мощностью при $N < 10$, позволяя не упустить статистически значимые связи между редкими морфологическими признаками.

3. Морфометрия и непрерывные данные.

При анализе линейных размеров, площадей срезов или количества клеток объем выборки часто ограничен 5–10 объектами. Тесты нормальности (например, тест Шапиро – Уилка) при таких объемах выборки имеют низкую мощность и не должны быть единственным основанием для выбора метода.

Рекомендация: если данные заведомо непрерывные (например, точная площадь ядра, измеренная с помощью систем цифровой морфометрии), разумным выбором по умолчанию является тест Уэлча. Подчеркнем, что это рекомендация не по результатам симуляции: в наших сценариях дисперсии групп были близки по построению, и преимущество поправки Уэлча проверено не было. Основание для нее иное – в реальных морфологических данных равенство дисперсий сравниваемых групп соблюдается редко, а цена ошибочного допущения о равенстве при малых выборках велика. Для данных, ограниченных диапазоном (например, процент позитивных

клеток, конвертированный в диапазон от 0 до 1), отличным современным инструментом является бета-регрессия [11]. Она показала более высокую эмпирическую мощность в заданных сценариях при малых выборках и позволяет гибко моделировать данные, не прибегая к грубым трансформациям.

4. Запрет на искусственное занижение шкалы измерения.

Категорически не рекомендуется без веской биологической причины понижать уровень шкалы измерения: превращать метрические данные (точную площадь) в порядковые («маленькое/большое»), а порядковые – в категориальные («есть признак/нет признака»). Это неизбежно ведет к существенной потере статистической информации и мощности, которая и без того критически низка при малых выборках [18].

5. Интерпретация отсутствия значимости.

При работе со малыми выборками ($N < 6$ в группе) отсутствие статистически значимых различий ($p > 0,05$) не должно автоматически трактоваться как доказательство отсутствия эффекта. В подавляющем большинстве случаев это свидетельствует о недостаточной мощности исследования (ошибке второго рода) [1]. В то же время применение точных статистических тестов в таких условиях имеет принципиальное значение, поскольку позволяет контролировать ошибку I рода без использования асимптотических приближений, достоверность которых при малых выборках сомнительна.

В некоторых случаях, однако, природа распределения статистик точных тестов является причиной невозможности достижения уровня значимости $p < 0,05$. Поэтому отсутствие статистической значимости при использовании точного теста в малой выборке следует интерпретировать с учетом как ограниченной мощности, так и дискретности p -значения, а не как свидетельство отсутствия биологического эффекта. Статистический анализ в морфологии малых выборок служит не для окончательного «доказательства» гипотезы, а для оценки вероятности случайности

наблюдаемого феномена и обоснования необходимости дальнейших, более крупных исследований. Дефекты, допущенные на этапе планирования исследования, невозможно исправить никаким объемом наблюдений [19].

Из этого следует не менее важный вывод о необходимости адекватного и достаточного соблюдения принципа 3R Рассела – Берча, о чем отмечалось вначале, с одной стороны, с другой стороны обязательного учета объективных возможностей набора достаточной численности материала и использования адекватных статистических методов и программ для получения обоснованных результатов и выводов. Все указанное в совокупности предъявляет строгие требования к ответственности и компетентности научно-методических советов и проблемных комиссий при планировании исследований для предотвращения нежелательных последствий по исправлению, доработке, пересмотру либо аннулированию их окончательных результатов. Квалифицированная статистическая экспертиза планируемых в биомедицинской морфологии исследований должна

стать органичной составляющей планирования и анализа их результатов.

ЗАКЛЮЧЕНИЕ

Статистический анализ малых выборок требует отказа от шаблонного использования классических асимптотических тестов. Результаты нашего симуляционного моделирования показывают, что для категориальных данных предпочтительны G-тест с поправкой Уильямса и тест Бошлу; для порядковых шкал (включая иммуногистохимический скоринг) – тест Бруннера–Мюнцеля; для метрических данных – бета-регрессия с коррекцией смещения и тест Уэлча. Предложенные подходы позволяют морфологам извлекать максимум достоверной информации из ограниченного объема биологического материала, снижая риск неконтролируемой ошибки I рода, повышая вероятность выявления эффекта в рассмотренных сценариях и укрепляя тем самым общую доказательную базу морфологических исследований.

Литература

References

- 1 Button KS, Ioannidis JPA, Mokrysz C et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci.* 2013;14:5(365–376). <https://doi.org/10.1038/nrn3475>
- 2 Morgan CJ. Use of proper statistical techniques for research studies with small samples. *Am J Physiol Lung Cell Mol Physiol.* 2017;313:5(873–877). <https://doi.org/10.1152/ajplung.00238.2017>
- 3 Jenkins DG, Quintana-Ascencio PF. A solution to minimum sample size for regressions. *PLoS One.* 2020;15:2(e0229345). <https://doi.org/10.1371/journal.pone.0229345>
- 4 Winter JCF de. Using the Student's t-test with extremely small sample sizes. *Pract Assess Res Eval.* 2013;18:10. <https://doi.org/doi:10.7275/E4R6-DJ05>
- 5 Tomchuk ON, Sen'chukova MA, Stadnikov AA i dr. Morfologicheskie osobennosti mikrososudov i limfoidnoy tkani v regionarnykh limfaticeskikh uzlakh pri rake zheludka. *Morfologicheskie vedomosti.* 2015;23(4):25–31. [https://doi.org/10.20340/mv-mn.2015.0\(4\):25-31](https://doi.org/10.20340/mv-mn.2015.0(4):25-31). In Russian
- 6 Govorukhina OA, Prokopenya NS, Kharitonchik VA i dr. Osobennosti morfologicheskoy diagnostiki bolezni Girshprunga. *Morfologicheskie vedomosti.* 2024;32:1(24–35). [https://doi.org/10.20340/mv-mn.2024.32\(1\).818](https://doi.org/10.20340/mv-mn.2024.32(1).818). In Russian
- 7 Bradley JV. Robustness? *Br J Math Stat Psychol.* 1978;31:2(144–152). <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
- 8 Demirhan H. rTableICC: an R package for random generation of 2×2×K and R×C contingency tables. *R J.* 2016;8:1(48–62). <https://doi.org/10.32614/RJ-2016-004>
- 9 Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a class of permutation tests: the coin package. *J Stat Softw.* 2008;28:8(1–23). <https://doi.org/10.18637/jss.v028.i08>
- 10 Pham TV, Piersma SR, Warmoes M, Jimenez CR. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics.* 2010;26:3(363–369). <https://doi.org/10.1093/bioinformatics/btp677>
- 11 Cribari-Neto F, Zeileis A. Beta regression in R. *J Stat Softw.* 2010;34:2(1–24). <https://doi.org/10.18637/jss.v034.i02>

- 12 Karch JD. Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Adv Methods Pract Psychol Sci*. 2021;4:2(2515245921999602) e-pub. <https://doi.org/10.1177/2515245921999602>
- 13 Brunner E, Munzel U. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal*. 2000;42:1(17–25). [https://doi.org/10.1002/\(SICI\)1521-4036\(200001\)42:1<17::AID-BIMJ17>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U)
- 14 Marateb HR, Mansourian M, Adibi P, Farina D. Manipulating measurement scales in medical statistical analysis and data mining: a review of methodologies. *J Res Med Sci*. 2014;19:1(47–56)
- 15 Liddell TM, Kruschke JK. Analyzing ordinal data with metric models: What could possibly go wrong? *J Exp Soc Psychol*. 2018;79(328–348). <https://doi.org/10.1016/j.jesp.2018.08.009>
- 16 Gvozdetskiy AN, Prokopovich GA, Dobrovol'skaya AE i dr. Primenimost' regressionnykh modeley dlya analiza porjadkovykh dannykh PANSS pri shizofrenii: kogortnoe issledovanie. *Consortium Psychiatricum*. 2026;7:2(10–21). <https://doi.org/10.17816/CP15731>
- 17 Fagerland MW, Lydersen S, Laake P. Statistical Analysis of Contingency Tables. 1-st ed. New York: CRC Press, 2017. <https://doi.org/10.1201/9781315374116>
- 18 Grzhibovskiy AM, Unguryanu TN, Savvina NV. Analiz kategorial'nykh dannykh s ispol'zovaniem programmnoho obespecheniya SPSS i STATA. *Narkologiya*. 2018;12:2(39–55). <https://doi.org/10.25557/1682-8313.2018.02.39-55>. In Russian
- 19 Indrayan A, Mishra A. The importance of small samples in medical research. *J Postgrad Med*. 2021;67:4(219–223). <https://doi.org/10.4103/jpgm.JPGM23021>

Авторы заявляют об отсутствии каких-либо конфликтов интересов при планировании, выполнении, финансировании и использовании результатов настоящего исследования

ИНФОРМАЦИЯ ОБ АВТОРАХ

Гржибовский Андрей Мечиславович, доктор медицины, Университет РЕАВИЗ Санкт-Петербург, Россия; Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Россия; ORCID: 0000-0002-5464-0498; e-mail: a.grjibovski@yandex.ru

Гвоздецкий Антон Николаевич, кандидат медицинских наук, Северо-Западный государственный медицинский университет им. И.И. Мечникова, Санкт-Петербург, Россия; ORCID: 0000-0001-8045-1220; e-mail: comisora@yandex.ru

Полиданов Максим Андреевич, Университет РЕАВИЗ, Санкт-Петербург, Россия; ORCID: 0000-0001-7538-7412; e-mail: maksim.polidanoff@yandex.ru

Хайруллин Радик Магзинурович, доктор медицинских наук, профессор, Университет РЕАВИЗ; ORCID: 0000-0002-2796-7508; e-mail: r.m.hayrullin@reaviz.online

The Authors declares that she has no any conflicts of interest in the planning, implementation, financing and use of the results of this study

INFORMATION ABOUT AUTHORS

Andrey M. Grzhibovsky, Doctor of Medical Sciences, REAVIZ Private University, Saint-Petersburg, Russia; Ammosov North-Eastern Federal University Healthcare Organization and Preventive Medicine Department, Yakutsk, Russia; ORCID: 0000-0002-5464-0498; e-mail: a.grjibovski@yandex.ru

Anton N. Gvozdetskiy, Candidate of Medical Sciences, Mechnikov North-Western State Medical University Psychiatry and Narcology Department, Saint Petersburg, Russia; ORCID: 0000-0001-8045-1220; e-mail: comisora@yandex.ru

Maksim A. Polidanov, REAVIZ Private University, Saint-Petersburg, Russia; ORCID: 0000-0001-7538-7412; e-mail: maksim.polidanoff@yandex.ru

Radik M. Khayrullin, Doctor of Medical Sciences, REAVIZ Private University, Saint-Petersburg, Russia; ORCID: 0000-0002-2796-7508; e-mail: r.m.hayrullin@reaviz.online