УДК 614.2
http://dx.doi.org/10.22328/2413-5747-2023-9-3-102-112

АНАЛИЗ НАДЕЖНОСТИ ОЦЕНКИ ДИХОТОМИЧЕСКИХ ИСХОДОВ: РАЗМЕР ВЫБОРКИ И РАСЧЕТ КАППА-СТАТИСТИКИ

*[1]Е. А. Митькина, [1]Ю. Г. Козлова, [1]М. А. Горбатова, [1,2]А. М. Гржибовский**
[1] Северный государственный медицинский университет, г. Архангельск, Россия
[2]Северо-Восточный федеральный университет, г. Якутск, Россия

Анализ надежности — это важный методологический инструмент, используемый в медицинских исследованиях для определения степени согласованности измерений, проводимых различными методами или несколькими исследователями. В данной статье авторы доходчиво представляют обзор основных концепций, связанных с анализом надежности, а также статистические критерии, используемые при его применении в медицине. Представлены сходства и отличия анализа валидности (достоверности) от анализа надежности (воспроизводимости). Демонстрируются принципы расчета каппа-статистики для простейшей ситуации с двумя исследователями и бинарными признаками как с помощью формул, так и с помощью пакета статистических программ SPSS, а также ее достоинства и ограничения. Статья предназначена для начинающих исследователей и молодых ученых и будет полезна при планировании исследований и подготовке сборщиков данных.

КЛЮЧЕВЫЕ СЛОВА: морская медицина, анализ надежности, достоверность, каппа-статистика, статистическая мощность, SPSS, размер выборки, калибровка

Для корреспонденции: *Гржибовский Андрей Мечиславович, e-mail: A.Grjibovski@yandex.ru*
For correspondence: *Andrej M. Grjibovski, e-mail: A.Grjibovski@yandex.ru*

Для цитирования: Митькина Е.А., Козлова Ю.Г., Горбатова М.А., Гржибовский А.М. Анализ надежности оценки дихотомических исходов: размер выборки и расчет каппа-статистики // *Морская медицина.* 2023. Т. 9, № 3. С. 102–112, doi: http://dx.doi. org/10.22328/2413-5747-2023-9-3-102-112
For citation: Mitkina E.A., Kozlova Yu.G., Gorbatova M.A., Grjibovski A.M. Reliability analysis of binary outcomes: sample size and calculation of kappa statistic // *Marine Medicine.* 2023. Vol. 9, № 3. P. 102–112, doi: https://dx.doi. org/10.22328/2413-5747-2023-9-3-102-112

RELIABILITY ANALYSIS OF BINARY OUTCOMES: SAMPLE SIZE AND CALCULATIONS OF KAPPA STATISTIC

*[1]Ekaterina A. Mitkina, [1]Yulia G. Kozlova, [1]Maria A. Gorbatova, [1,2]Andrej M. Grjibovski**
[1] Northern State Medical University, Arkhangelsk, Russia
[2] North-Eastern Federal University, Yakutsk, Russia

Reliability analysis is an important methodological tool used in medical research to assess the degree of agreement between measurements taken by different methods or by multiple investigators. In this article, we provide an easy-to-understand overview of the basic concepts associated with reliability analysis, as well as the statistical criteria used in its application in biomedical research. The similarities and differences between the analysis of validity and the analysis of reliability are also presented. The principles of calculating Cohen's kappa for the simplest situation with two researchers and binary variables are demonstrated both by using the formulas and by applying the SPSS software. Advantages and disadvantages of using kappa statistic are discussed. The article is intended for novice researchers and young scientists and will be useful for planning of research projects and training data collectors.

KEYWORDS: marine medicine, reliability analysis, validity, kappa-statistic, statistical power, SPSS, sample size, calibration

Введение. Достоверность и надежность — две важнейшие составляющие успешных исследований. Эти понятия тесно взаимосвязаны, поэтому большинство начинающих исследователей часто либо объединяют их в одну мысль, либо и вовсе путают. В любой исследовательской работе даже ненамеренно может быть допущено большое количество ошибок, поэтому сведение их числа до минимума является важным этапом планирования исследования. Надежность собранных данных – важнейший компонент общей уверенности в достоверности исследования.

Анализ надежности (reliability analysis) — расчет меры согласованности, показывающей, будет ли результат одинаковым каждый раз при повторной оценке у того же участника исследования тем же методом. Следовательно, высокая надежность свидетельствует о высокой воспроизводимости данных, а низкая может говорить о высокой доле случайности в полученных результатах измерений, обусловленной несовершенством механизма измерения или недостаточной подготовкой персонала.

Оценка валидности, или достоверности (validity analysis) — мера точности результатов или степень их соответствия объективной реальности. Валидность (достоверность) отражает, насколько правдиво результаты исследования соответствуют установленным критериям, по которым оцениваются научные данные (часто в сравнении с «золотым стандартом») [1]. Именно исследования, в которых результаты соответствуют объективной реальности, считаются достоверными, а не те, в которых демонстрируются формулировки «$p < 0,05$». Достоверность исследования труднее поддается оценке, чем надежность, так как для измерения многих признаков не существует «золотого стандарта», но в любом случае исследователям следует помнить, что для получения объективных результатов исследования методы сбора данных должны быть валидными.

Оценка надежности и валидности необходима для понимания, насколько хорошо методология, техника сбора и анализ данных были спланированы для измерения изучаемых признаков [2]. Надежность оценивается путем проверки согласованности результатов либо во времени, либо между разными наблюдателями. Существует несколько типов надежности: среди нескольких сборщиков данных, что называется межоценочная (межкорреляционная, межэкспертная), или одного сборщика данных в разное время – внутриоценочная (внутрикорреляционная, внутриэкспертная) [3].

В сфере здравоохранения для сбора данных чаще всего привлекают несколько человек. При этом возникает вопрос о согласованности или согласии между людьми, собирающими данные, вследствие вариативности показателей среди наблюдателей [4]. Поэтому хорошо спланированные исследования должны включать процедуры, измеряющие согласованность результатов между исследователями при измерении одного и того же признака.

В литературе имеется достаточное количество данных для оценки точности и согласованности измерений, то есть валидности и надежности, но этот аспект часто либо полностью игнорируется, либо не обсуждается должным образом при планировании исследований, особенно молодыми учеными и начинающими исследователями [5]. В значительном количестве публикаций исследователи не обсуждают надежность своих инструментов, что может критически настроенных читателей заставить усомниться в воспроизводимости результатов и общем качестве работы [6, 7]. Чаще всего это ограничение связано с недостатком знаний среди исследователей о том, как оценивают валидность и надежность два этих инструмента.

Целью данной работы является описание анализа надежности, статистических критериев, применяемых для данного вида анализа, представление расчета каппа-статистики с помощью формул и программного обеспечения, а также демонстрация расчета, необходимого размера выборки для калибровки исследователей для выбранных уровней каппа-статистки.

При планировании исследования важно включить обучение сборщиков данных и их калибровку [8], то есть провести измерение признаков на тестовой выборке или стандарте и оценить степень согласованности полученных результатов. Степень согласия между сборщиками данных называется надежностью. Анализ надежности играет важную роль в проведении исследований из-за того, что несколько человек, собирающих данные, могут по-разному воспринимать и интерпретировать интересующие явления. Вследствие этого основной целью калибровки является обеспечение одинакового толкования и понимания всеми исследователями критери-

ев оценки различных явлений и состояний [9], которые подлежат выявлению и регистрации, а также уверенность в том, что каждый исследователь может осматривать участников исследования с постоянной точностью.

Программа обучения и калибровки сборщиков данных может включать:

1) обзор критериев и протокола исследования;

2) наглядный обзор (например, презентация Power Point) с рассмотрением критериев обследования;

3) инструкции по регистрации данных;

4) оценку знаний обучаемых по выделенным критериям;

5) клиническое обучение, инструктаж с демонстрационными обследованиями;

6) практические обследования будущими оценщиками;

7) калибровку исследователей [10].

Пункты с 1 по 6 относятся к обучающей части программы. Перед каждой встречей каждый исследователь обычно получает информационный материал, объясняющий цель программы обучения, протоколы подготовки и калибровки, а также письменные инструкции по клиническому обследованию пациентов или объектов калибровки. Обучающиеся должны быть проинформированы о том, что эффективность программы обучения будет оцениваться [11], например, с помощью письменного теста, который должен быть успешно пройден на определенный пороговый процент, установленный заранее [10]. До начала объяснения и калибровки обучающий эксперт должен лично просмотреть учебный материал, прежде чем встретиться с испытуемыми для обучения и калибровки. Желательно, чтобы эксперт имел большой опыт проведения калибровочных программ. Пункт 7 выполняется на заключительной калибровочной части, когда происходит оценка схожести результатов между исследователями команды, то есть оценивается надежность исследования. К тому же на данном этапе можно сравнить результаты каждого исследователя с «золотым стандартом» для оценки достоверности или валидности. Как правило, обзор критериев и протокола исследования, обучение критериям проходят за пару дней. Последующие несколько дней отводятся на калибровку специалистов для того, чтобы оценить надежность обследования. Каждый исследователь должен сначала попрактиковаться, к примеру, осмотрев группу из 10 человек, а затем – группу из 20 или более человек, и сравнить свои результаты с данными, полученными другими членами команды при осмотре той же группы. Важно, чтобы команда специалистов могла проводить осмотры с приемлемым постоянством, используя общепринятые стандарты [12]. Уровень постоянства для больших исследований должен быть, по данным ВОЗ, в пределах 85–95 %. Если результаты какого-либо специалиста постоянно отличаются от результатов большинства в значительной степени, он исключается из команды исследователей. Оценщик также может быть исключен в силу других обстоятельств.

Для оценки достоверности исследования необходимо выбрать валидатора команды исследователей, который был заранее обучен в соответствии с предлагаемой методологией проведения обследований. Для команды исследователей валидатор является «золотым стандартом», с которым специалисты могут сравнивать свои результаты обследования [13].

Программу обучения и калибровки исследователей можно рассмотреть на примере стоматологического обследования населения командой специалистов. При проведении обследования важно, чтобы все специалисты были подготовлены таким образом, чтобы одинаково оценить стоматологический статус пациентов. Рассмотрим гипотетический пример калибровки исследователей для эпидемиологического стоматологического обследования населения для демонстрации записи данных, их анализа, а также применения пакета статистических программ SPSS для расчета показателей надежности.

Итак, допустим, что для проведения калибровки двух специалистов были обследованы 13 испытуемых (7 мужчин и 6 женщин от 12 до 25 лет). Для процесса калибровки было отобрано 2 специалиста-стоматолога. Группу пациентов из 13 человек обследовали на наличие кариеса в полости рта. Были установлены следующие критерии оценки: 1 – наличие кариозного процесса в полости рта пациента; 0 – отсутствие кариозного процесса в полости рта пациента. Результаты исследования можно оформить в виде таблицы (табл. 1), в которой столбцы представлены оценщиками, а в строках – испытуемыми, то есть пациентами. Ячейки в таблице содержат результаты исследования, полученные оценщиками для каждого испытуемого. Для оценки валидности исследования можно доба-

вить столбец, в котором представлен валидатор команды («золотой стандарт», или эталон).

Существует ряд показателей для оценки надежности и достоверности результатов [3, 14]. В табл. 2 представлены наиболее часто используемые из них.

В нашем случае речь идет о дихотомических исходах, поэтому в работе мы рассмотрим только те статистические показатели, которые актуальны для бинарных переменных – общий процент согласия, каппа-статистика и взвешенная каппа-статистика.

Таблица 1

Результаты оценки обследования пациентов двумя исследователями и валидатором

Table 1

Results of the patient examination by two investigators and a validator (gold standard)

Испытуемый	Исследователь 1	Исследователь 2	Валидатор
1	1	0	1
2	0	1	1
3	0	0	0
4	1	1	1
5	1	1	1
6	0	1	0
7	1	1	0
8	0	1	1
9	0	0	0
10	0	1	0
11	1	1	1
12	0	0	1
13	1	0	1

Таблица 2

Статистические коэффициенты, используемые для анализа надежности и достоверности

Table 2

Statistical procedures used in analysis of reliability and validity

Показатель	Используется для оценки	
	надежности	достоверности
Общий процент согласия	+	±
Процент положительного результата	+	±
Каппа-статистика	+	±
Взвешенная каппа-статистика	+	±
Коэффициент внутриклассовой корреляции	+	±
Коэффициент конкордации Кендалла	+	±
Коэффициент корреляции Пирсона	±	±
Коэффициент корреляции Спирмена	±	±
Чувствительность/специфичность	-	+
J – статистика (индекс) Юдена	-	+

Примечание: + – показатель используется для анализа; ± – показатель может быть использован, но его показания неоднозначны; - – показатель не используется

Note: + – the indicator is used for analysis; ± – the indicator can be used, but its readings are ambiguous; - – the indicator is not used

Дихотомические, или бинарные исходы – это данные, которые могут быть выражены только двумя альтернативными переменными [15]. Если, например, взять варианты из нашей гипотетической калибровки специалистов, такими данными будут: «наличие кариозного процесса» и «отсутствие кариозного процесса». Эти величины являются взаимоисключающими (рис. 1).

Оценка результатов надежности, как уже говорилось выше, может быть получена многими путями, простейшим из которых является процент соглашений между показателями, то есть процент пациентов, по которым два исследователя зарегистрировали одинаковую величину показателя: либо оба отметили наличие кариеса, либо оба отметили его отсутствие. Его также можно рассчитать не только для бинарных исходов, но и для любого количества категорий качественных (категориальных) признаков. Общий процент согласия рассчитывается как количество баллов согласия, деленное на общее количество баллов, что можно записать как: общий процент согласия (percent agreement) = $(a + d) / (a + b + c + d) \cdot 100$ [3].

Несмотря на простоту расчета, общий процент согласия не может являться наилучшей опцией для анализа надежности, так как он не учитывает случайность совпадений ответов исследователей или простое угадывание [16]. Для решения этой проблемы используют более точный статистический инструмент – каппа-статистику. Ее преимущество перед другими методами состоит в том, что она учитывает вероятность случайного согласия между исследователями.

Рассмотрим вариант «угаданных» результатов на примере гипотетического исследования.

		Исследователь 2	
		Есть признак	Нет признака
Исследователь 1	Есть признак	a	b
	Нет признака	c	d

Рис. 1. Четырехпольная таблица для расчета статистических показателей анализа надежности
Fig. 1. Two-by-two table for calculating coefficients used in reliability analysis

Предположим, что в качестве «исследователей» были взяты случайные люди, которые оценивают наличие или отсутствие кариозного процесса наугад. Первый исследователь определил положительный результат – 46 % (6 из 13) всех случаев, а второй – 62 % (8 из 13). Ситуации, в которых «исследователи» ставят одинаковый результат – 54 % (7 из 13). Это число должно говорить о том, что случайные сборщики данных хорошо справляются с диагностикой, потому что они достигли умеренного согласия в большинстве случаев. Однако этот гипотетический пример показывает, что оценка согласия в реальных исследованиях сильно завышена за счет случайного согласия [17]. Это завышение должно быть устранено для понимания истинной картины.

Каппа как статистический инструмент широко используется во многих областях здравоохранения для сбора исследовательских или клинических лабораторных данных. Первоначально она была введена Джейкобом Коэном, выдающимся американским статистиком, который в 1960 году разработал величину для измерения межрейтинговой надежности. Коэн обратил внимание на то, что между сборщиками данных может быть определенный уровень согласия, даже если они не знают правильного ответа, а просто угадывают его. Он предположил, что определенное количество догадок будет совпадать, и что статистика надежности должна учитывать это случайное согласие [18].

Каппа-статистика, обозначаемая строчной греческой буквой к, – это степень согласия между двумя или более исследователями по сравнению с величиной согласия, которое можно было бы ожидать в результате случайности, если бы оценки были статистически независимыми. Самое простое использование к для ситуации – когда, например, два врача дают заключение одному и тому же пациенту или когда один врач делает два заключения в разные моменты времени, что представляет собой межэкспертную и внутриэкспертную надежность соответственно. Каппа-статистика также может быть адаптирована для ситуации с наличием более двух экспертов или врачей, однако внимание будет сосредоточено на простой ситуации, когда два исследователя дают одну независимую оценку для каждого пациента относительно наличия бинарного исхода.

Каппа-статистика измеряет согласованность данных, которую упрощенно можно представить как свободную от случайности, и определяется как:

$$к = \frac{\textit{Доля наблюдаемого согласия} - \textit{Доля ожидаемого случайного согласия}}{1 - \textit{Доля ожидаемого случайного согласия}} = к = \frac{Pr\,(н) - Pr\,(о)}{1 - Pr\,(о)},$$

где Pr (н) представляет собой фактическую (наблюдаемую) вероятность согласия, а Pr (о) – это ожидаемое согласие при независимой оценке.

Наблюдаемую и ожидаемую вероятности согласия или совпадения заключений исследователей можно рассчитать с помощью простых формул на основании четырехпольной таблицы (рис. 2):

$$Pr\,(о) = \frac{\frac{(f1 \times g1)}{n} + \frac{(f2 \times g2)}{n}}{n}$$

$$Pr\,(н) = \frac{(a) + (d)}{n}.$$

Каппа-статистика может варьировать от -1 до 1, где 0 означает, что согласие между результатами совершенно случайно, а 1 представляет собой идеальное согласие, указывающее на то, что исследователи полностью согласны в классификации каждого случая. Отрицательные значения говорят о согласии, которое хуже, чем может быть обусловлено случайностью, то есть речь может идти о каких-то систематических различиях между исследователями, которые необходимо либо нивелировать, либо пересмотреть список исследователей [19]. Как и для коэффициентов корреляции, для абсолютных значений каппа-статистики есть несколько вариантов интерпретации, представленных в табл. 3. Несмотря на большое количество предложенных интерпретаций, ни одна из них не является общепринятой.

Любая каппа ниже 0,60 указывает на недостаточное согласие между исследователями или сборщиками данных, поэтому результатам таких исследований не стоит доверять. Всемирная организация здравоохранения, например, предлагает считать минимально допустимым согласием значение к = 0,81. При калибровке исследователей необходимо проводить обучение до тех пор, пока не будет достигнуто значение минимум 0,81 или выше.

Если в распоряжении исследователей имеется «золотой стандарт» для измерения изучаемого признака, то исследование надежности уже можно считать исследованием валидности, или достоверности. В табл. 1 это будет сравнение не между исследователями, а сравнение с валидатором. Статистическая часть, если речь идет о каппа-статистике, не изменится, но дополнительно следует рассчитывать чувствительность, специфичность, а также прогностическую ценность положительного и отрицательного результатов [20].

Несмотря на то что все расчеты можно провести вручную, мы покажем, как рассчитать каппа-статистику и достигнутый уровень значимости на основании данных табл. 1 с помощью пакета статистических программ SPSS, который все еще является одним из наиболее попу-

		Исследователь 2		
		Есть признак	Нет признака	**Итог исследователя 1**
Исследователь 1	Есть признак	a	b	g1
	Нет признака"	c	d	g2
	Итог исследователя 2	f1	f2	n

Рис. 2. Четырехпольная таблица для расчета каппа-статистики с помощью формул.
Fig. 2. Two-by-two table for manual calculation of kappa statistic.

Таблица 3

Интерпретация значений каппа-статистики

Table 3

Interpretation of kappa values

Значение	Уровень согласия
Landis & Koch, 1977	
0,00–0,20	Незначительный
0,21–0,40	Удовлетворительный
0,41–0,60	Умеренный
0,61–0,80	Существенный
0,81–1,00	Практически идеальный
Fleiss, 1981	
< 0,40	Плохой
0,41–0,75	От удовлетворительного до хорошего
0,76–1,00	Превосходный
Altman, 1991	
< 0,20	Плохое согласие
0,21–0,40	Удовлетворительное согласие
0,41–0,60	Умеренное согласие
0,61–0,80	Хорошее согласие
0,81–1,00	Отличное согласие
Burt, 1996	
0,00–0,20	Плохой
0,21–0,40	Незначительный
0,41–0,60	Удовлетворительный
0,61–0,80	Хороший
0,81–0,92	Отличный
0,93–1,00	Превосходный

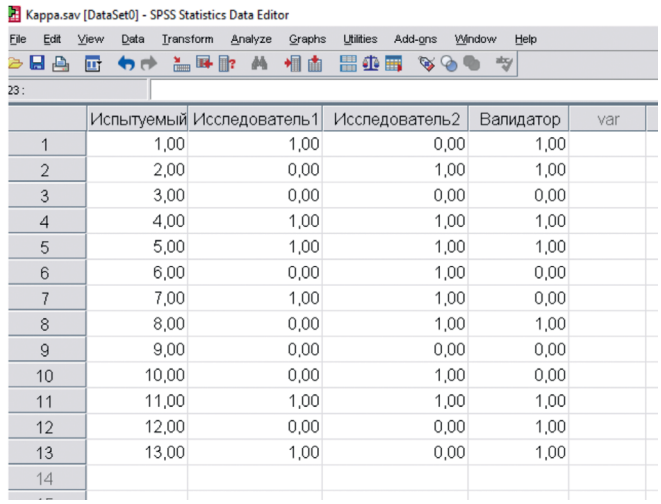

Рис. 3. Вид в программе SPSS данных примера из таблицы 1.

Fig. 3. Example from Table 1 in SPSS data window.

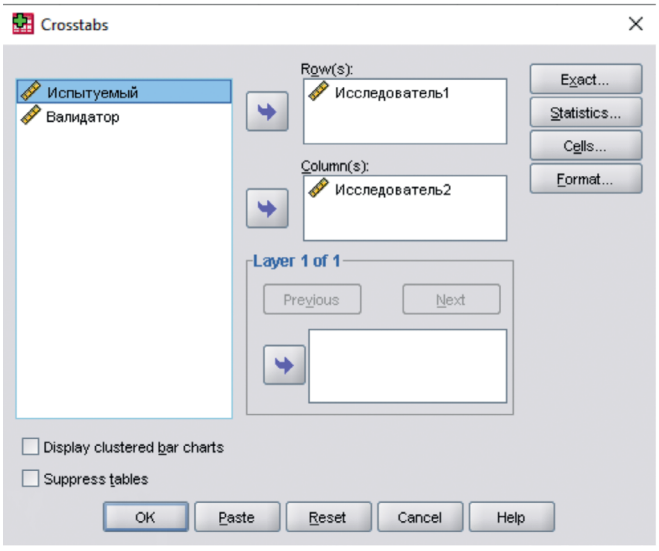

Рис. 4. Диалоговое окно для построения таблиц сопряженности.

Fig. 4. Dialog box for crosstabulation.

лярных в отечественном медицинском научном сообществе.

В окне данных табл. 1 будет иметь тот же вид (рис. 3).

Для построения таблицы сопряженности и расчета каппа-статистики следует в верхней строке окна данных выбрать Analyze, в выпадающем меню выбрать Descriptive statistics, в котором, в свою очередь, выбрать Crosstabs. Это приведет к появлению диалогового окна, в котором из левого поля надо перевести переменные «Исследователь 1» и «Исследователь 2» в поля справа, как показано на рис. 4, после чего в том же диалоговом окне нажать на кнопку Statistics и галочкой выбрать нужный нам статистический критерий – Kappa (рис. 5). После выбора критерия вернуться в прежнее окно можно нажатием на кнопку Continue.

Запуск анализа осуществляется нажатием на кнопку OK в диалоговом окне, показанном на рис. 4. Можно также сохранить синтаксис для выбранных манипуляций, нажав кнопку Paste. Для нашего примера синтаксис будет выглядеть следующим образом:

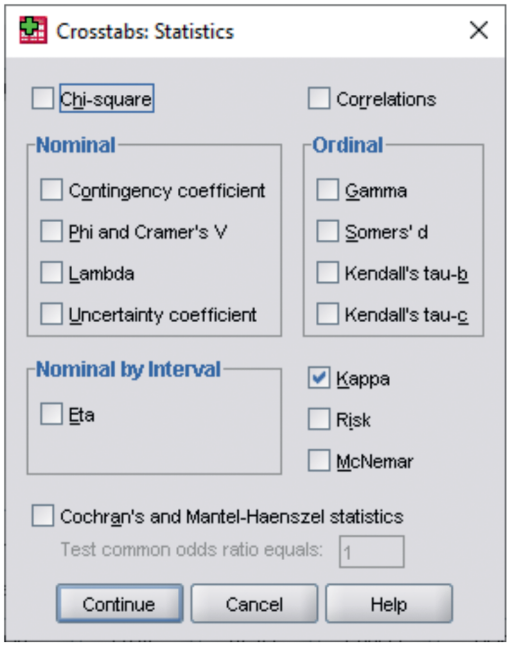

Рис. 5. Диалоговое окно Crosstabs: Statistics и выбор каппа-статистики.
Fig. 5. Dialog box Crosstabs: Statistics with selection of kappa statistic.

Исследователь 1 * Исследователь 2 Crosstabulation

Count

		Исследователь 2		Total
		,00	1,00	
Исследователь1	,00	3	4	7
	1,00	2	4	6
Total		5	8	13

Рис. 6. Таблица сопряженности в SPSS для примера, показанного в табл. 1.
Fig. 6. Contingency table in SPSS with the data from Table 1.

```
CROSSTABS
/TABLES = Исследователь1 BY Исследователь 2
/FORMAT = AVALUE TABLES
/STATISTICS = KAPPA
/CELLS = COUNT
/COUNT ROUND CELL.
```

После запуска анализа программа выдаст три таблицы, первая из которых описательная и не будет рассматриваться для экономии места. Вторая таблица (рис. 6) представляет собой таблицу сопряженности, в которой представлены абсолютные числа.

В следующей таблице дано абсолютное значение каппа-статистики и результат проверки нулевой гипотезы о полученном коэффициенте, равном нулю (рис. 7).

SPSS выдает абсолютное значение каппа-статистики (0,093), значение стандартной ошибки показателя (0,262), а также уровень значимости, полученный при проверке нулевой гипотезы (0,725). Полученные результаты следует интерпретировать как отсутствие связи между исследователями в оценке кариеса, за исключением случайных совпадений, так как коэффициент к очень мал (менее 0,1), а уровень значимости 0,725 значительно превышает общепринятое в биомедицинских исследованиях значение критического уровня значимости 0,05, то есть мы вынуждены принять нулевую гипотезу о равенстве каппа-статистики нулю. Следует обратить внимание на то, что в данном примере общая доля согласия составляет 7/13 = 54 %, но ее можно отнести к случайности.

Часто задаваемый вопрос – сколько испытуемых необходимо набрать для проведения кали-

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	,093	,262	,352	,725
N of Valid Cases		13			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Рис. 7. Результат, выдаваемый SPSS о результатах расчета каппа-статистики.
Fig. 7. SPSS output with the results of kappa statistic calculation.

бровки исследователей? Если принять во внимание общепринятые в медицине уровни альфа- и бета ошибок 5 % и 20 % соответственно, то для ситуации с распространенностью признака 50 % достаточная статистическая мощность для отклонения нулевой гипотезы для каппа-статистики, равной 0,8, будет достигнута уже при выборке в 10 человек. Рассчитанный размер выборки для каппа-статистики от 0,1 до 0,9 показан на рис. 8.

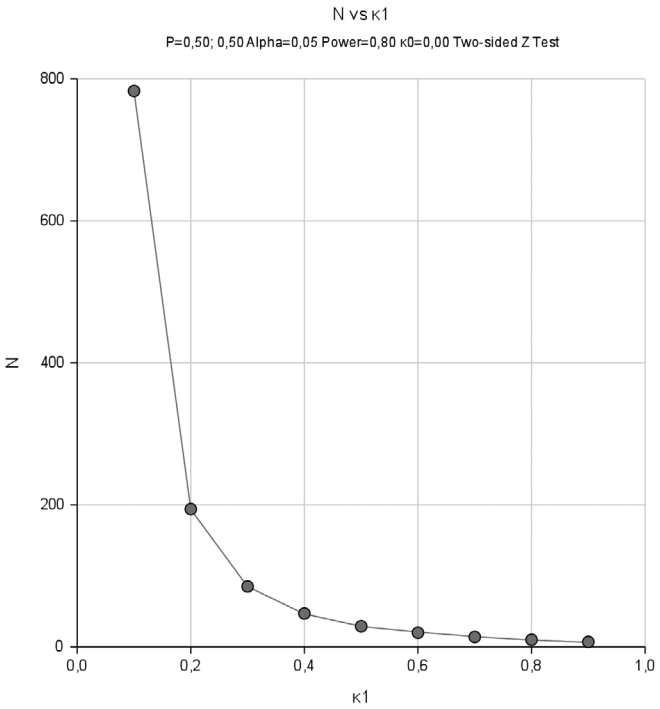

Рис. 8. Размер выборки для расчета каппа-статистики.
Fig. 8. Sample size for kappa statistic calculation.

Таблица 4

Размер выборки, позволяющий отклонить нулевую гипотезу, к = 0,8 для каппа-статистики от 0,90 до 0,99 при распространенности изучаемого признака 50 %

Table 4

Sample size for kappa values from 0,90 through 0,99 sufficient to reject H0: к = 0,8 with a prevalence of the studied outcome of 50%

Каппа-статистика	Размер выборки	Нулевая гипотеза (каппа)
0,90	239	0,80
0,91	193	0,80
0,92	158	0,80
0,93	131	0,80
0,94	110	0,80
0,95	93	0,80
0,96	78	0,80
0,97	66	0,80
0,98	56	0,80
0,99	47	0,80

Следует отметить, что для других показателей распространенности признака числа будут другие. Например, при распространенности признака 10 % или 90 % необходимо будет включить в исследование не 10, а 13 человек, то есть на 30 % больше. Однако приведенные выше расчеты подразумевают значение к = 0 в качестве нулевой гипотезы.

Если же в процессе планирования крупного исследования ставится задача добиться соответствия между исследователями как минимум 0,81, то в качестве нулевой гипотезы можно использовать значение к = 0,8, что значительно увеличивает размер выборки, но это оправданно для получения надежных результатов в крупных популяционных исследованиях (табл. 4).

Таким образом, можно заключить, что каппа-статистика является широко используемой мерой согласованности в медицинских исследованиях. Однако у нее есть некоторые недостатки, например, чувствительность к дисбалансу классов или распространенности изучаемого признака. Высокая распространенность признака может уменьшить каппа-статистику, так как в этом случае вероятность случайного совпадения между исследователями становится выше. В этом случае даже небольшое количество ошибок в классификации может привести к значительному снижению согласованности между наблюдателями и, следовательно, к уменьшению значения каппа-статистики [21], о чем следует помнить при интерпретации результатов. Более того, каппа-статистика учитывает не порядок классов, а только их согласованность. Это значит, что она может дать одинаковые результаты для разных порядков классов. Для учета порядка классов можно использовать взвешенную каппа-статистику или другие меры согласованности, такие как взвешенный индекс Жаккара (Weighted Jaccard Index).

Тем не менее несмотря на то что в современном статистическом арсенале имеются средства, учитывающие недостатки каппа-статистики, этот метод по-прежнему широко распространен в медицинских исследованиях, причем не только для калибровки исследовате-

лей перед началом крупных научных проектов. Каппа-статистика может использоваться для оценки согласованности между различными методами диагностики заболеваний, такими как диагностические тесты, физикальные исследования и инструментальная диагностика, подразумевающими категориальный ответ. Например, каппа-статистика может использоваться для оценки согласованности между различными методами определения эффективности лечения: клинические испытания, результаты лабораторных тестов и жалобы пациентов. Также этот метод может использоваться для оценки согласованности между наблюдателями при интерпретации изображений, таких как рентгенограммы, МРТ и УЗИ. Вот несколько конкретных примеров использования каппа-статистики: 1) оценка согласованности между патологоанатомами при определении стадии рака; 2) оценка согласованности между врачом и разработанной компьютерной программой с использованием технологий машинного обучения при диагностике интересующего врача заболевания; 3) оценка согласованности между двумя экспертами при интерпретации рентгенограммы пациента при постановке диагноза и прочее.

Заключение. Таким образом, каппа-статистика может быть полезной не только в медицине, но и в других областях, где необходимо оценить согласованность между двумя или более исследователями (наблюдателями, экспертами, регистраторами, датчиками и пр.) при изучении категориальных признаков, а использование программного обеспечения делает расчет каппа-статистики доступным любому начинающему исследователю. Авторы надеются, что этот обзор поможет читателям лучше понять сущность анализа надежности, а также важность оценки согласованности между наблюдателями при подготовке исследований.

Сведения об авторах:

Митькина Екатерина Андреевна – студентка стоматологического факультета федерального государственного бюджетного образовательного учреждения высшего образования «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации, 163069, г. Архангельск, Троицкий проспект, д. 51; e-mail: miekandr@yandex.ru; ORCID 0000-0002-5631-5197.

Козлова Юлия Геннадьевна – студентка стоматологического факультета федерального государственного бюджетного образовательного учреждения высшего образования «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации, 163069, г. Архангельск, Троицкий проспект, д. 51; e-mail: iuliak0z@yandex.ru; ORCID 0009-0003-6496-2719.

Горбатова Мария Александровна – кандидат медицинских наук, магистр общественного здоровья, доцент кафедры стоматологии детского возраста федерального государственного бюджетного образовательного учреждения высшего образования «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации, 163069, г. Архангельск, Троицкий проспект, д. 51; e-mail: marigora@mail.ru ; ORCID 0000-0002-6363-9595.

Гржибовский Андрей Мечиславович — доктор медицинских наук, начальник управления научно-инновационной работы, заведующий центральной научно-исследовательской лабораторией федерального государственного бюджетного образовательного учреждения высшего образования «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации, 163069, г. Архангельск, Троицкий проспект, д. 51; профессор кафедры общественного здоровья, здравоохранения, общей гигиены и биоэтики федерального государственного автономного образовательного учреждения высшего образования «Северо-Восточный федеральный университет имени М. К. Аммосова», 677007, Республика Саха (Якутия), г. Якутск, ул. Кулаковского, д. 42; e-mail: A.Grjibovski@yandex.ru; ORCID 0000–0002–5464–0498; SPIN 5118–0081.

Information about the authors:

Ekaterina A. Mitkina – student of the faculty of dentistry Northern State Medical University; Troitsky Av., 51, 163069 Arkhangelsk, Russian Federation; e-mail: miekandr@yandex.ru; ORCID 0000-0002-5631-5197

Yulia G. Kozlova - student of the faculty of dentistry, Northern State Medical University, Troitsky Av., 51, 163069, Arkhangelsk, Russian Federation; e-mail: iuliak0z@yandex.ru; ORCID 0009-0003-6496-2719

Maria A. Gorbatova – Cand. of Sci. (Med.), MPH, Associate professor at the Department of Pediatric Dentistry, Northern State Medical University; Troitsky Av., 51, 163069, Arkhangelsk, Russian Federation; e-mail: marigora@mail.ru ; ORCID 0000-0002-6363-9595.

Andrej M. Grjibovski — D-r of Sci. (Med), Master of International Community Health, Head of the Directorate for Research and Innovations, Director of the Central Scientific Research Laboratory, Northern State Medical University; 51, Troitskiy Av., Arkhangelsk, 163069, Russian Federation; Professor at the Department of Public Health, Public Health, General Hygiene and Bioethics, North-Eastern Federal University; 42 Kulakovskogo St., Yakutsk; Sakha Republic (Yakutia), 677007, Russian Federation; e-mail: A.Grjibovski@yandex.ru; ORCID 0000–0002–5464–0498; SPIN: 5118–0081.

Вклад авторов. Все авторы подтверждают соответствие своего авторства, согласно международным критериям ICMJE (все авторы внесли существенный вклад в разработку концепции, проведение исследования и подготовку статьи, прочли и одобрили финальную версию перед публикацией).

Authors' contributions. All authors meet the ICMJE authorship criteria: all authors significantly contributed to concept and design, performed the research and drafted the manuscript. All authors approved the final version of the paper.

Потенциальный конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Disclosure of conflicts of interest. The authors declare that they have no competing interests.

Финансирование. Авторы не имеют финансовой заинтересованности в представленных материалах или методах.
Funding. No author has a financial or property interest in any material or method mentioned.

Поступила / Received: *11.08.2023*
Принята к печати / Accepted: *02.09.2023*
Опубликована / Published: *30.09.2023*

ЛИТЕРАТУРА / REFERENCES

1. Whittemore R., Chase S.K., Mandle C.L. Validity in Qualitative Research. *Qualitative Health Research*, 2001, Vol. 11, № 4, pp. 522–537. doi: 10.1177/104973201129119299.
2. Ahmed I., Ishtiaq S. Reliability and validity: Importance in Medical Research. *J Pak Med Assoc*, 2021, Vol. 71, № 10, pp. 2401–2406. doi: 10.47391/JPMA.06-861.
3. McHugh Mary L. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 2012, Vol. 22, № 3, pp. 276–282.
4. Tang W., Hu J., Zhang H., Wu P., He H. Kappa coefficient: a popular measure of rater agreement. *Shanghai Arch Psychiatry*, 2015, Vol. 27, № 1, pp. 62–67. doi: 10.11919/j.issn.1002-0829.215010.
5. Noble H., Smith J. Issue of validity and reliability in quantitative research. *Evid Based Nurs*, 2015, Vol. 18, № 2, pp. 34–35. doi: 10.1136/eb-2015-102054.
6. Aoki K., Hall T., Takasaki H. Reporting on the level of validity and reliability of questionnaires measuring Katakori severity: A systematic review. *SAGE Open Med*, 2019, Vol. 7, pp. 1–13. doi: 10.1177/2050312119836617.
7. Akturk Z. Reliability and validity in medical research. *Dicle Med J*, 2012, Vol. 39, № 2, pp. 196–202. doi: 10.5798/diclemedj.0921.2012.02.0150.
8. Fyffe H.E., Deery C., Nugent Z.J., Nuttall N.M., Pitts N.B. Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee selectable threshold method for caries diagnosis (DSTM). *Community Dent Oral Epidemiol*, 2000, Vol. 28, № 1, pp. 42–51. doi: 10.1034/j.1600-0528.2000.280106.x.
9. Рождественская Е.Ю. Надежность качественных методов и качество данных // *INTER*. 2014. Т. 8. С. 16–28 [Rozhdestvenskaya E.Y. Reliability of qualitative methods and data quality. *INTER*, 2014, № 8, 16–29 (In Russ.)].
10. Rechmann P., Jue B., Santo W., Rechmann B.M.T., Featherstone J.D.B. Calibration of dentists for Caries Management by Risk Assessment Research in a Practice Based Research network - cambra pbrn. *BMC Oral Health*, 2018, Vol. 18, №2. doi: 10.1186/s12903-017-0457-3.10.1186/s12903-017-0457-3
11. Tavakol M., Sandars J. Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part II. *Medical Teacher*, 2014, Vol. 36, № 10, pp. 838–848. doi: 10.3109/0142159X.2014.915297.
12. Warren J.J., Weber-Gasparoni K., Tinanoff N., Batliner T.S., Jue B., Santo W., Garcia R.I., Gansky S.A., Early Childhood Caries Collaborating Centers. Examination criteria and calibration procedures for prevention trials of the Early Childhood Caries Collaborating Centers. *Public Health Dent*, 2015, Vol. 75, № 4, pp. 317–326. doi: 10.1111/jphd.12102.
13. Amarante B.C., Arima L.Y., Pinheiro E., Carvalho P., Michel-Crosato E., Bönecker M. Diagnosis training and calibration for epidemiological studies on primary and permanent teeth with hypomineralization. *Eur Arch Paediatr Dent*, 2022, Vol. 23, № 1, pp. 169–177. doi: 10.1007/s40368-021-00686-3.
14. Shoukri M. Measurement of Agreement. *Wiley StatsRef: Statistics Reference Online*, 2015, pp. 1–31. doi: 10.1002/9781118445112.stat05301.pub2.
15. Donner A., Rotondi M.A. Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters. *The International Journal of Biostatistics*, 2010, Vol. 6, № 1. doi: 10.2202/1557-4679.1275.
16. Hyunsook H., Yunhee C., Seokyung H., Sue K.P., Byung-Joo P. Nomogram for sample size calculation on a straightforward basis for the kappa statistic. *Annals of Epidemiology*, 2014, Vol. 24, № 9, pp. 673–680. doi: 10.1016/j.annepidem.2014.06.097.
17. Guggenmoos-Holzmann I. The meaning of kappa: probabilistic concepts of reliability and validity revisited. *Clin Epidemiol*, 1996, Vol. 49, № 7, pp. 775–782. doi: 10.1016/0895-4356(96)00011-x.
18. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, Vol. 20, №1, pp. 37–46. doi: 10.1177/001316446002000104.
19. Sim J., Wright C.C. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 2005, Vol. 85, № 3, pp. 257–268. doi: 10.1093/ptj/85.3.257.
20. Кригер Е.А., Гржибовский А.М., Постоев В.А. Оценка распространенности заболеваний с учетом диагностической эффективности тестов на примере использования серологических тестов для диагностики новой коронавирусной инфекции (COVID-19) // *Экология человека.* 2022. Т. 29, № 5. С. 301–309 [Kriger E.A., Grjibovski A.M., Postoev V.A. Prevalence assessment adjusted for laboratory test performance using an example of the COVID-19 serological tests. *Ekologiya cheloveka [Human Ecology],* 2022, Vol. 29, № 5, 301–309 (In Russ.)]. doi: 10.17816/humeco108116.
21. Zec S., Soriani N., Comoretto R., Baldi I. High Agreement and High Prevalence: The Paradox of Cohen's Kappa. *Open Nurs J,* 2017, Vol. 11, pp. 211–218. doi: 10.2174/1874434601711010211.